Comparison of AI Models and Methods for Infilling Hydrological Time Series Data

MSc Metropolitan Analysis, Design & Engineering
TU Delft & Wageningen University

Sam Groen sam.groen@wur.nl 1184261

Academic Supervisors
Dr. Lixia Chu lixia.chu@wur.nl
Wageningen University
Agrotechnology & Food Sciences - Environmental Technology (ETE)

Dr. Ir. Olivier Hoes o.a.c.hoes@tudelft.nl
TU Delft
Civil Engineering & Geosciences - Water Management

Internship Supervisors
Michel Zuijderwijk michel.zuijderwijk@witteveenbos.com
Witteveen+Bos
Water Management

Arie de Niet arie.de.niet@witteveenbos.com
Witteveen+Bos
Coasts, Rivers & Land Reclamation

August 20, Amsterdam

Abstract

Monitoring and modelling water systems and potential interventions help to manage water more effectively. However, due to defects and operational failures in sensors, the reliability of monitoring and modelling is affected by incomplete data. Because traditional statistical methods are often ineffective in infilling large and multivariate gaps, several machine learning (ML) techniques have been explored. However, a comprehensive overview of ML techniques and their performance across different methods and hydrologic regimes is missing. Therefore, this thesis project tackles the question: *What ML approach(es) are most suitable for infilling gaps in hydrological time series in The Netherlands?*

This thesis evaluates ML models and their behaviour with different methods and hydrologic regimes. This study will help to understand the generalisability of the ML models, while also enabling improvements in (urban) water management. In this thesis, a multicriteria analysis (MCA) was used to assess a wide range of ML models. Next, during two case studies, a subset of these models were applied to infill gaps in groundwater, sewage water and surface water levels using the intra-station and inter-station methods.

In the MCA, it was found that Support Vector Regressor (SVR), Random Forest (RF), Gradient Boosting Trees (GBT), Multilayer Perceptron (MLP), Self-Organising Map (SOM) and Long-Short Term Memory (LSTM) were sufficiently suitable.

In the first case study, the use of these models with the intra-station approach led to mixed results for infilling small artificial gaps. Generally, acceptable MSE scores were achieved but poor NSE and KGE scores implied limited scalability.

The second case study showed more promising results with the inter-station method on an artificial gap of seven months. This method proved to be more scalable as all metrics indicated acceptable performance.

In the end, it was concluded that both the RF and GBT models performed most robustly. The MLP and LSTM models showed great potential but suffered from inconsistency, potentially caused by too little training data. It was also found that the inter-station method proved more scalable as compared to the intra-station approach. Furthermore, it was found that success is dependent upon conditions of the hydrologic regime such as human intervention.

Keywords

Machine Learning; Hydrology; Water Management; Gap Infilling; Time Series Data; Random Forest; Gradient Boosting Trees; Neural Networks

Table of Contents

Abstract 2

Table of Contents 3

1. Introduction 5

1.1 Context 5

1.2 Problem Statement 5

1.3 Research Aim 6

1.4 Relevance 6

1.5 Scope 7

1.6 Reading Guide 7

2. Literature Review 9

2.1 Hydrologic Regimes 9

2.2 Existing Infilling Methods 9

2.3 Machine Learning 11

3. ML Model Exploration & Analysis 19

3.1 Methodology 19

3.2 Results 21

3.3 Discussion 25

4. Case Study 1 - Intra-Station Infilling 27

4.1 Data & Methods 27

4.2 Results 30

4.3 Discussion 37

5. Case Study 2 - Inter-Station Infilling 41

5.1 Data & Methods 41

5.2 Results 44

5.3 Discussion 48

6. General Discussion 53

6.1 ML Model Comparison 53

6.2 Systematic Infilling Methods 54

6.3 Comparison of Infilling Performance for Hydrologic Regimes 55

6.4 Limitations 56
7. Conclusion 57
7.1 Conclusion 57
7.2 Recommendations 58
References 61
Appendix A - Resources 67
Appendix B - ChatGPT Conversations 68
Appendix C - Reflection on AI Use 69

1. Introduction

1.1 Context

Monitoring and modelling are powerful tools to improve water management and assess the effects of (potential) measures such as dykes, water level decisions and adequate sewage systems. Monitoring allows for real-time insights into the current state of water bodies and evaluation of whether implemented measures are working as expected (Yang & Liu, 2020; Luo & Wood, 2007). This provides insight into potential weaknesses of the water systems (i.e. floods and droughts), which can be addressed through the implementation of targeted measures. Modelling can help to iteratively improve these potential interventions for optimal results by simulating multiple scenarios (Yang & Liu, 2020). Together, monitoring and modelling provide a strong approach to strengthening water systems and ensuring their resilience.

Monitoring and modelling relies on hydrological time series data such as water level and water flow (Arriagada et al., 2021; Yang & Liu, 2020). However, these hydrological time series datasets can sometimes be incomplete or lack quality (Dembélé et al., 2019; Arriagada et al., 2021). These anomalies can be caused by defect sensors, maintenance, technical failures or other operational problems (Ren et al., 2022; Longman et al., 2020; Dembélé et al., 2019).

1.2 Problem Statement

Incomplete time series affect hydrological monitoring and modelling through ineffective model calibration, unreliable models and biassed statistics (Arriagada et al., 2021; Dembélé et al., 2019). As a result, the (potential) effects of policy measures cannot be modelled and monitored adequately when data is incomplete (Arriagada et al., 2021; Harvey et al., 2012; Wu et al., 2022).

Gaps in time series data come in different patterns. Gaps can be univariate, where one variable is missing, or multivariate, where multiple variables are missing (Hamza et al., 2020). Furthermore, missing data can have different types. First, gaps can occur at random. Its probability of being absent is independent of itself and other variables (Hamza et al., 2020). Second, gaps can occur non-randomly. The probability of these gaps is dependent upon itself or other variables (Hamza et al., 2020).

Example in Sammerspolder, The Netherlands. In Figure A, it can be seen that the water level recorded for Sammerspolder likely rose above -0.1 m NAP during an extreme weather event. However, due to a failure in measurement, no data is available and thus no conclusion can be drawn regarding the effects of this event on the water system.

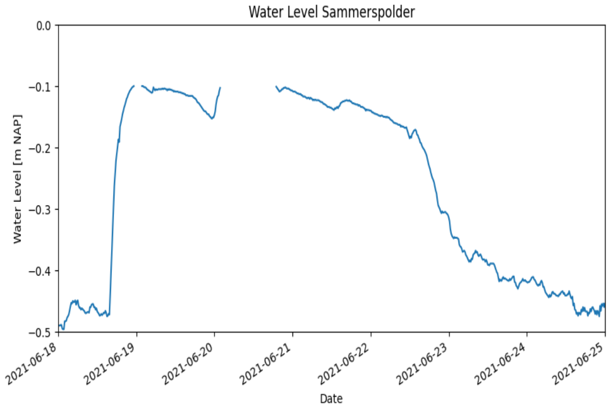


Figure A, water level data at Sammerspolder.

Infilling, the process of filling gaps, of hydrological time series datasets is not straightforward. Hydrological processes are very complex and have a high degree of spatiotemporal variability (Post & Jones, 2001). In addition, different water types (i.e. ground water, surface water and sewage water) have different dynamics. These dynamics are the result of numerous natural and anthropogenic factors such as geology, climate and infrastructure which make up the hydrologic regime (McGregor, 2019; Nowak & Ptak, 2019; Devia et al., 2015; Mackay et al., 2014).

Several statistical methods have already been explored for infilling hydrological time series. These methods range from manual observations to more mathematical infilling solutions (Harvey et al., 2012). However, the effectiveness of these methods deteriorates as gap length increases and gaps become multivariate (He et al., 2020). Furthermore, the missing values often occur in continuous blocks, making statistical infilling methods useless (He et al., 2020). Consequently, these methods are not scalable.

Machine learning (ML) models have emerged as promising solutions as infilling methods. ML enables computers to learn from data without being explicitly programmed (Samual, 1959). Furthermore, ML models are very good at recognising complex patterns and dependencies in datasets (Kazijevs & Samad, 2023; Tang & Ishwaran, 2017). As a result, ML could prove to be a scalable solution to the infilling problem. Several approaches and models have been explored for using ML as an infilling method. These approaches relied upon a station's own data, stations nearby, other relevant variables such as meteorological data, or a mix of these three approaches (Janbain et al., 2023). However, a comprehensive comparison of the performance of ML techniques and approaches was missing. Additionally, a comparison of their performance on different hydrologic regimes was missing.

1.3 Research Aim

The aim of this research project was to compare and assess the infilling performance of a wide range of models and methods on different water types with different hydrological regimes. To guide this research project, the main research question was:

What ML approach(es) are most suitable for infilling gaps in hydrological time series in The Netherlands?

To answer this question, three sub questions are formulated. First, *what ML techniques are most suitable for infilling water level time series data?* The objective for this question was to get an overview of ML techniques already implemented in other research. Additionally, the characteristics of these techniques were assessed to select a set of models for the next phases of the project.

Second, *how effective is infilling based on the use of intra-station data?* The objective in this second part was to explore whether infilling can be achieved with a station's own historic data, instead of relying on other data sources.

Third, *how effective is infilling based on the use of inter-station data?* In this third part, the objective was to use data from other measurement stations to estimate water level of the target station.

This research was conducted in the capacity of a 6 month internship at Witteveen+Bos. The two objectives of this internship were to produce a thesis report and AI-based infilling tool.

1.4 Relevance

1.4.1 Scientific Relevance

This research project is scientifically relevant for several reasons. First, it is helpful to create an overview of the performance of multiple AI models and methods that can be used for infilling incomplete time series data of different water types and hydrological regimes such as groundwater bodies, sewage water systems and surface water bodies. This will help to understand the generalisability of the different ML models and methods, as well as their limitations.

Second, not only the field of hydrology suffers from missing data. Hence, the findings of this overview could be transferable to other domains such as energy, health and mobility. As a result, researchers in these other domains can use these findings to complete their dataset and improve the reliability of their findings (Tang & Ishwaran, 2017).

1.4.2 Societal Relevance

This research project is also socially relevant for several reasons. First, more reliable monitoring and modelling could improve the resilience of our urban and rural areas as water can be managed more effectively. Effective water management can help to protect urban systems against climate impacts like droughts and floods (Dolman, 2021). These climate change effects have disrupted supply chains, hindered transportation, and strained energy grids, leading to significant economic losses and adverse effects on public health and safety (Yang & Liu, 2020). By reliably monitoring and modelling the effects of (potential) interventions such as dykes, adequate sewage systems and water level decisions, floods and droughts can be prevented or acted upon (Yang & Liu, 2020; Chan et al., 2020; Albuquerque et al., 2019).

Second, the findings of this research project could be transferable to different regions in the world. Often, data is missing more frequently in poorer areas of the world where floods and droughts have more detrimental effects due to fragile infrastructure (Arriagada et al., 2021).

1.5 Scope

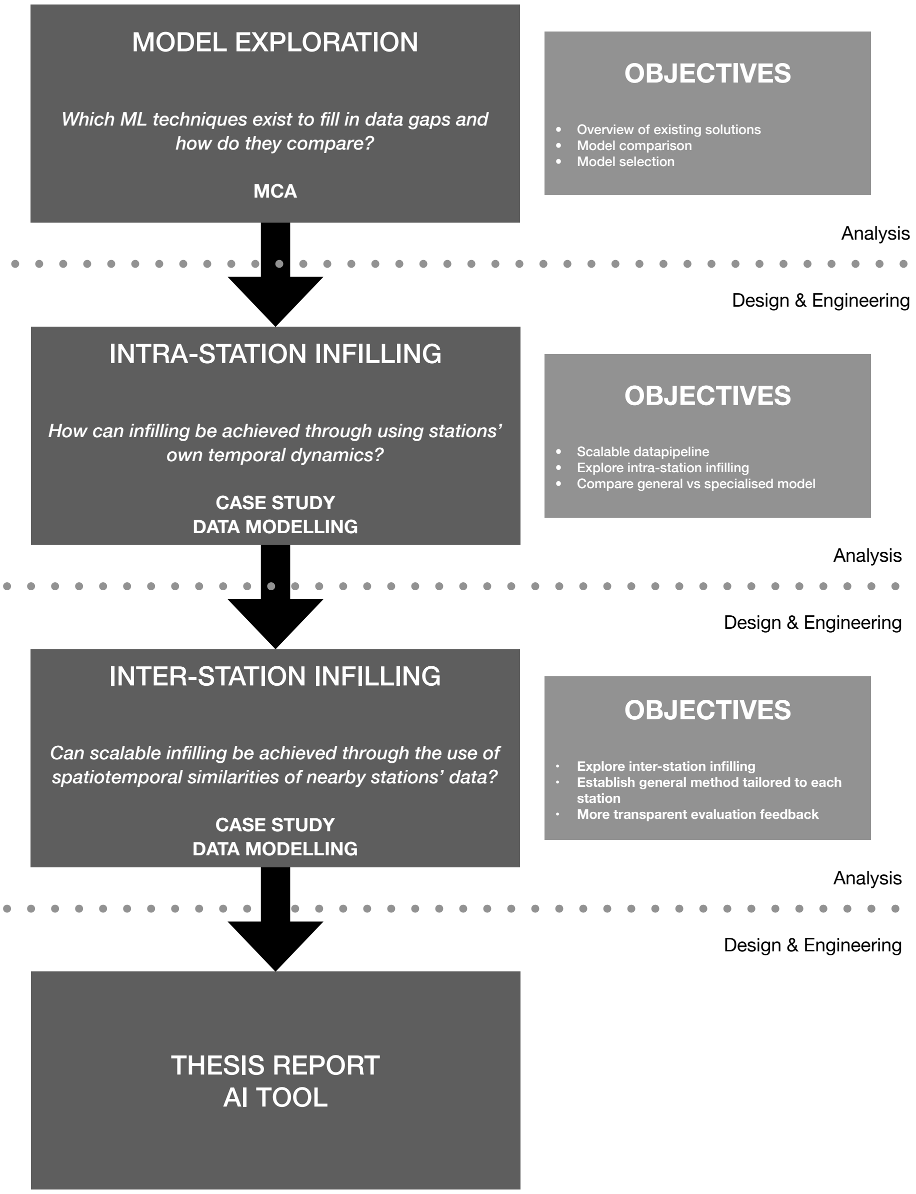
This project focused on a set of nine ML models, found in section 2.3.2. All of these models were subjected to a multicriteria analysis (MCA) to assess their suitability for infilling hydrological time series data. A subset of these models were implemented to infill gaps in water level data sets containing three different types of water: groundwater, sewage water and surface water. These water bodies were located in the centre of Almere, a Dutch city in the province of Flevoland.

The gaps under investigation varied in length from just six hours to roughly seven months. These gaps were artificially implemented to ensure robust assessment of the models with the ground truth. Furthermore, this exploration focused on both multivariate and univariate infilling in the theoretical part (Chapter 3), but during implementation the focus was on univariate infilling to ensure equal comparison of the models. Additional findings regarding multivariate infilling are included in the discussion of the results.

1.6 Reading Guide

The next chapter provides a literature review and theoretical framework for using ML to infill gaps in hydrological time series data. The report discusses the methodologies, results and discussion for each sub research question separately. First, the exploration and comparison of performance of different ML models used for ML models through a multicriteria analysis (MCA) is presented in Chapter 3. Second, the case study in which intra-station infilling is explored is presented in Chapter 4. Third, the case study in which inter-station infilling is explored and presented in Chapter 5. At last, CH6 and CH7 contain a discussion of the results in relation to the main research question, including limitations that spread across research phases, and the conclusion of this thesis, synthesising the findings of all research phases.

An overview of the thesis project can be seen in Figure 1.



MODEL EXPLORATION

Which ML techniques exist to fill in data gaps and how do they compare?

MCA

OBJECTIVES

- Overview of existing solutions
- Model comparison
- Model selection

Analysis

Design & Engineering

INTRA-STATION INFILLING

How can infilling be achieved through using stations' own temporal dynamics?

CASE STUDY
DATA MODELLING

OBJECTIVES

- Scalable datapipeline
- Explore intra-station infilling
- Compare general vs specialised model

Analysis

Design & Engineering

INTER-STATION INFILLING

Can scalable infilling be achieved through the use of spatiotemporal similarities of nearby stations' data?

CASE STUDY
DATA MODELLING

OBJECTIVES

- **Explore inter-station infilling**
- **Establish general method tailored to each station**
- **More transparent evaluation feedback**

Analysis

Design & Engineering

THESIS REPORT
AI TOOL

Figure 1, an overview of the research project.

2. Literature Review

In this section, relevant concepts will be discussed and a theoretical framework will be developed. First, a description of the hydrological regime is presented to gain a better understanding of the dynamics of the water level. Second, existing statistical infilling solutions and their shortcomings will be discussed. Thereafter, a brief overview of ML concepts will be provided to understand the basics of this field. Additionally, the ML models under investigation are explained. At last, concepts will be synthesised into the ML workflow framework as presented by Oakes et al. (2022).

2.1 Hydrologic Regimes

It is important to understand the dynamics of water level because the goal of this project was to accurately infill water level time series. Water level is a component of the hydrologic regime. Hydrologic regimes can be defined as the temporal and spatial relationship between precipitation and discharge of a watershed (Post & Jones, 2001). The water level regime can be defined as the frequency, timing and duration of changes in water level (Leira & Contonati, 2008). An overview of the various factors influencing the hydrologic regime can be seen in Figure 2.

Water level is influenced by natural and anthropogenic factors. Natural factors include meteorological variabilities such as precipitation, temperature, and evapotranspiration (McGregor, 2019; Nowak & Ptak, 2019; Devia et al., 2015), and geographic factors such as geology, vegetation and land use (McGregor, 2019; Nowak & Ptak, 2019). Anthropogenic factors vary from interventions stemming from industrial and agricultural activities and urbanisation. Some examples are irrigation, energy production and flood protection (Mackay et al., 2014; Now & Ptak, 2019; Devia et al., 2015). The addition of these anthropogenic factors to the natural influences increases the complexity of infilling data gaps (Harvey et al., 2012; Arriagada et al., 2021).

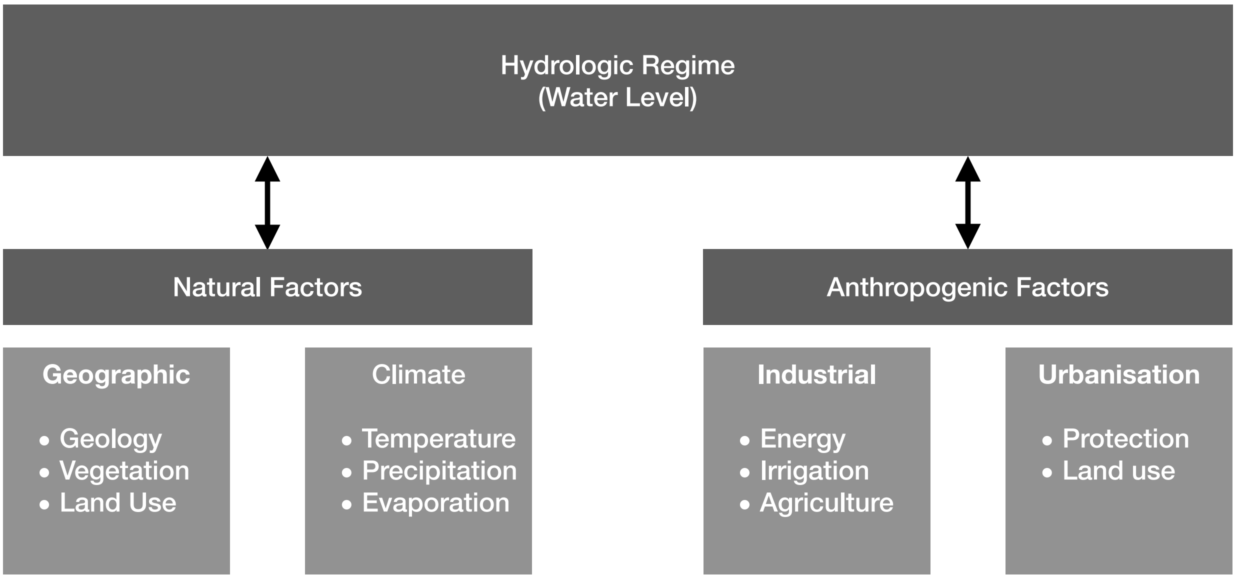


Figure 2, a simplified overview of the hydrologic regime, own visualisation.

2.2 Existing Infilling Methods

To address the problem of data gaps, several infilling methods have been applied and explored. Below, the statistical methods are briefly described along with the different systematic approaches used to apply these methods.

2.2.1 Statistical Methods

A wide range of statistical methods have already been explored for infilling hydrological time series. These methods range from manual observations to more mathematical infilling solutions. Harvey et al. (2012) classified several infilling methods for hydrological data, as presented in Table 1. Infilling methods have various limitations and may not always perform well due to assumptions about the distribution of missing values. Hence, understanding the pattern and mechanism of missingness is crucial before choosing an infilling method (Tabachnick & Fidell, 2014).

With increasing gap length and multivariate gaps, uncertainty rises. This affects the effectiveness of these statistical methods (He et al., 2020). Furthermore, the missing values often are not random but often occur in continuous blocks, making statistical infilling methods useless (He et al., 2020). Consequently, these methods are not scalable to bigger and multivariate gaps.

Table 1, an overview of different infilling methods (Harvey et al., 2012).

Infilling Methods	Explanation
Manual Inference	Estimation of flows through a visual comparison
Serial Interpolation	Linear, polynomial or spline interpolation
Scaling Factors	Using a water body's specific characteristics and the differences and ratios of these variables between different measuring stations
Equipercentile Techniques	Determining the percentile of different measurement station values and then converting them into target flow values based on existing data
Linear Regression	A regression formula for the missing measurement is derived based upon at least one different measurement station
Hydrological Modelling	Methods that are more black-box approaches

2.2.2 Systematic Methods

There are several systematic methods that have been attempted to infill gaps in hydrologic time series (Janbain et al., 2023). Each systematic method has a distinct scope based on the data system it uses.

The first method is the intra-station system. Through this approach, gaps are filled using a model trained solely on data from the specific measurement station where the time series is recorded (Janbain et al., 2023). This method avoids reliance on external data sources (Janbain et al., 2023; Song et al., 2020; Sarafanov et al., 2022), thereby accounting for the unique hydrological regime at the station's location. This represents a very closed system approach.

The second method is the inter-station system. Here, values for target stations are estimated using data from nearby stations with similar hydrological regimes (Janbain et al., 2023; Dastoni et al., 2010; Taie Semiromi & Koch, 2019). Due to their proximity, these stations are likely to share comparable hydrological conditions, taking into account local natural and man-made factors (McGregor, 2019).

A third approach is more holistic, taking into account the hydrologic regime. It involves estimating the target variable based on its relationship with external variables such as precipitation and evaporation (Janbain et al., 2023). This method heavily relies on the hydrological regime to select relevant features and has been explored by a diverse range of researchers (Kim et al., 2015; Contractor & Roughan, 2018; Ren et al., 2022).

Example in Sammerspolder, The Netherlands.
In the Sammerspolder example, the gap could be filled by using only data from Sammerspolder (Figure B) or with data from stations nearby (Figure C).

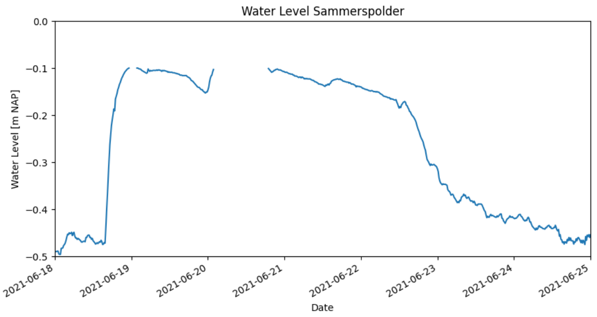


Figure B, water level in Sammerspolder.

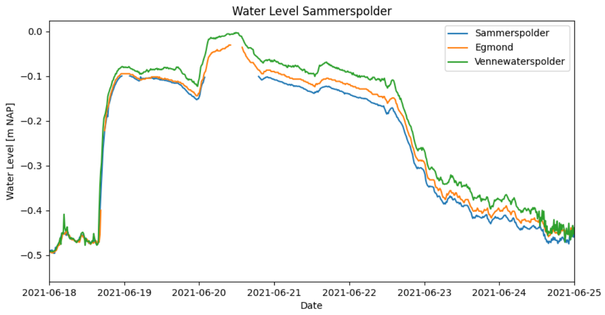


Figure C, water level at Sammerspolder and nearby stations in Egmond and Vennewaterspolder.

2.3 Machine Learning

ML models could be classified under the hydrological modelling approaches mentioned in Table 1, since they tend to be black-box methods. Indeed, ML modes are not implicitly programmed and thus difficult to understand (James et al., 2013). ML is well-suited for infilling hydrological time series data due to its flexibility and ability to handle complex relationships inherent in hydrological datasets (Hamza et al., 2020; Kazijevs & Samad, 2023; Tang & Ishwaran, 2017).

In this section, the basics of ML will be explained in section 2.3.1. In section 2.3.2, an explanation and overview of the ML techniques under investigation is presented. At last, in section 2.3.3, a framework for multidisciplinary ML projects is provided.

2.3.1 Machine Learning Basics

ML focuses on developing algorithms capable of learning from and making predictions or decisions based on data (James et al., 2013). These algorithms can perform classification or regression tasks. Classification concerns the categorisation of data into classes, whereas regression tasks concern continuous outcomes (James et al., 2013). These predictions are done based on a set of features: independent variables that help to estimate the dependent variable (James et al., 2013).

ML models have several characteristics. First, they can be classified into supervised or unsupervised models. Supervised models learn based on a set of inputs and corresponding outputs, or labelled data. In contrast, unsupervised models deal with unlabelled data and try to uncover patterns, such as clustering data into groups (James et al., 2013). Second, models can be parametric or nonparametric. A parametric model uses a mathematical formula to establish its prediction, whereas a nonparametric model does not (James et al., 2013). Due to these differences between models, some are more suitable for certain problems than others.

Table 2 provides an overview of ML model characteristics.

Table 2, an overview of ML model characteristics.

Learning	**Inference**	**Task**
Supervised: training output is know	Parametric: use of mathematical formula	Classification: predict categorical output
Unsupervised: training output is unknown	Nonparametric: no use of mathematical formula	Regression: predict numerical output

2.3.2 Infilling Hydrological Time Series with Machine Learning

A wide range of ML models has been researched already. In this section, the models under investigation in this thesis will be explained and their use cases are described briefly. A deeper inquiry into the potential of different ML models is provided in Chapter 3, including an assessment on their suitability for infilling water level through an MCA.

Multiple Linear Regression (MLR)

An MLR model is a supervised parametric model primarily used for regression tasks. It fits a linear equation to the observed data by minimising the sum of the residuals, which are the differences between the observed and predicted values. The goal is to find the best-fitting line that explains the relationship between the dependent variable and multiple independent variables. Once the model is trained, it can estimate outputs by plugging the input values into this linear equation. This allows for the prediction of continuous outcomes based on the input data (James et al., 2013). A visualisation of MLR can be seen in Figure 3.

MLR models have been explored for infilling streamflow data (Hamzah et al., 2020), precipitation data (Portuguez-Mauratua et al., 2022), and temperature data (Körner et al., 2018).

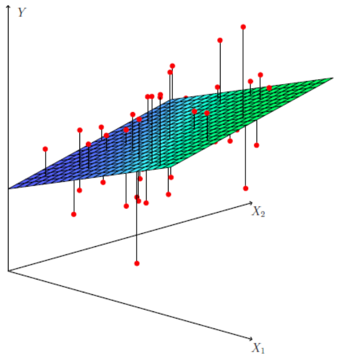


Figure 3, a visualisation of the MLR model (James et al., 2013).

Support Vector Regressor (SVR)

SVRs are supervised parametric models that rely on a mathematical equation for their output. It is slightly similar to MLR but can handle more complexities. This is enabled by their use of kernels, which are mathematical operations to handle non-linearities. Like other models, SVRs aim to minimise the residual between the training points and its mathematical equation (James et al., 2013). A visualisation can be seen in Figure 4.

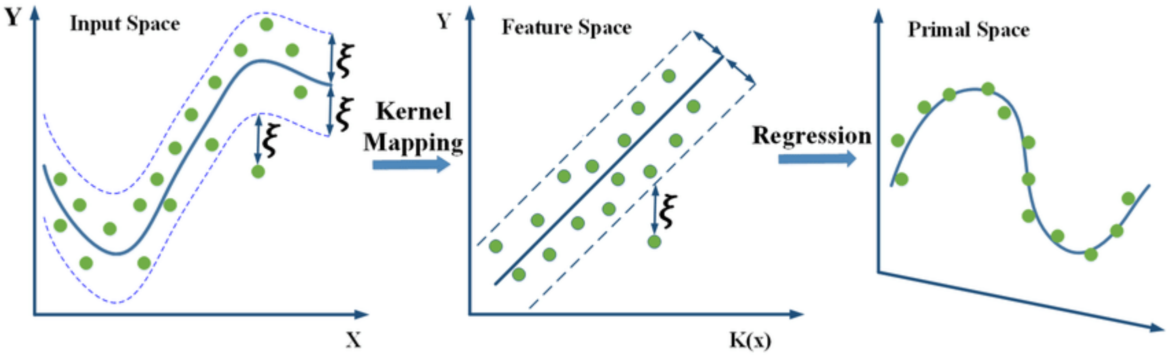


Figure 4, a visualisation of the SVR model (Moradzadeh et al., 2020).

SVR models have been used by He et al. (2020) for infilling groundwater level data, and by Dahmani & Latif (2024) for streamflow data infilling.

k-Nearest Neighbour (kNN)

A kNN model is a supervised non-parametric model that can be used for both regression and classification tasks. The model predicts its output by calculating the Euclidean distance between the input data and its training data. In regression, it uses the mean of the *k*-nearest data points for its prediction (James et al., 2013). Figure 5 presents a visualisation of the kNN model.

These models have been used in infilling groundwater level data by He et al. (2020). Additionally, Hamzah et al. (2020) used a kNN model for infilling streamflow data. Several researchers have used kNN for infilling meteorological data such as precipitation (Portuguez-Mauratua et al., 2022; Sahoo & Ghose, 2022), and temperature data (Sharma & Yuden, 2021).

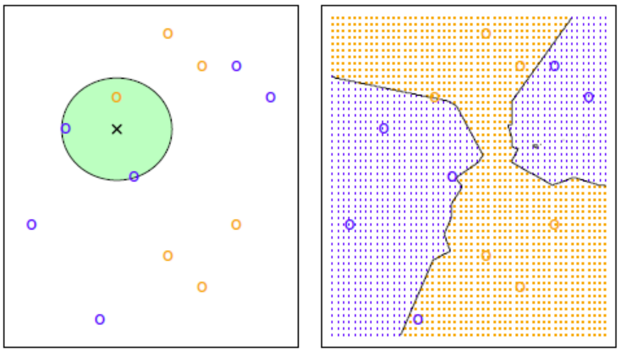


Figure 5, a visualisation of the kNN model (James et al., 2013).

Tree-Based Models

Tree-based models are supervised non-parametric models. They use decision trees instead of mathematical equations for their predictions. For regression tasks, the training set is recursively split into two subsets with the aim to minimise the variance or another criteria (MSE, RMSE, etc.) at each step (James et al., 2013). The output of tree-based models is then based on the resulting decision tree(s). The process is visualised in Figure 6.

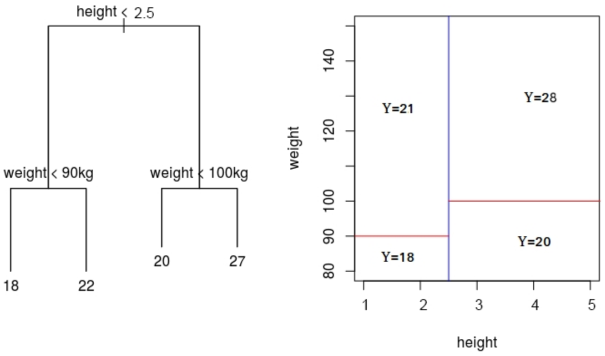


Figure 6, a visualisation of how decision trees are made (Bhatnagar, 2019).

Random Forest (RF)

Random Forest (RF) models are a random ensemble of decision trees, with each tree focusing on a different subset of the data and its features. The output represents the average of the ensemble (James et al., 2013).

RFs have been researched to infill precipitation data (Portuguez-Mauratua et al., 2022), streamflow data (Hamza et al., 2020) and univariate and multivariate infilling of water level data (Umar & Gray, 2023).

MissForest

MissForest is a variant of RF. In a MissForest model, a RF model is constructed for each missing variable. Next, the missing values are estimated iteratively until the estimations converge (Stekhoven & Bühlmann, 2012).

The MissForest model has been explored for infilling streamflow data by Arriagada et al. (2021), water level data (Umar & Gray, 2023) and also for precipitation and temperature data (Sharma & Yuden, 2021).

Gradient Boosting Trees (GBT)

Gradient Boosting Trees (GBTs) are an non-random ensemble of decision trees as a tree is added to minimise training error during training. Afterwards, the mean of the outputs of the decision trees is its output (James et al., 2013).

GBTs have been used to infill precipitation data (Portuguez-Mauratua et al., 2022) and other meteorological variables such as humidity, wind speed and temperature (Körner et al., 2018).

Artificial Neural Networks

Artificial Neural Networks (ANNs) are parametric models made up of interconnected nodes. Each connection between nodes has a weight, and each node uses an activation function to determine its output. Based on the information provided by previous nodes, these activation functions calculate the value that the node needs to propagate to connected nodes. Together, these weights and functions create a set of mathematical functions that take input data and produce an output (James et al., 2013).

Multi-Layer Perceptrons (MLP)

Multi-Layer Perceptrons (MLPs) are a supervised network of nodes consisting of an input layer, a configurable number of hidden layers and nodes, and an output layer (James et al., 2013). As the problem becomes more complex, additional hidden layers and nodes can be added. As a result, a more complex mathematical equation is established to capture the relation between input and output (James et al., 2013). Figure 7 shows a visual representation of a MLP.

MLPs have been researched for infilling streamflow data (Coutinho et al., 2018; Aghelpour & Varshavian, 2020; Kim et al., 2015) and a wide range of meteorological data (Sharma & Yuden, 2021; Sahoo & Ghose, 2022; Nkiaka et al., 2016).

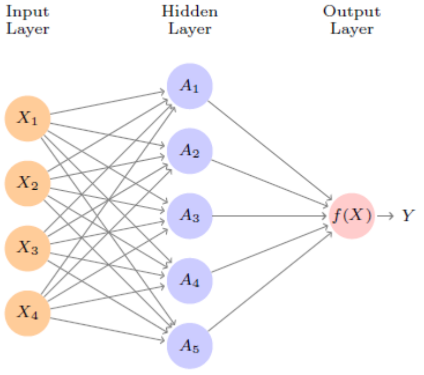


Figure 7, a visualisation of an MPL (James et al., 2013).

Long Short Term Memory (LSTM)

Long-Short Term Memory (LSTM) are more complex supervised networks that are able to memorise data. The LSTM has three different gates: input, forget and output. The input gate processes the input, the forget gate determines whether its memory should be updated or that information should be forgotten, and an output gate makes the prediction. As a result, LSTMs can learn about trends and temporal dynamics very well (van Houdt et al., 2020). The LSTM block is visualised in Figure 8.

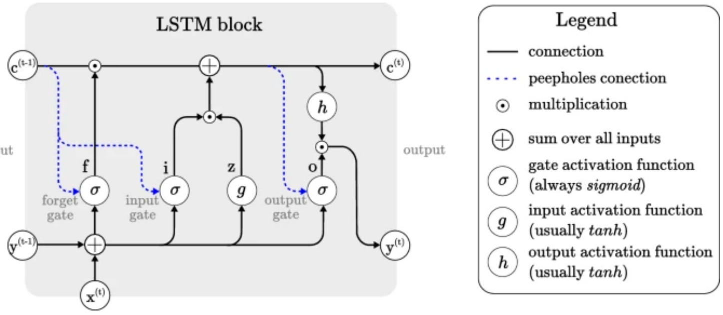


Figure 8, a visualisation of an LSTM cell (van Houdt et al., 2020).

LSTMs have been used for infilling water level data (Janbain et al., 2023), streamflow data (Ren et al., 2022), and also for anomaly detection (Kulanuwat et al., 2021).

Self-Organising Map (SOM)

Self-Organised Maps (SOMs) are unsupervised grids of neurons, thereby reducing the dimension of the data. These neurons contain a vector with information. During training, for each data point, the distance to these neurons is calculated. The data point is assigned to the closest neuron, the winning neuron or best matching unit (BMU), and the vector of this neuron is adjusted to better fit the input data. To make predictions, new data points are mapped onto the SOM grid by finding the closest neuron. The prediction is then made using the value associated with the position of the target variable's vector on this node (Nanda et al., 2017; Sahoo & Ghose, 2022). Figure 9 contains a visualisation of the SOM and how it is used for estimation.

SOMs have been by Kim et al. (2015) and Hamza et al. (2020) for infilling streamflow data, by Nanda et al. (2017) and Umar & Gray (2023) for infilling water level data. Several researchers also used SOMs for meteorological infilling (Sahoo & Ghose, 2022; Nanda et al., 2017; Nkiaka et al., 2016).

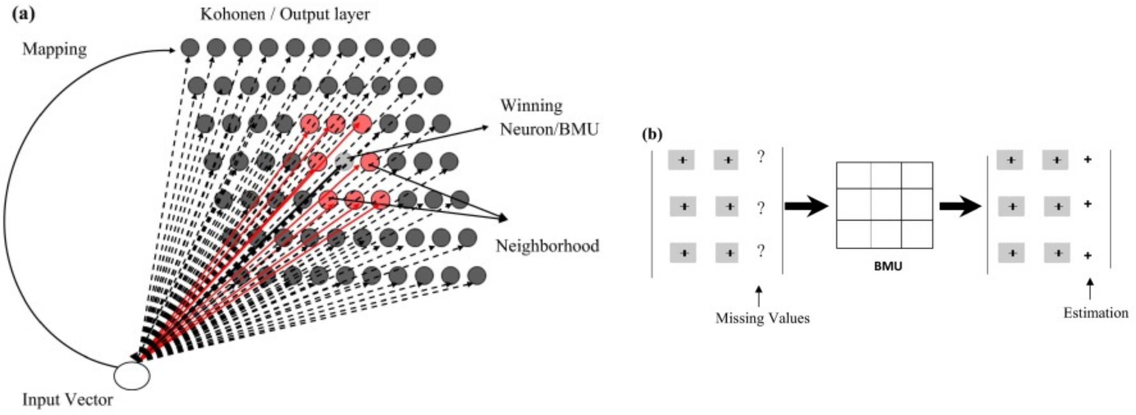


Figure 9, a visualisation of a SOM and how inference works for SOMs (Nanda et al., 2017).

A wide number of ML models have thus been researched for infilling purposes. However, these studies primarily focused on evaluating different ML models using a single systematic approach (see section 2.2.2). Additionally, these studies were limited to infilling data for a single water type or body. Consequently, the generalisation capabilities of the ML models across various contexts remain largely unexplored. For instance, it is unclear whether certain models demonstrate superior performance when applied to intra-station (within the same station) versus inter-station (across different stations) scenarios. Moreover, the effectiveness of different systematic approaches may vary depending on the type of hydrological regime, but this aspect has not been thoroughly investigated. Therefore, there is a significant knowledge gap regarding the adaptability and robustness of ML models across diverse hydrological conditions and the comparative efficacy of systematic approaches tailored to different hydrological regimes.

2.3.3 Machine Learning Workflow

ML can be used for a wide variety of tasks and is thus a very multidisciplinary domain. In 2022, Oakes et al. presented a comprehensive framework for multidisciplinary ML projects that aim to solve domain specific problems. The framework consists of three layers: problem, solution and implementation. Each layer consists of four regions: general, domain-specific, ML and the multidisciplinary region. These layers and regions do not have strong boundaries, as there is a lot of interaction between both the layers and the regions, making it a multidisciplinary and non-linear process. A visualisation can be seen in Figure 10.

Problem Space

The problem space layer concerns the problem formulation. Here, domain-specific knowledge and theory can be translated into ML concepts such as potential features, relations, and relevant models (Oakes et al., 2022). This then translates into a mix of both problems in the multidisciplinary region.

Hence, the domain specific problem in this project is scalable infilling hydrological time series datasets. Furthermore, it can be noted that the ML problem under investigation here is a regression problem, and that a supervised learning approach works best here. Consequently, the multidisciplinary problem of this research project is to compute missing values for water level data to infill incomplete hydrological time series datasets.

Solution Workflow Space

Next, the solution space layer concerns the workflows that are developed to solve the problem. Here, the development, evaluation and comparison of several ML models occurs (Oakes et al., 2022). The domain-specific workflow concerns the use of industry standards and workflows (Oakes et al., 2022). The specifics of hydrological workflow standards will not be discussed to avoid too much detail.

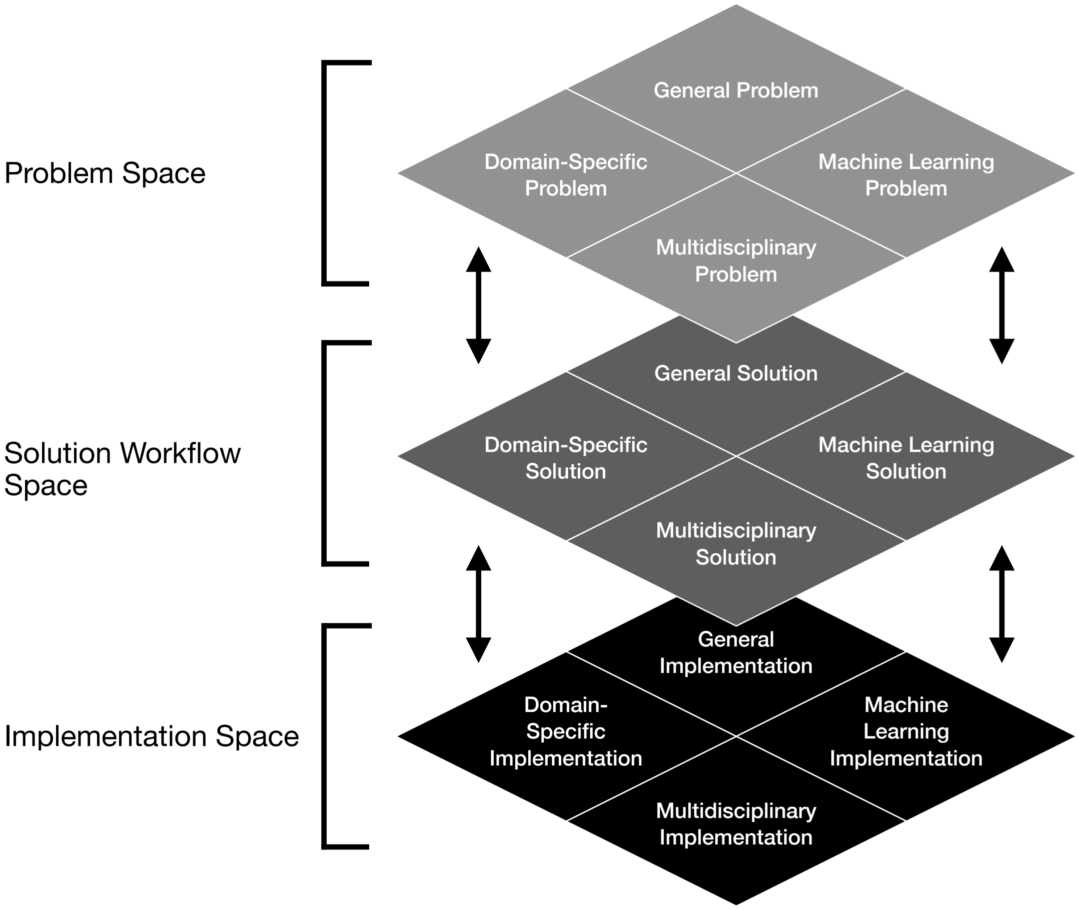


Figure 10, a visualisation of the ML workflow as described by Oakes et al. (2022).

Data Preprocessing & Feature Selection

The development of an ML model requires several steps (James et al., 2013; Oakes et al., 2022). First, the relevant data needs to be acquired. Then, the data needs to be preprocessed, meaning it is cleaned and formatted correctly (Keller et al., 2020; Oha et al., 2020; Oakes et al., 2022). Thereafter, it is important to explore the data. This concerns the different features (independent variables), potential missing data and the relationships between features. Next, it is important to select the features that are most likely to produce accurate outputs. Features can also be engineered by using existing features to derive new features (Dong & Liu, 2018). After preparing the data, the set is split into different subsets for training, validation and testing purposes.

Model Selection

Model selection is also an important part of ML. This concerns the exploration of the best hyperparameter set for a model. Hyperparameters are the configurable settings for an ML model, such as the number of neighbours in a KNN, the number of trees of a RF and the number of hidden layers in an MLP (Raschka, 2018; James et al., 2013). During hyperparameter tuning, a set of models with different hyperparameter settings are trained on the training data. After training, the models are evaluated and selected by assessing their performance on the validation and test set (James et al., 2013; Oakes et al., 2022; Raschka, 2018).

Model Evaluation

It is crucial that evaluation is done with unbiased data. Therefore, validation and test datasets are often used to assess the model's accuracy. By leaving the validation and test datasets out during training, the model is assessed on its accuracy using data it has not seen yet. As a result, using validation and test sets give a more representative understanding of the model's performance (James et al., 2013; Raschka, 2018; Tatachar, 2021).

The performance of ML models can be evaluated and compared in several ways, focusing on computational performance and prediction accuracy (Raschka, 2018). Computational performance of ML models is often measured in training time, inference time, memory size, and CPU/GPU utilisation, among others (Pykes, 2023). The prediction accuracy of regression models can be evaluated using the mean square error (MSE) and root mean square error (RMSE) or by the R-squared (R^2) and adjusted R^2 (Tatachar, 2021; James et al., 2013). Both MSE and RMSE calculate the average difference between the predicted value and the observed value. Both R^2 metrics indicate how well the data fits the regression model by taking into account the variance of both predictions and the data. Adjusted R2 takes into account the number of features used for predictions (James et al., 2013).

In hydrology, reliability of models can be evaluated using the Kling-Gupta efficiency (KGE) and Nash Sutcliff efficiency (NSE) (Knoben et al., 2019). KGE metric accounts for spatial variability, normalised variance, and Pearson's correlation. It ranges between negative values and 1, in which a higher KGE indicates strong accuracy (Knoben et al., 2019). NSE takes into account the mean squared error of observed and referenced data. It ranges from negative values to one, which indicates a perfect result (Knoben et al., 2019).

In Table 3, an overview of relevant evaluation metrics can be found.

Table 3, an overview of evaluation metrics for ML and hydrological models.

Metric	Explanation
MSE / RMSE	(Root) of the mean residual between a prediction and observation of test set.
R^2 / Adjusted R^2	Measures the proportion of variance in the dependent variable explained by the independent variables
NSE	Assesses predictive skill by comparing model predictions to observed data mean,
KGE	Evaluates model accuracy considering correlation, bias, and variability, with values closer to 1 being better

This ML workflow is a highly iterative process and can follow several cycles (Oakes et al., 2022). An overview of the ML workflow can be seen in Figure 11.

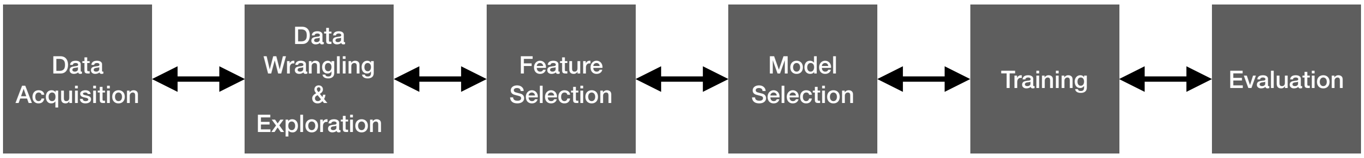


Figure 11, an overview of the ML pipeline, own visualisation.

Implementation Space

At last, the implementation space concerns the implementation of the solution. For example, the ML model is added to the operations of an organisation (Oakes et al., 2022). Usually, this is initially done at a small scale through a proof of concept (Aho et al., 2020). It is important that the implemented solution is monitored to assess its performance. This monitoring concerns the data fed to the model, the model's accuracy and its operational performance over time (Pykes, 2023).

For this project, the desired domain-specific implementation is a script to complete hydrologic datasets for hydrologic modelling. Hydrological models help to understand and predict the complex hydrological processes, albeit that these models only partly describe reality (Wagener et al., 2001; Devia et al., 2015). Water level data serves as input and calibration tool for these models (Wagener et al., 2001; Devia et al., 2015; Arriagada et al., 2021).

The relevant ML implementation is a model that accurately predicts the missing values of hydrologic time series data. Consequently, the multidisciplinary implementation concerns an AI tool for infilling hydrological time series datasets.

All of these concepts can be synthesised in a theoretical framework as shown in Figure 12.

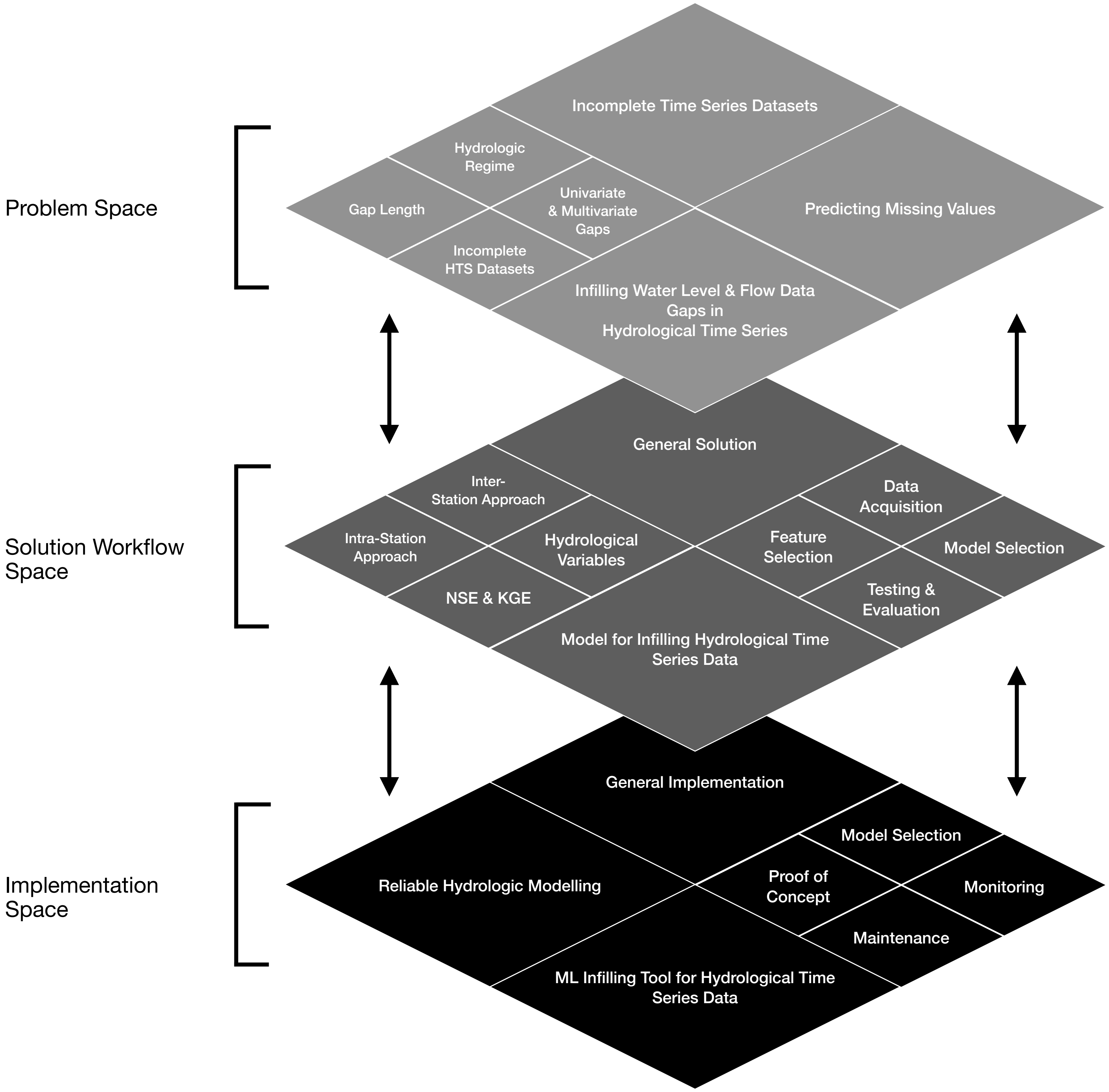


Figure 12, a visualisation of the TF used for this thesis project, based on Oakes et al. (2022) conceptual framework for ML workflows.

3. ML Model Exploration & Analysis

There are a wide range of ML models that could be used for infilling hydrological time series data. However, most ML models are slight deviations from a few foundational architectures. Additionally, since there is limited time to try and build all these models, it is important to compare and assess these models to select a subset for the next stages. Therefore, an MCA was executed on a selected list of widely researched models that also serve as foundations for other models. Through this MCA, models were compared and assessed on their suitability for practical implementation.

In this chapter, the first phase of this thesis project is discussed, addressing the first research question: *what ML techniques are most suitable for infilling water level time series data?* First, the methodology section (3.1) begins by explaining the steps and reasoning behind the MCA process used to answer this question. Following this, the results section (3.2) presents the findings from the MCA, providing a detailed comparison of the techniques. Finally, the chapter concludes with a discussion (section 3.3) of these results, synthesising the findings to provide a comprehensive answer to the research question.

3.1 Methodology

To explore which ML techniques have been implemented to infill (hydrological) time series data gaps, the following steps were taken. First, a list of the different techniques was established. This list was developed with the help of Google Scholar, ChatGPT and Perplexity. An overview of used search terms is presented in Table 4.

Due to their stochastic nature and potential hallucinations (Yao et al., 2023), both ChatGPT and Perplexity are used as supportive tools only, never as leading sources. A list with links to the relevant conversation with these AI models can be found in Appendix B.

Table 4, an overview of potential search terms for the literature review.

AI infilling time series data	hydrological time series data infilling	ML infilling methods time series	hydrological time series ML infilling
ML imputation methods time series	hydrological time series imputation methods	time series infilling methods	imputation methods for incomplete time series data

After establishing a comprehensive list of nine different ML techniques, mentioned in section 2.3.2, more in-depth research was done into each ML technique for a MCA. This MCA was performed to choose the techniques that would be the focus of this project. The MCA method was chosen because it is a helpful tool for comparing and ranking machine learning algorithms (Ali et al., 2017).

3.1.1 MCA Criteria

Four criteria were used for the MCA. First, the *accuracy* of the models was assessed. This takes into account how well the models predicted missing values and infilled gaps. This is important because the gaps need to be infilled with great accuracy, improving the reliability of modelling and monitoring. This criterion was assessed by looking at evaluation metrics such as (R)MSE, R2, NSE, and KGE obtained by other researchers.

Second, the models' *scalability* was assessed. The meaning of scalability here is threefold: how well accuracy holds up as gap length increases, its ability to handle multivariate gaps and the potential to capture complex dynamics. This criterion was chosen because it is important that the ML models perform adequately in a wide range of circumstances (i.e. large and small gaps, univariate and multivariate gaps). Scalability was assessed by looking at accuracy degradation, the ability to process missing input variables and the degree of (mathematical) complexity the model is potentially able to capture with its inner workings.

Third, *data requirements* for training a model were taken into account. Different models require different amounts of data and features to work properly (James et al., 2013). Therefore, this criterion is chosen because data might be scarce in some situations, leading to suboptimal predictions for some models. Data requirements are assessed through the amount of training data and whether the model is able to perform well with a limited amount of features.

Computational load is the last criterion, taking into account the time and resources needed by a model to train itself and for inference, the processing of input data and making predictions. ML models have different underlying methods, some rely on complex mathematical operations, others on simpler mechanisms such as decision trees (James et al., 2013). It is useful to take this into account, since computation can be costly (Pykes et al., 2023). Indicators of this criterion are training and inference time and also the processes used for training and inference.

3.1.2 MCA Weights & Scores

The different criteria were assigned different weights to take into account their relative importance. Since accuracy is most important for infilling time series data, it received a weight of 40%. Scalability received a weight of 30% because the AI tool should work with long and multivariate gaps. The remaining 30% was divided equally among data requirements and computational load as these are essential but secondary concerns.

After assessing each technique, the obtained scores were standardised on a scale of zero to one. Then, the standardised scores were added up in accordance with their weight to obtain a total score. Next, the ML techniques were ranked based on this total score. An overview of the criteria, their score range, weights and indicators can be seen in Table 5.

Table 5, an overview of the criteria used in the MCA.

Criterium	Score Range	Weight	Cost / Benefit	Indicators
Accuracy	0 - 1	40%	Benefit	R^2, NSE, KGE, RMSE, etc.
Scalability	Very Low - Very High	30%	Benefit	Gap Length, Complexity, etc.
Data Requirements	Very Low - Very High	15%	Cost	Training Data, Number of Features, etc.
Computational Requirements	Very Low - Very High	15%	Cost	Training Time, Inference Time, etc.

3.1.3 Sensitivity Analysis

Afterwards, a sensitivity analysis was performed to assess the robustness of the MCA results. In this analysis, the weights of the criteria were adjusted. These adjustments can be seen in Table 6.

Table 6, an overview of the weight changes for the sensitivity analysis.

Criterium	Original Weight	Weight A	Weight B
Accuracy	40%	25%	30%
Scalability	30%	25%	30%
Data Requirements	15%	25%	20%
Computational Requirements	15%	25%	20%

3.1.3 Limitations

This methodology suffered from multiple limitations. These limitations will be discussed here, so that the reader knows how to interpret the results up front. First, the list of nine models was not an exhaustive list. Due to the enormous amount of ML models, this would be too time consuming. However, the nine models were chosen because they appeared most often during the literature review. Additionally, some of these models are often seen as the foundations of other ML models, of which different variations exist. This can be observed by the close resemblance of RF, GBT and MissForest and also for MLP, SOM and LSTM. However, the validity of this study is compromised by this non-exhaustive list.

Second, the literature focused not only on infilling of water level data, but also on infilling performance for other hydrological and meteorological variables. As a result, infilling performance for water level could be different, hurting the validity of the research. Furthermore, the results obtained by the referenced researchers had differences in methods such as training methods (features, model selection, etc.), gap length and evaluation metrics. Some of the methods were unclear and thus some uncertainty arises in the scores.

Third, there is some subjectivity involved in scoring the models for some criteria. This subjectivity results from the absence of hard metrics and indicators. For example, there is no clear distinction between a large amount of data and a moderate amount of data. Additionally, this distinction would be different for problems because of the unique environment of different problems. To counter this subjectivity, and to improve the replicability and reliability of the results, several rounds of validation were done with supervisors and AI tools such as ChatGPT.

Fourth, the criterias' weights suffer from the same subjectivity. To improve the reliability and validity of the results, the same measures were taken: validation with supervisors and AI tools.

3.2 Results

In this section, the results are presented per model. An overview of the results can be found in Table 7 on page 24.

3.2.1 Multiple Linear Regression (MLR)

The use of simple MLR models for infilling time series datasets have not yielded significant results. When evaluating its performances, MLR models tend to score around 0.4-0.5 as R^2 value (Portuguez-Maurtua et al., 2022; Körner et al., 2018). Therefore, MLR scores poorly on accuracy and a *4* is given.

Considering scalability, MLR cannot handle missing variables since this would result in an incomplete formula (James et al., 2013). Contrarily, accuracy seems to be rather constant as gap length increases (Portuguez-Maurtua et al., 2022; Körner et al., 2018). Although MLRs are capable of finding complex relationships through polynomial extension, its ability to detect highly complex relations are limited (James et al., 2013). Hence, MLR scores *low* on scalability.

With regards to data requirements, MLR can be used with a small number of features. However, a small number of features can give dissatisfactory results due to a limited scope of the input data (James et al., 2013). MLR does not require lots of data to perform well, as its parametric character enables the ability to extrapolate. However, this extrapolation power should not be overestimated (James et al., 2013). However, performance will increase with sample size, as is the case with most ML models (Knofczynski & Mundfrom, 2008). As a result, MLR scores *moderate* on data requirements.

MLR scores *moderate* on computational requirements. Complexity of the relationships and the number of variables involved influence the computational loads proportionally, because then the mathematical equations become more complex and time consuming (Körner et al., 2018; James et al., 2013).

3.2.2 Support Vector Regressor (SVR)

Research on the use of SVRs for infilling time series data has resulted in good results. It has scored low (R)MSE scores, while obtaining R^2 scores of roughly 0.8 (He et al., 2020; Dahmani & Latif, 2024; Mounce et al., 2011). Therefore, SVR receives a score of *8* for its accuracy.

SVRs' scalability with regards to gap length has proven to be somewhat robust, although its accuracy decreases as gap length becomes significant (He et al., 2020). However, SVRs are unable to process missing input variables as this affects the underlying mathematical formula (James et al., 2013). These architectures are very flexible and can handle complex relationships between variables (Mounce et al., 2011; James et al., 2013). Therefore, SVR scores *moderately* on scalability.

Furthermore, SVRs are good at handling limited training data due to its ability to avoid over-learning and local minima (Mounce et al., 2011). However, SVRs can be prone to overfitting due to their goal to minimise the residual between training data points and its mathematical equation (James et al., 2013). Therefore, SVR got a score of *low* for its data requirements.

As training data and the non-linearity of the relationships between variables increase, SVRs can be computationally intensive because of their mathematical complexity (Mounce et al., 2011; James et al., 2013). A *high* score is thus obtained by SVR.

3.2.3 k-Nearest Neighbour (kNN)

The implementation of kNN models have resulted in moderate results for infilling time series data, with R^2 ranging between 0.35 and 0.55 (He et al., 2020; Portuguez-Maurtua et al., 2022). kNNs thus score insufficient for accuracy, and a score of *5* is given.

kNNs have demonstrated to be vulnerable in relation to increases in gap size (He et al., 2020; Umar & Gray, 2023). This architecture is not able to handle missing variables as input (James et al., 2013). Furthermore, since the model is non-parametric, it can capture complex relationships between variables that are difficult to capture in mathematical equations. However, in a high dimensional feature space kNN suffers from the curse of dimensionality. This means that as the dimension of the feature space increases, the distance between data points becomes bigger, and the model's accuracy is affected (James et al., 2013). Therefore, kNN's scalability is considered *low*.

Since kNN is non-parametric and uses historical data as input for its estimation, its extrapolation is very limited. It is thus important that the training data sufficiently covers the variance of the data. Hence, as the amount of data needed is proportional to the variance of the data (James et al., 2013). Therefore, kNN scores *moderate* on data requirements.

The computational load of kNNs is proportional to the training data, since the model needs to calculate the distance of an input to every training data point. Especially in situations where there is a lot of variance in the data, and thus a lot of data is needed to cover this variance or optimal performance, computational loads for inference are very high (James et al., 2013; Sahoo & Ghose, 2022). Hence, kNNs obtained a score of *high* for computational requirements.

3.2.4 Tree-Based Models

Random Forest (RF)

RFs, due to their non-parametric character, are capable of learning dynamics that are difficult to capture in mathematical formulas (James et al., 2013). With regards to infilling, RF models have demonstrated acceptable performance as it obtained R^2 scores in the range of 0.6 (Portuguez-Maurtua et al., 2022). Therefore, a score of *6* is awarded to the RF for its accuracy.

As gap size increases, RFs are not able to keep up this good performance (Umar & Gray, 2023). Furthermore, RFs are capable of handling complex relationships and also missing data in features (Hamza et al., 2020; Sharma & Yuden, 2021). This handling of missing data is enabled by the structure of the RF: since each tree focuses on a subset of the training dataset, not all trees are affected by a missing value (James et al., 2013; Umar & Gray, 2023). However, RFs have limited extrapolation capabilities due to their nonparametric nature (James et al., 2013). So, RFs score *high* on scalability.

RFs require a *moderate* amount of data, but as with all ML models, more data is preferred (James et al., 2013). Since RFs are non-parametric, they have limited extrapolation ability. Therefore, the training data should cover the variance of the data sufficiently (James et al., 2013).

RFs can be considered *moderately* computationally demanding, depending on the number of trees and other factors such as tree depth (Sharma & Yuden, 2021; Sahoo & Ghose, 2022; Umar & Gray, 2023).

MissForest

MissForest models have proven to be good infillers as it managed to get R^2 scores of 0.6 (Arriagada et al., 2021; Umar & Gray, 2023). As a result, a score of *6* was awarded for its accuracy.

Since it is a looped version of RF, it shares the same scalability and data characteristics as RF (Arriagada et al., 2021; Stekhoven & Bühlmann, 2012). However, it is more computationally demanding as several RFs are trained (Stekhoven & Bühlmann, 2012). Hence, it scores *high* on computational load.

Gradient Boosting Trees (GBT)

GBTs have shown good infilling performance. It has achieved R^2 and NSE scores of 0.7 (Portuguez-Maurtua et al., 2022; Körner et al., 2018). As a result, it received a score of *7* for its accuracy.

GBT shares the same characteristics as RF when it comes to scalability, data requirements and computational requirements (Portuguez-Maurtua et al., 2022; James et al., 2013).

3.2.5 Artificial Neural Networks

Multi-Layer Perceptron

MLPs have shown to perform very well when it comes to infilling time series data. In various research projects, it has consistently scored R^2 and NSE scores of 0.8, and occasionally scoring 0.9. In addition, it has achieved very low scores for (R)MSE metrics. (Coutinho et al., 2018; Dahmani & Latif, 2024; Aghelpour & Varshavian, 2020; Kim et al., 2015; Sahoo & Ghose, 2022). As a result, MLPs score an *8* on accuracy.

This accuracy deteriorated as gap length increased, although performance was still acceptable (Park et al., 2023). MLPs are not able to handle missing data because this would distort the functions between the nodes and layers (James et al., 2013). Furthermore, MLPs have excellent scalability when it comes to complex problems due to their flexibility: the underlying architecture (nodes, layers, activation functions, etc.) can be tailored to match the complexity of the problem (Aghelpour & Varshavian, 2020; Sahoo & Ghose, 2022). However, MLPs are not able to handle missing variables as this affects the mathematical operations of the network (James et al., 2013). So, MLPs score *high* on scalability.

Neural nets need large amounts of data to allow the weights of the model to converge for optimal predictions. Because these weights are initialised at random, this convergence can take some time. This also depends on the learning rate with which the model updates its weights (James et al., 2013). However, a high or low learning rate can lead to suboptimal results and, therefore, more training data enables a more efficient learning rate (James et al., 2013). However, the network could rely on only a few features (Kim et al., 2015; Sharma & Yuden, 2021; Nkiaka et al., 2016). Thus, MLPs obtained a *high* score for data requirements.

At last, MLP architectures also require significant computational loads, potentially having a long training and inference time (Körner et al., 2018; Nkiaka et al., 2016). This is proportional to the number of layers and nodes. As the network contains more layers and nodes, more connections are established and thus longer mathematical formulas (James et al., 2013). Hence, MLPs got a score of *high* for computational requirements.

Long Short Term Memory (LSTM)

LSTM models have demonstrated exceptionally high performance, with KGEs and NSEs surpassing 0.92 and very low RMSEs (Ren at al., 2022; Janbain et al., 2023). Hence, LSTMs obtained a score of *9* for its accuracy.

These architectures have proven scalable, as they can take into account spatial, temporal and seasonal information (Ren et al., 2022). However, model performance can suffer as gaps become bigger, although this is minimal (Ren et al., 2022; Janbain et al., 2023; Kulanuwat et al., 2021). LSTMs also have difficulty with processing missing data (van Houdt et al., 2020). Hence, LSTMs score *high* on scalability.

As discussed above, ANNs require large amounts of training data, and thus LSTMs also require significant training data, although this can be reduced by cleverly feeding input (Ren et al., 2022; Janbain et al., 2023). As a result, LSTMs received a score of *high* for data requirements.

LSTMs are computationally demanding, due to their complex architecture (van Houdt, 2020). Hence, it obtained a *high* score for its computational load.

Self-Organising Map (SOM)

SOMs have also shown promising results, with high R^2 and NSE scores in the range of 0.6 and 0.9 (Kim et al., 2015; Sahoo & Ghose, 2022; Nkiaka et al., 2016). Therefore, SOMs received a score of *8* for accuracy.

Additionally, SOMs are capable of capturing statistical relationships of high dimensional data into a low dimension with limited information loss (Hamza et al., 2020; Nkiaka et al., 2016). This architecture is also able to handle high variability in data, and thus is scalable (Nanda et al., 2017; Nkiaka et al., 2016). As the gap length increases, SOMs performance is diminishing slowly (Nkiaka et al., 2016). Furthermore, SOMs can handle missing data with slight modifications (Folguera et al., 2015). So, SOMs received a *very high* score on scalability.

As with most ANN architectures, lots of training data is needed (Sahoo & Ghose, 2022; Nkiaka et al., 2016). Due to SOMs limited ability to extrapolate, the amount of training data required is proportional to the variance of the data (Nanda et al., 2017). However, SOMs can also function well with only a few features (Kim et al., 2015; Nkiaka et al., 2016). As a result, SOMs received a *moderate* score for data requirements.

With regards to computational loads, smaller SOMs are not super computationally intensive (Nkiaka et al., 2016; Umar & Gray; 2023). However, as the optimal map size increases with variance of the data, SOMs become very computationally intense because there are more distances to be calculated (Nanda et al., 2017). Therefore, SOMs obtain a score of *high* for computational load.

Table 7, an overview of the MCA results.

ML Technique	Accuracy	Scalability	Data Requirements	Computational Requirements	Total Score
Multiple Linear Regression	4	Moderate	Moderate	Moderate	0.39
Support Vector Regressor	8	High	Low	High	**0.62**
k-Nearest Neighbour	5	Moderate	Low	High	0.39
Random Forest	6	High	Moderate	Moderate	**0.62**
MissForest	6	High	Moderate	High	0.58
Gradient Boosting Trees	7	High	Moderate	Moderate	**0.66**
Multi-Layer Perceptron	8	High	High	High	**0.62**
Long Short Term Memory	9	High	High	High	**0.66**
Self-Organised Maps	8	High	Moderate	High	**0.66**

3.2.6 Sensitivity Analysis

When adjusting the weights for the sensitivity analysis, a few changes arise, as seen in Table 8. What is interesting is that the RF architecture comes in as fifth best in both of these analyses. When all criteria are all equally weighted, this is at the expense of the MLP architecture. When a little more importance is given to data and computational requirements, the SOM architecture scores lower than RF.

Table 8, an overview of the results for the sensitivity analysis (Bold = top 5 result).

ML Technique	Original Score	Score Analysis A	Score Analysis B
Multiple Linear Regression	0.39	0.41	0.4
Support Vector Regressor	**0.62**	**0.58**	**0.59**
k-Nearest Neighbour	0.39	0.38	0.38
Random Forest	0.62	**0.59**	**0.61**
MissForest	0.58	0.53	0.56
Gradient Boosting Trees	**0.66**	**0.61**	**0.64**
Multi-Layer Perceptron	**0.62**	0.51	0.57
Long Short Term Memory	**0.66**	**0.54**	**0.60**
Self-Organised Maps	**0.66**	**0.58**	**0.59**

3.3 Discussion

The aim of this chapter was to explore and analyse a set of ML models that have been researched for infilling hydrological time series. From this exploration, several remarks can be made regarding suitability of the models for infilling hydrological time series and how they compare.

3.3.1 General Scores

First, it is no surprise that the simplest models, MLR and kNN, score lowest. Their simple architecture is not enough to capture the complex dynamics of a hydrological regime. Second, it is not surprising that the models based on decision trees score roughly similar and adequately. Due to the complexity of the hydrological process, it is difficult to capture these dynamics into a mathematical formula. The non-parametric nature of decision tree-based models enables these models to capture the complex dynamics of the hydrological regime. Third, the high scores or the neural network based models is also not surprising. Due to their complex architecture of nodes and layers, these models can be adapted to the complexity of the problem of infilling water level time series.

3.3.2 Model Similarities & Differences

Some models have similar foundations, but slight differences in methods made a big difference in score. First, MLR and SVR both use equations that minimise the residuals between the training points, yet their scores differ notably. This difference can be explained by the increased flexibility of SVR, which uses kernels to create non-linear equations. In contrast, even with polynomial extensions, MLR cannot achieve this level of flexibility.

Next, kNN and SOM both consider the Euclidean distance between observations. However, their scores also differ significantly for several reasons. Firstly, SOMs do not suffer from the curse of dimensionality because they can map high-dimensional data onto a lower-dimensional grid. Secondly, SOMs are less prone to noise, thanks to their iterative training process that averages out noise. On the contrary, kNNs are more susceptible to noisy neighbours, which can be outliers. As a result, SOMs outperform kNNs in infilling hydrological time series data.

When comparing the models that can be grouped together, the subtle differences in score are also explainable. The small differences in scores or the tree-based models can be attributed to their training method. Indeed, for GBTs the trees are added to the ensemble to improve its performance on the test set. On the contrary, for RFs the trees are added at random, regardless of their effect on the performance of the model. Furthermore, the slightly lower score for MissForest is mostly due to its computational intensity, resulting from being a set of RF algorithms.

Next, despite their different use of neural nets, the MLP, SOM and LSTM perform very similarly. Here, it is also not surprising that LSTM scores highest on accuracy, given its recurrent foundation that allows it to remember and learn trends. However, it is somewhat surprising that SOMs outperform MLPs, given the limited extrapolation power of SOMs. This outperformance can be attributed to the lower data requirements for SOMs.

3.3.3 Implications

The six highest scoring models were selected for further scrutiny and practical implementation during the first case study: SVR, RF, GBT, MLP, SOM and LSTM. Despite its similar scores, MissForest was not selected because of its high similarity to RFs. Therefore, selecting it as a seventh model seemed redundant.

4. Case Study 1 - Intra-Station Infilling

In this chapter, the case study to answer the second research question is presented: *how effective are the models in infilling based on the use of intra-station data?* For this first case study, the first objective was to create a data pipeline that could be used throughout the project. This would lay the foundation for the AI tool, as it provided a scalable method for preprocessing the raw data. Second, the objective was to explore the intra-station approach. Due to its independence on other data sources, this approach seemed most suited for a robust AI tool. The third goal was to see whether a general model (a model trained on data of several stations) would outperform a specific model (a model trained on one station's data). This would help to decide the approach for the AI tool. A last objective was to compare how the best models of the MCA performed in practice. As it was not sustainable and time efficient to keep developing a wide range of models, this helped to further shave down the number of models.

This chapter is structured as follows. First, the data and methods section (4.1) explains the steps taken during the implementation of the models in this first case study. Second, the infilling results will be presented and analysed in the results section (4.2). At last, a clear answer to the research question and some commentary on the objectives are presented in the discussion section (4.3).

4.1 Data & Methods

4.1.1 Study Area

The study area of the first case study was Kruidenwijk and Waterwijk, two adjacent neighbourhoods located in Almere, The Netherlands. The area is relatively heterogeneous in its land use. It contains a large built environment and some greener areas, as can be seen in Figure 13.

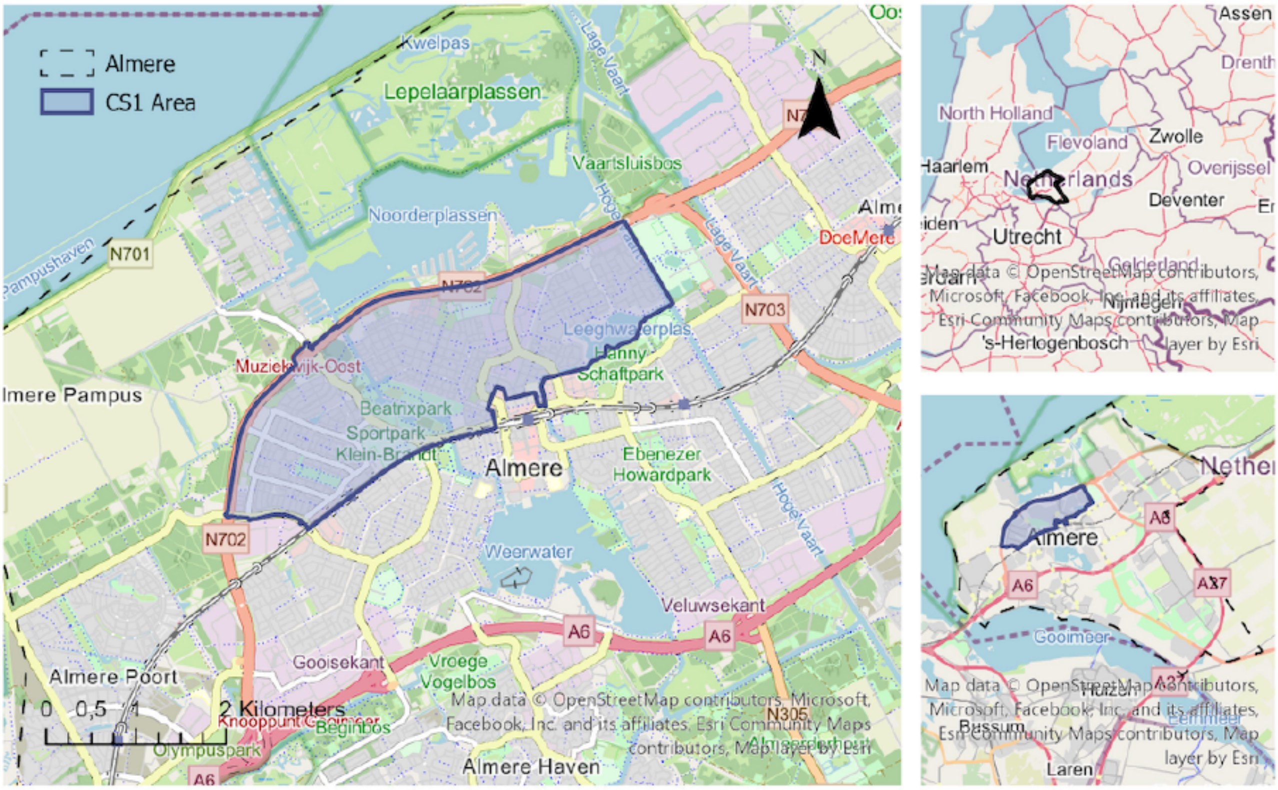


Figure 13, a map of the study area in case study 1.

This area was chosen for several reasons. First, it contained a good mix of water stations. All water types had multiple measurement points, except for surface water level which was only measured at one location in this area. As a result, this data mix provides a good set for answering the second research question. Furthermore, it can be tested whether a model trained on multiple stations (ground and sewage water models) outperforms a model trained on one station (surface water). Additionally, a range of different hydrological regimes could be investigated, namely groundwater, sewage water and surface water. At last, the data in this neighbourhood was mostly complete, enabling accurate testing because the ground truth is known.

4.1.2 Data Acquisition & Preprocessing

For this case study, data was acquired from three groundwater stations, six sewage water stations and one surface water station. Both ground and surface water had time intervals of one hour, whereas sewage water contained data per minute. The time interval of sewage water data was later modified to ten minutes to reduce training time. Additionally, precipitation data per minute of three weather stations was acquired. These data sets contained data from March 2022 till April 2024. From these three stations, the mean precipitation was derived. At last, daily evaporation data for the Almere municipality was acquired. Since this evaporation data was only available till 2023, it did not match the length of the water level and precipitation data set. To solve this issue, a random forest model was trained on meteorological data to estimate evaporation for the remaining days. This meteorological data came from the same three stations as the precipitation data. This model scored an NSE and KGE of 0.9 and thus the estimations can be considered accurate (Knoben et al., 2019). The specifics of the data can be seen in Table 9. A visualisation of the data for the different water level stations can be seen in Figure 14.

Table 9, an overview of the data in case study 1.

Variable	Unit	Stations	Time interval
Groundwater level	m NAP	3	Hourly
Sewage water level	m NAP	5	10 minutes
Surface water level	m NAP	1	Hourly
Evaporation	0.1 mm	1	Daily
Precipitation	mm	3	1 minute

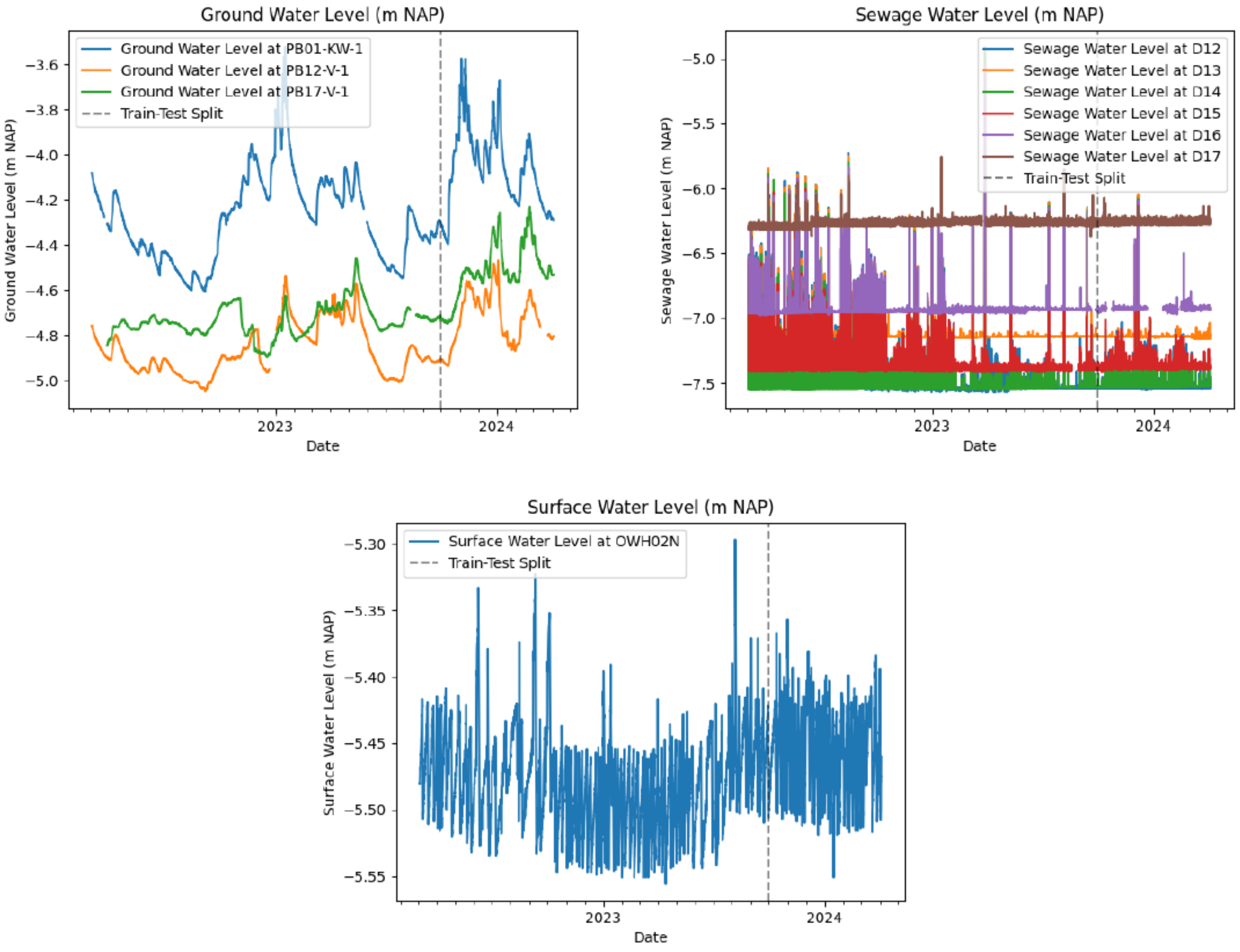


Figure 14, a visualisation of the water levels in the study area.

4.1.3 Feature Selection

Next, several features were extracted. Since the approach taken for this case study concerned the use of intra-station data, a lookback period of five previous measurements was used. In addition, the precipitation and evaporation since the last measurement was used to take into account the meteorological conditions. As a result, the features used as input for the models were the five previously recorded water levels and the precipitation and evaporation since the last measurement. Hence, this method is a mix between the intra-station system method and the more holistic system method as discussed in section 2.2.2. For optimal results, the input data was standardised for the MLP, SOM and SVR models to avoid an imbalance caused by differences in units. The other models are not vulnerable to these differences.

4.1.4 Model Selection

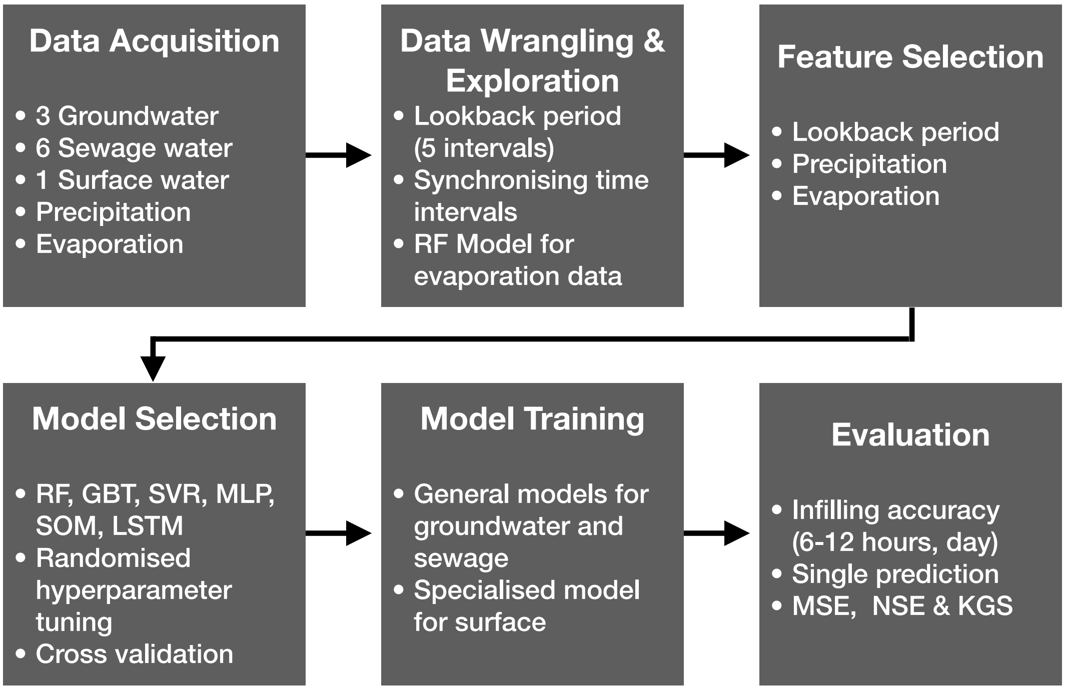
Model selection was done based on the findings of the MCA (Chapter 3). Therefore, the models selected for this case study were RF, GBT, SVR, MLP, SOM and LSTM. Each model was trained on four training datasets: groundwater, sewage water, surface water and a dataset containing all water level data. As a result, 24 models were created. These training sets contained data from March 2022, to October 2023. To optimise the models' performance, hyperparameter tuning was performed using cross validation and randomised search. To save time, a randomised search approach was used instead of a grid search and the number of cross validations was reduced to three instead of the more often used five times.

4.1.5 Model Evaluation

All these models were then evaluated in two ways. First, random artificial gaps were implemented in the test dataset, varying from six hours to a day in length. These relatively small gaps were chosen to start simple. Then, the infilling accuracy of the models was assessed by comparing the predicted series to the series containing the ground truth.

Second, a traditional approach was taken to see how well the models performed when it comes to estimating the water level at a single point in time. This evaluation was done on a test set, kept apart from the training data. This test set contained 7 months of data (October 2023 till April 2024). In both cases, the set of evaluation metrics used was MSE, NSE and KGE.

Figure 15 presents an overview of the methodology in relation to the theoretical framework. More details about the resources used can be found in Appendix A. Additional resources such as ChatGPT and other AI tools were used as support for bug fixes, refactoring code and explanation of some concepts, with a selection of conversations seen in Appendix B.



Data Acquisition
- 3 Groundwater
- 6 Sewage water
- 1 Surface water
- Precipitation
- Evaporation

Data Wrangling & Exploration
- Lookback period (5 intervals)
- Synchronising time intervals
- RF Model for evaporation data

Feature Selection
- Lookback period
- Precipitation
- Evaporation

Model Selection
- RF, GBT, SVR, MLP, SOM, LSTM
- Randomised hyperparameter tuning
- Cross validation

Model Training
- General models for groundwater and sewage
- Specialised model for surface

Evaluation
- Infilling accuracy (6-12 hours, day)
- Single prediction
- MSE, NSE & KGS

Figure 15, an overview of the methodology for Case Study 1.

4.1.6 Limitations

This methodology led to several limitations that affected the robustness of these results. First, despite adding precipitation and evaporation features, the hydrological process was not accurately represented. The precipitation and evaporation data between intervals was taken. However, due to geological processes, there is a lag between precipitation and evaporation and their effect on the water level of a water body. Hence, the use of these two features might have distorted the estimations from the models.

Second, the resolution of the evaporation variable, the resolution of water level data and the meteorological variables were not similar. Evaporation data was acquired for a day. To match the timescale of the water levels, this value for a day was divided linearly. As a result, the implemented solution of evaporation data did not take into account the non-linearity of the evaporation data, which is affected by daily cycles of the weather (Pagano & Sorooshian, 2002). Therefore, evaporation data used was not accurate on the timescale used.

Third, the sewage data resolution was changed to 10 minutes, instead of the original resolution of 1 minute per measurement. As a result, some information about the sewage water level was lost. This change was made because the training of some models took too long, and this reduction of resolution would result in a dataset ten times smaller, speeding up training significantly.

Fourth, model selection through randomised search was done based on MSE. Hence, the selection procedure was biassed towards models which might have sacrificed NSE and KGE scores in favour of increased accuracy.

At last, limited training data was available, potentially limiting the performance of some models. This limitation is discussed in more detail in the discussion section as the effects vary per model and the limitation was applicable for both case studies.

4.2 Results

In this section, the results of the case study will be presented. First, the infilling results in general are presented (section 4.2.1). Thereafter, a closer look at the individual models will be taken, including their accuracy when predicting water level at a single point in time (sections 4.2.2-4.2.7).

4.2.1 General Infilling Results

The infilling results were mixed, as can be seen in Figure 16. When looking at MSE, both the RF and GBT models showed great accuracy consistently while the accuracy for the MLP was very bad. The SVR, LSTM and SOM models did a mixed job as some of these models showed high accuracy and others did a bad job, as indicated by their longer range in the box plots in Figure 16. When looking at NSE and KGE for more information about the accuracy of the models, it can be observed that all models performed poorly. None of the models achieved a positive NSE, while only the LSTM achieved a positive KGE on a rare occasion. These mixed results indicate that the models are not good in predicting the variance of the observed water levels.

Taking a look into the different gap lengths, the following results were obtained (Figure 17). First, the best median MSE is obtained for the gaps with a day in length, while the worst MSE is obtained for a gap of twelve hours. When looking at NSE and KGE scores, infilling performance was best for the gap of twelve hours, and worst for the small gap of six hours.

When comparing the performance across water types in Figure 18, it can be seen that the result of generally high infilling accuracy, but low ability to capture variance is obtained for all water types. Only for groundwater, a positive KGE is achieved on a rare occasion. Furthermore, for surface water the infilling performance is worst.

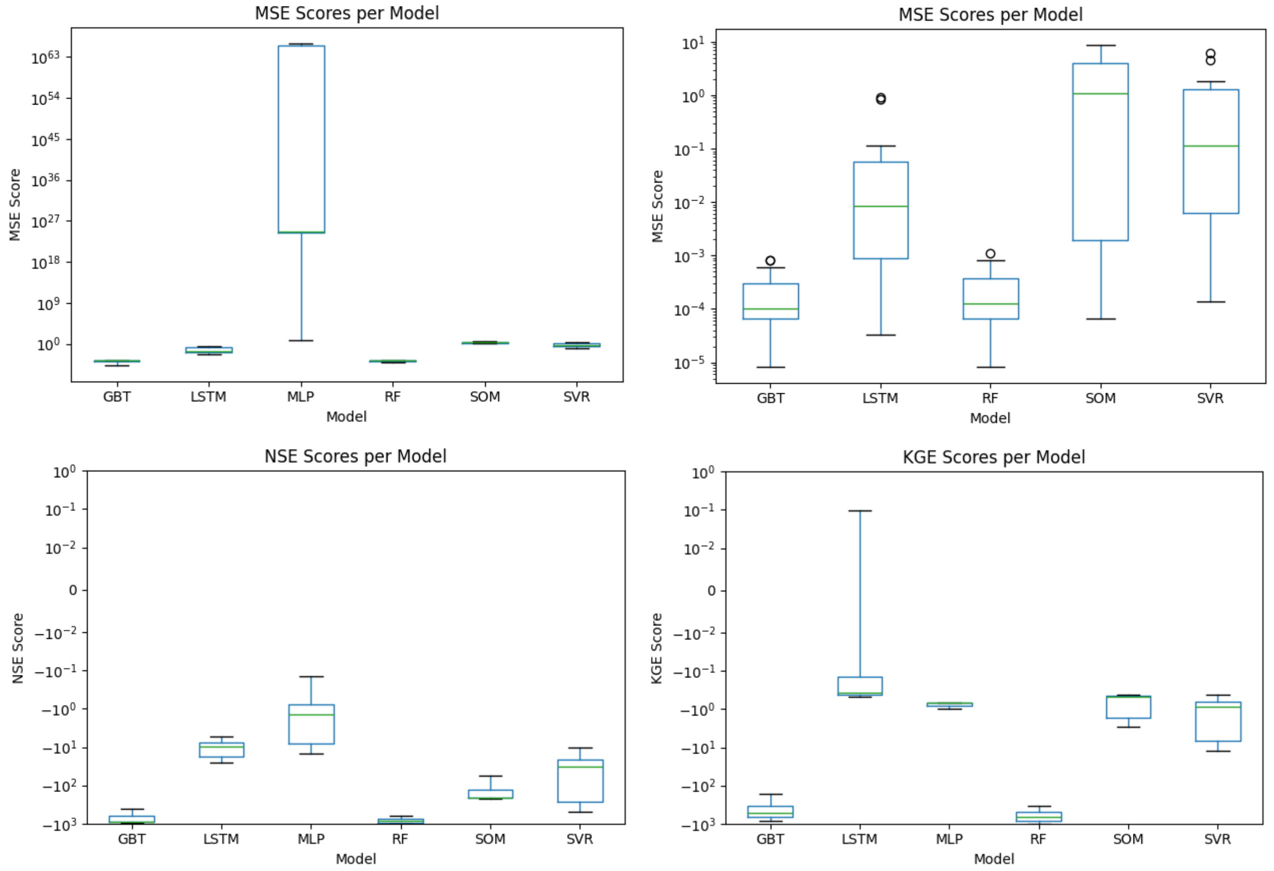


Figure 16, box plots of the obtained infilling results (upper left: MSE, upper right: MSE without MLP, down left: NSE, down right: KGE).

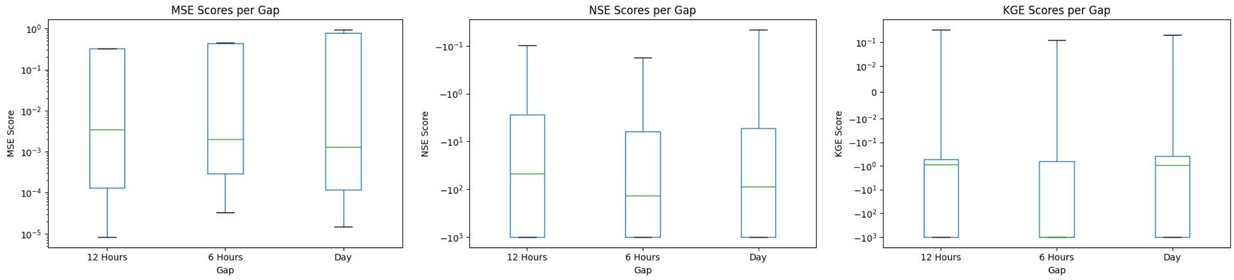


Figure 17, box plots of the obtained infilling results per gap length (left: MSE, center: NSE, right: KGE).

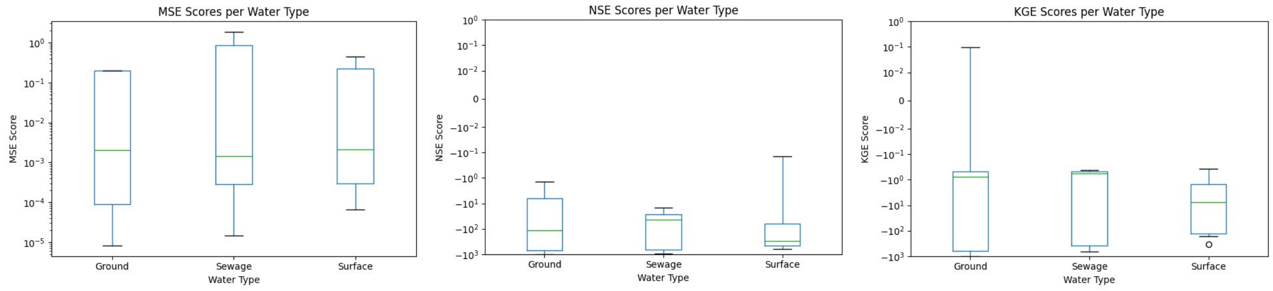


Figure 18, box plots of the obtained infilling results per water type (left: MSE, center: NSE, right: KGE).

With these general results in mind, it is helpful to take a closer look into the infilling performance of the models. This provides insight into the different characteristics of the models.

4.2.2 Support Vector Regressor (SVR) Results

The SVR model (Figure 19) is able to create a series with variance. However, it predicts variance that is not observed. Furthermore, the estimated water levels are somewhat accurate with most estimates within a range of 15 centimetres of the observed measurement, with outliers to 30 and 70 centimetres.

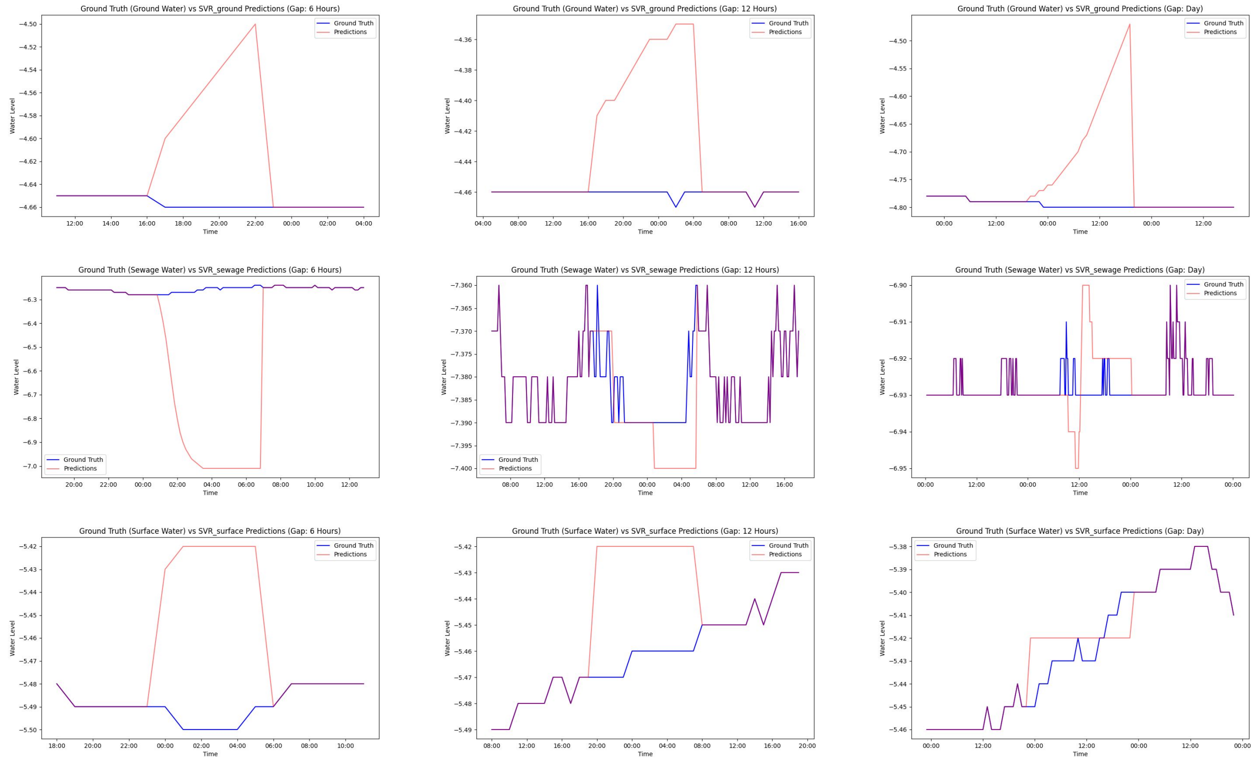


Figure 19, an overview of the infilling results obtained by the SVR models (upper row: groundwater, middle row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

When looking at the single prediction accuracy of the SVR models in Figure 20 to explain this behaviour, a few things stand out. First, the model tends to overestimate lower and higher water levels of ground and sewage water, while underestimating water levels closer to the median. This could explain why it often overestimates or underestimates gaps.

Second, it performs poorly on surface water data as it generally overestimates the water level. Furthermore, it can be seen that the model is range bound. When looking at the training data (Figure 14), this behaviour is difficult to explain, given that the range of the water level is higher and the water level also seems to be lower in the training set.

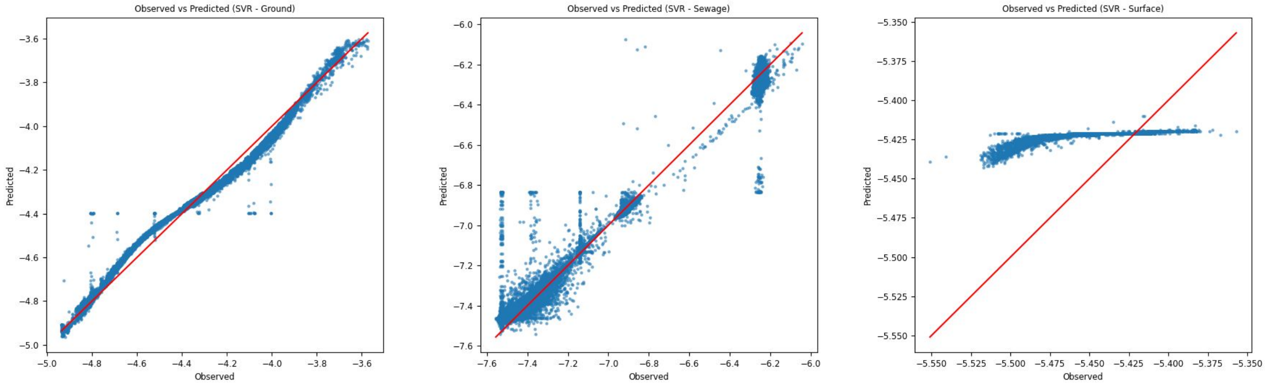


Figure 20, Q-plots for the groundwater (left), sewage water (center) and surface water (right) SVR models.

4.2.3 Random Forest (RF) Results

The results for the RF are mixed as seen in Figure 21. When looking at the MSE, its accuracy is very high as most estimates are within one to three centimetres of the observed water level.

However, the RF model fails to capture the dynamics of the water level during the artificial blackout. As a result, its NSE and KGE scores are low.

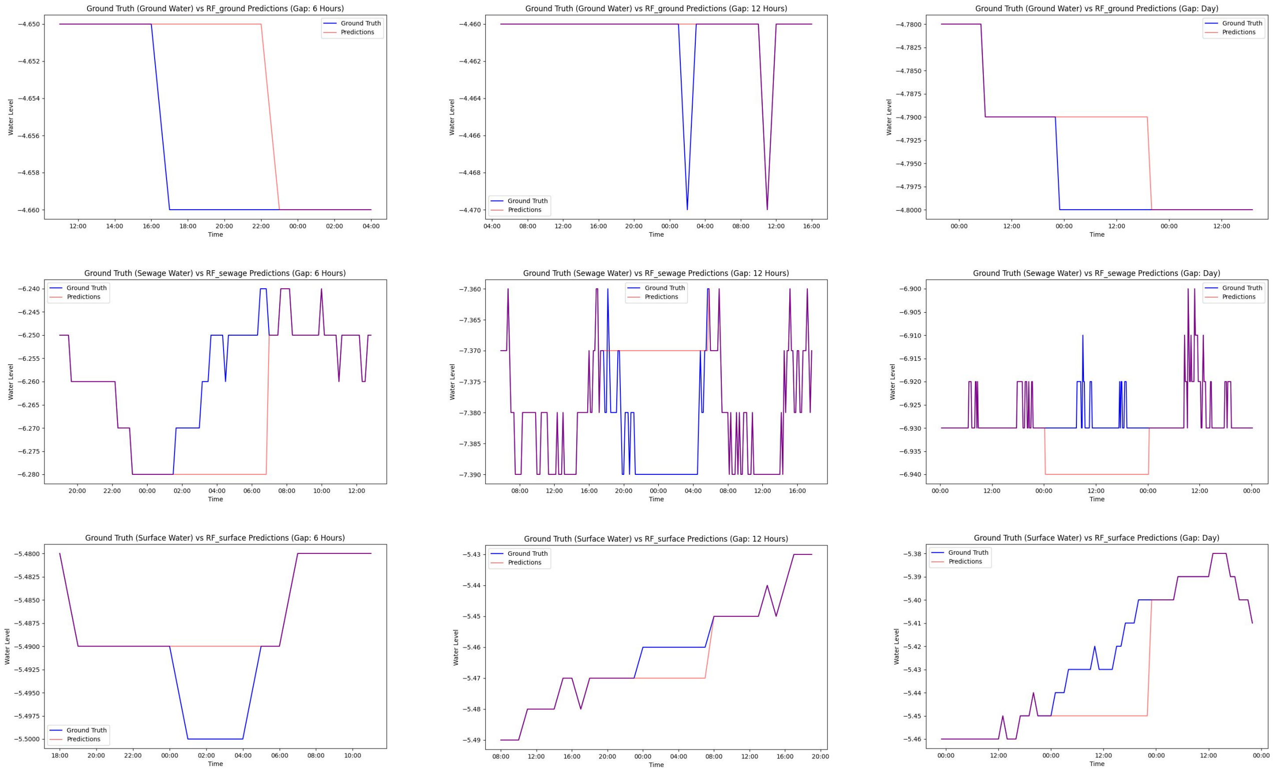


Figure 21, an overview of the infilling results obtained by the RF models (upper row: groundwater, middel row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

Looking at the prediction accuracies of the RF models for a single point in time in Figure 22, it stands out that for the groundwater and sewage water the model is relatively static, as is evident from the horizontal lines in the Q-plot. Only a few outputs do the trick. This could explain the low NSE and KGE scores of the RF models. Its set of outputs is simply too small to capture the variance, which is usually only a few centimetres, correctly.

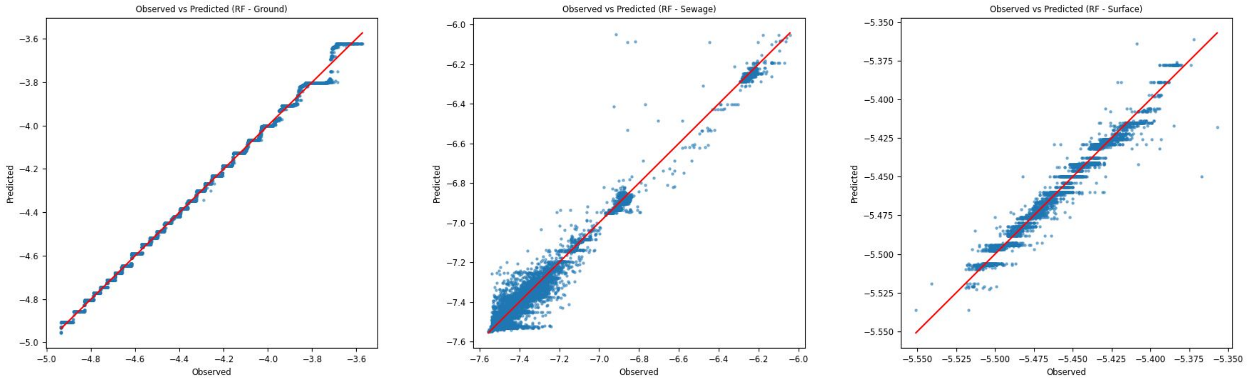


Figure 22, Q-plots for the groundwater (left), sewage water (center) and surface water (right) RF models.

4.2.4 Gradient Boosting Trees (GBT) Results

For GBT, the infilling results paint a similar story as RF: high accuracy in terms of MSE, low accuracy in terms of NSE and KGE. Hence, this model also fails to capture the variance of the water level in the artificial gap. Infilling results for the GBT are shown in Figure 23.

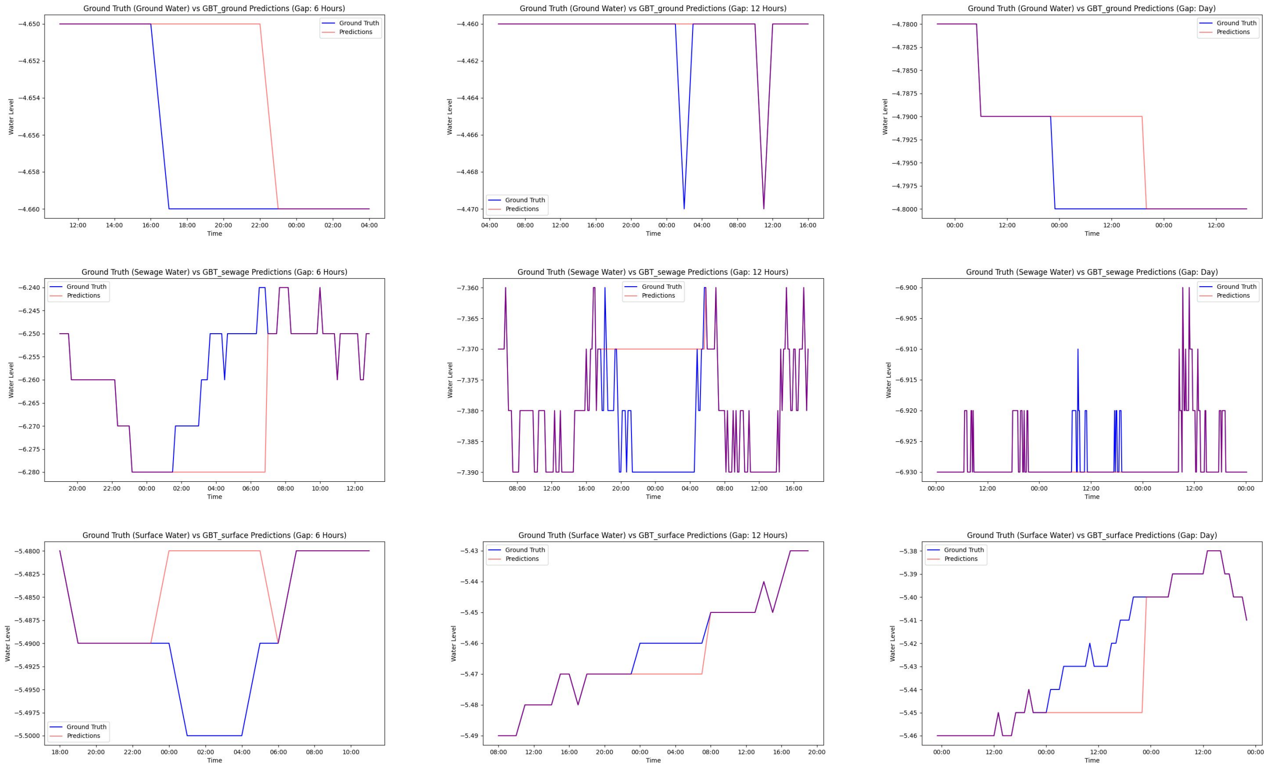


Figure 23, an overview of the infilling results obtained by the GBT models (upper row: groundwater, middel row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

For the GBT models it can be observed in Figure 24 that it makes more or less the same predictions as the RF but that it is more flexible and accurate. This flexibility is seen as the horizontal lines are not visible here, and thus its set of outputs is bigger. However, this flexibility and ability to capture variance correctly was not transferable to infilling accuracy.

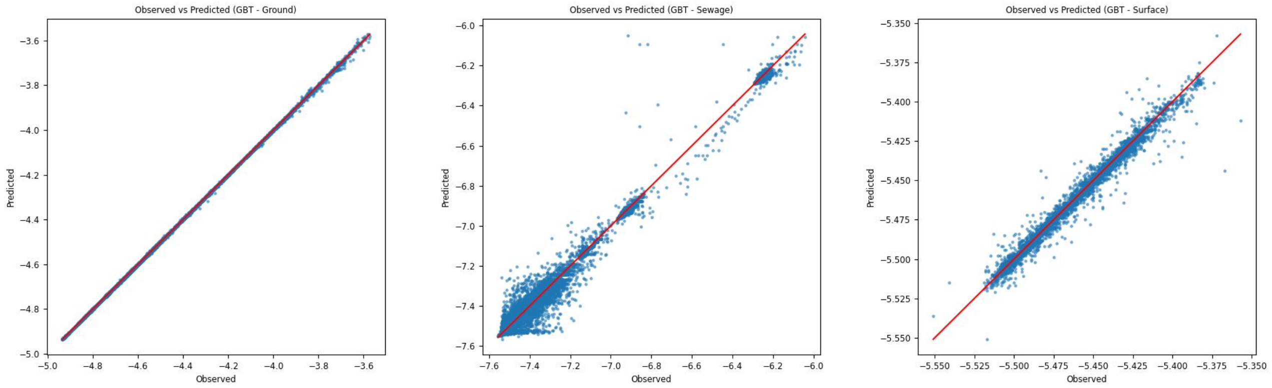


Figure 24, Q-plots for the groundwater (left), sewage water (center) and surface water (right) GBT models.

4.2.5 Multi-Layer Perceptron (MLP) Results

The MLP models had the poorest results. The majority of the MLP models create an exponential function when filling in the gaps, and thus its accuracy is very bad (Figure 25). Occasionally, the MLP's estimation comes within centimetres of the observed value, while for others it is off by a few decimetres. However, for some of the surface gaps it is off by orders of magnitude.

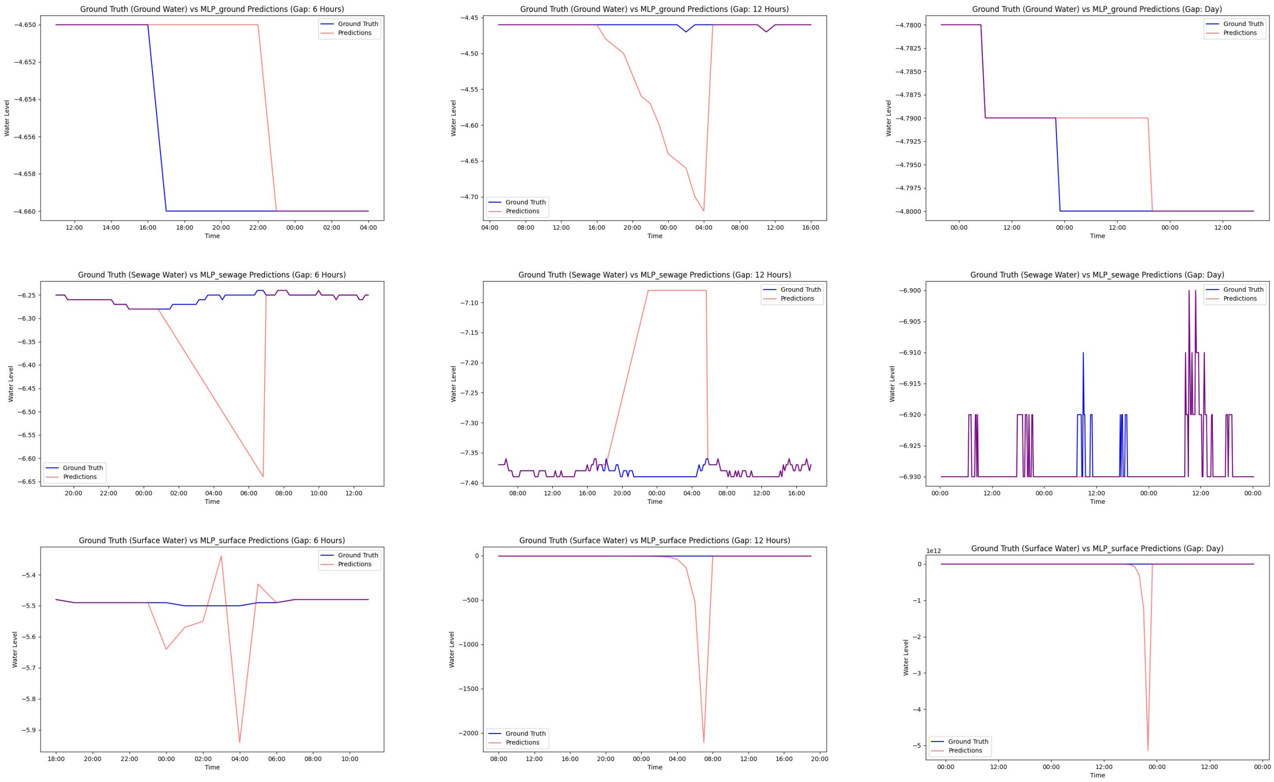


Figure 25, an overview of the infilling results obtained by the MLP models (upper row: groundwater, middel row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

As far as the accuracy for estimating water level at a single point in time is concerned, the MLP models perform well, except on surface water data where it underestimates the water level (Figure 26). This could be explained by the observation that the training set contains lower water levels than the test set (Figure 14). Furthermore, a lack of training data might have prevented the convergence of the weights of the model. The range of the model and its exponential infilling are difficult to explain as it is significantly lower than the data in the training set. One potential explanation could be the oscillation of the surface water level (Figure 14). The MLP might have learned that the water level rises and falls, but did not learn the maxima and minima of these cycles.

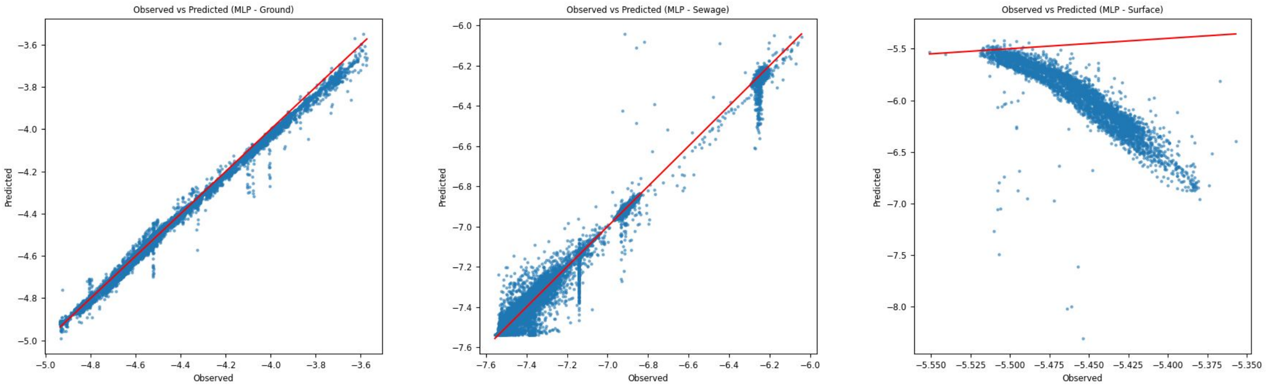


Figure 26, Q-plots for the groundwater (left), sewage water (center) and surface water (right) MLP models.

4.2.6 Long Short Term Memory (LSTM) Results

The LSTM (Figure 27) tends to diverge from the ground truth while its predictions are not very far off. This divergence tends to be biassed to underestimations, although overestimations also occur. Similar to MLP, LSTMs also show signs of exponential behaviour, although it is far less inaccurate than the MLPs. The estimates of the LSTM are mostly within centimetres of the recorded water levels, with some estimates being off by almost a metre.

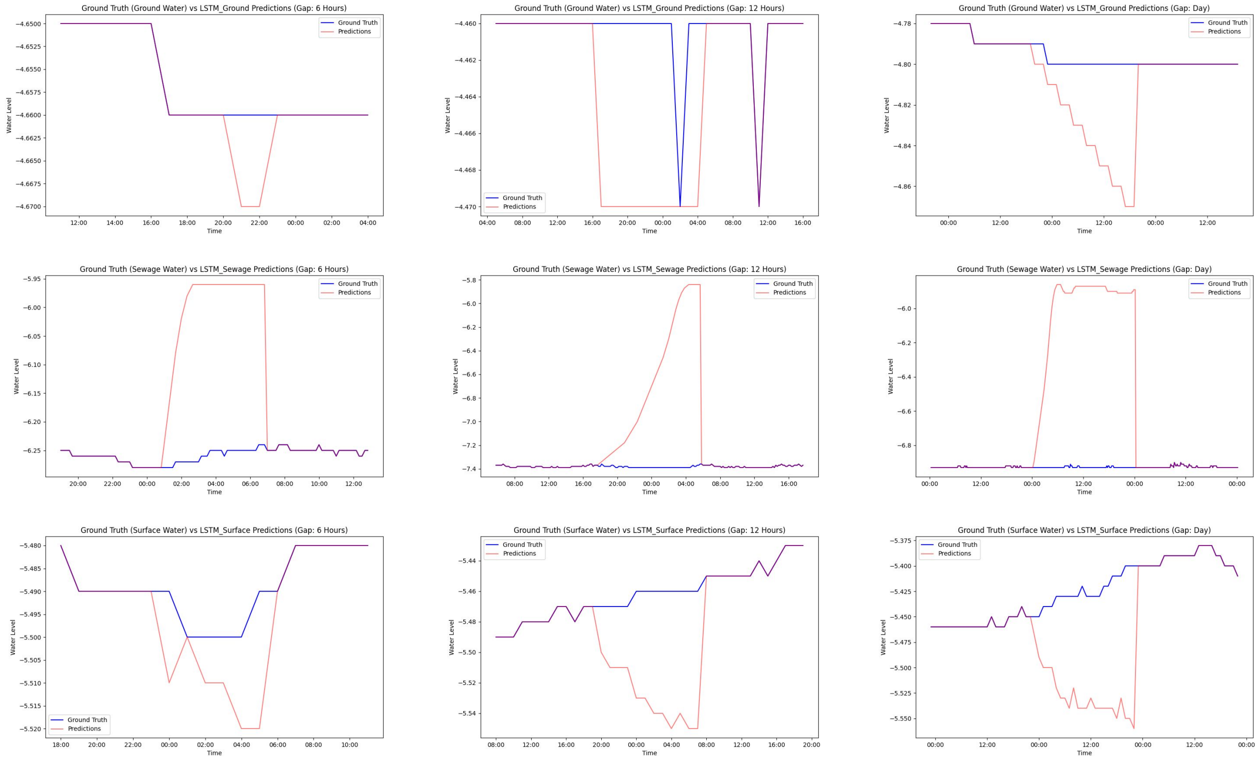


Figure 27, an overview of the infilling results obtained by the LSTM models (upper row: groundwater, middel row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

For the LSTM’s single predictions, it can be seen in Figure 28 that it sometimes overestimates the water level for groundwater. Its underestimation of surface water is probably due to the discussed difference in training and testing data. Both behaviours could cause its tendency to diverge from the ground truth, either upwards or downwards. Here again, the lack of sufficient training data could have prevented the weights of the model to converge sufficiently for the surface water model.

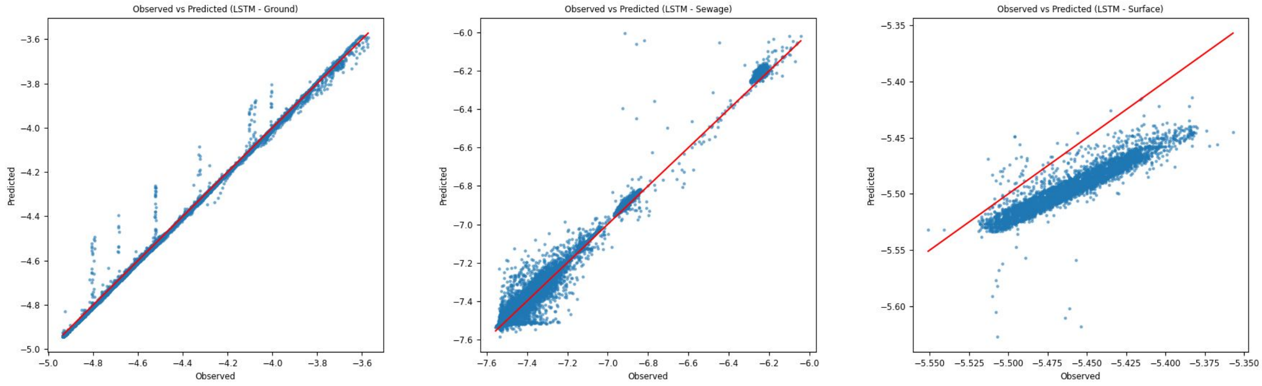


Figure 28, Q-Plots for the groundwater (left), sewage water (center) and surface water (right) LSTM models.

4.2.7 Self-Organising Map (SOM) Results

Taking a look at the SOM models’ infilling behaviour (Figure 29), it suffers from inflexibility and poor extrapolation. However, occasionally it is able to create dynamic time series but suffers from poor estimates and predicts inaccurate variance. Its estimates are off by a few centimetres to a few decimeters in some occasions. It both underestimates and overestimates the water level in the gaps.

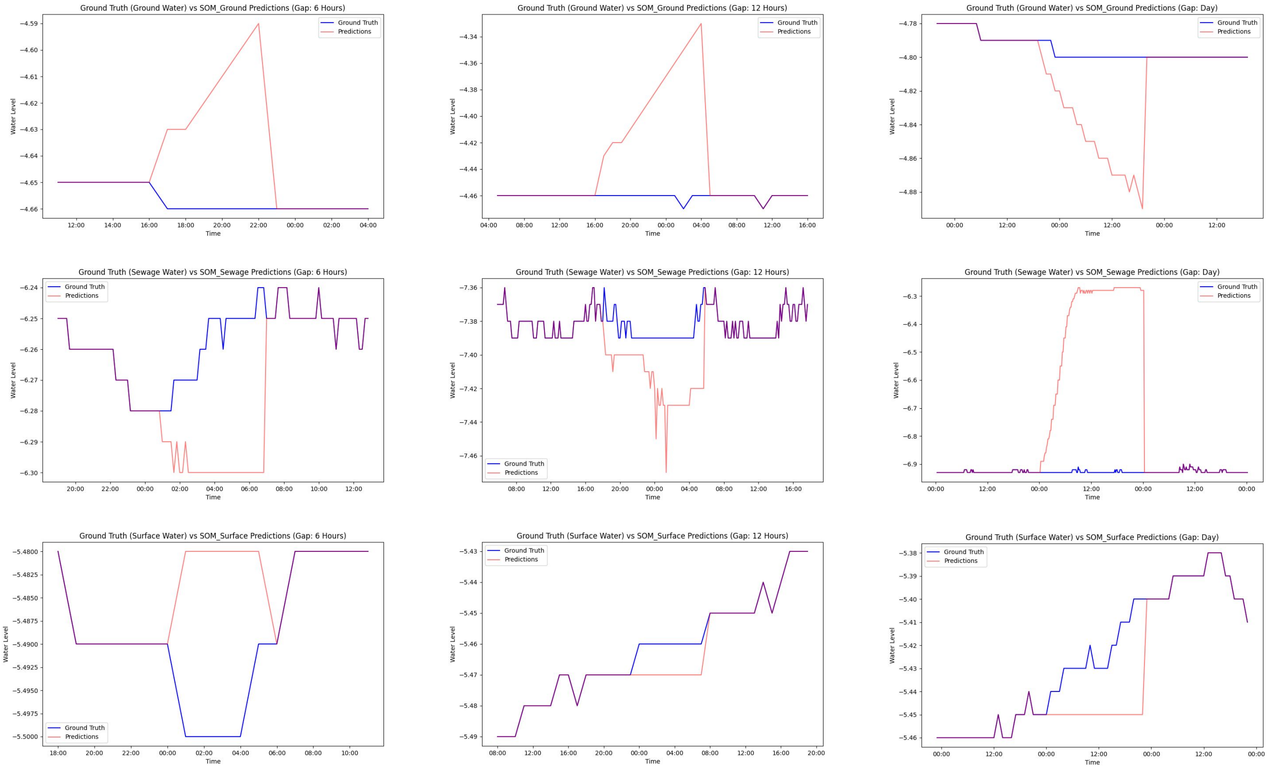


Figure 29, an overview of the infilling results obtained by the LSTM models (upper row: groundwater, middel row: sewage water, lower row: surface water & left column: 6 hour gap, middle column: 12 hour gap, right column: day gap).

Looking at Figure 30 to assess accuracy for a single point in time, it can be seen that the SOM models are range bound and show a similar tendency towards inflexibility, seen by the lack of smooth curves, especially for surface water. For example, for ground water it under estimates water level lower than -4.1. This indicates very poor extrapolation capability. A potential explanation of this behaviour is that in the training set there is only a minority of data points at certain levels, and thus the nodes of the map have not converged towards these relatively rare water levels. This inflexibility and poor extrapolation can be explained by the architecture of the SOM as it only contains a finite number of nodes and thus outputs.

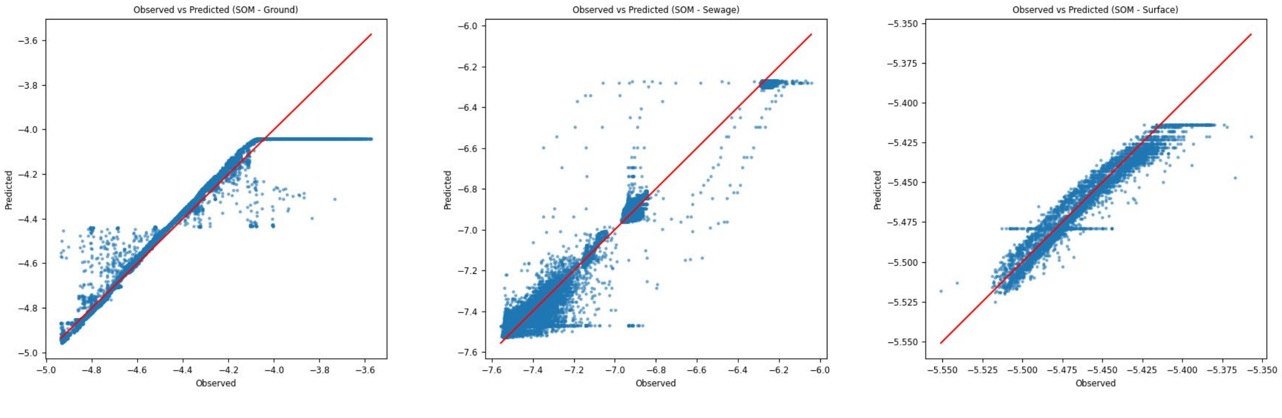


Figure 30, Q-Plots for the groundwater (left), sewage water (center) and surface water (right) SOM models.

4.3 Discussion

The implementation of the six selected models using the intra-station method delivered some interesting results. First, the difference in accuracy of the models themselves. The RF and GBT models achieved highest accuracy, while the MLP achieved lowest accuracy. This will be discussed in the General Discussion section (6.1) to avoid repetitions. Second, the difference in accuracy as implied by the evaluation metrics: MSE scores painted a brighter picture than the NSE and KGE scores. This implies that the models had difficulty with estimating the variance of the water level, hinting at poor scalability. Third, the method of training models on either multiple stations or a single station also provided mixed results. Both training methods did not lead to superior results. At last, the practical implementation of the models brought to light some important limiting factors for some models.

4.3.1 Intra-Station Infilling

What was striking about these results is the two sided picture painted by the evaluation metrics: good ability to estimate water level, low ability to estimate its variance, generally. This implies that although most models can estimate water level with acceptable accuracy, they are not able to estimate the variance of the water level during the period of the gap. Hence, the scalability towards bigger gaps is probably poor. There are numerous potential explanations for these mixed results.

A first potential explanation is related to feature importance for the models. Looking at Figure 31, it can be seen that the previously measured or estimated water level is by far the most important feature used for estimation for most models, except LSTM and SOM. LSTM gives most importance to precipitation and almost equal importance to the water levels previously measured or estimated. For SOM, all features are equally important since the input values were standardised. Hence, distances between values on the map are not distorted by difference in scale.

By primarily focusing on the previous water level, the models fail to take into account the dynamics of the water level during the lookback period and thus potential variance during the period of the gap. This thus could explain the lower NSE and KGE scores. It also explains the low MSE scores for most models because, in general, water levels are unlikely to fluctuate massively in between consecutive measurement intervals. Furthermore, mistakes in estimation are propagated throughout the period of the gap. If the model makes a mistake in its first estimation, then this mistake weighs too heavily on the next estimation, repeating until the gap is filled.

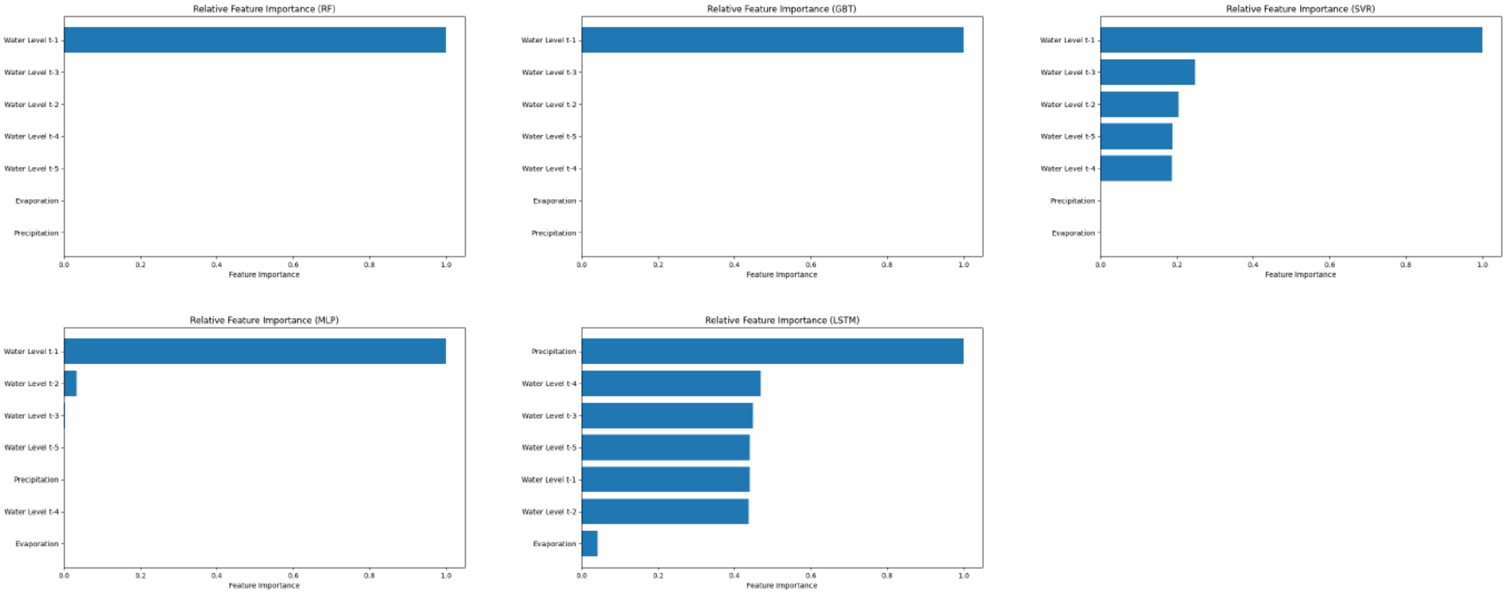


Figure 31, an overview of the feature importance analysis for the RF (upper left), GBT (upper center), SVR (upper right), MLP (lower left) and LSTM (lower center).

This failure to estimate variance most likely implies that the intra-station method has difficulties with infilling greater gaps accurately. For other researchers, the intra-station infilling method had more success. There are various potential explanations that could explain the difference between the results obtained in this thesis as compared to the results of other researchers.

4.3.2 General vs Specialised Model

A first explanation for the failure to estimate the observed variance during the gap periods may lie in the approach taken for training. The models for groundwater and sewage water were trained on multiple stations. As a result, these models might not have learned the spatiotemporal variances that are unique to each station. Hence, the poor estimation of this variance could result from generalising these variances too much.

However, caution is needed here as the results can be compared to a specialised approach. Indeed, the surface water models were trained on one station, OWH02N, potentially allowing to learn the temporal variance unique to this station. These models showed similar results, although there might be different causes here. Namely, the models SVR, MLP and LSTM underperformed for this water type. This underperformance could potentially be caused by insufficient training data, given that this training set was one third the length of the set of groundwater (which was trained on three stations).

The use of general models also had another downside. Because the training set became very big, especially for the sewage dataset and the dataset containing all data, training and inference times also became very long for some of the models. This will be elaborated on in 4.3.3.

4.3.3 Performance Challenges

A second explanation for the underperformance achieved by the intra-station method in this research project is the limited amount of data. In this project, the models were trained on a dataset containing roughly 1.5 years of data. In other projects, the dataset with training data contained data of two decades (Janbain et al., 2023). As a result, the models in this project are biassed towards the conditions of only 1.5 years, while these conditions can vary per year. For example, the weather is not the same every year, or interventions could be made after a year, etc.

Third, more specifically focused on the LSTM, other researchers used more complex forms of this model. For example, Janbain et al. (2023) used several LSTM units. In this project, the LSTM model was limited to only one unit. For other models, it could be that other researchers also used more complex variations (i.e. more hidden layers for MLP, bigger maps for SOM).

Fourth, other researchers could have used a bigger lookback period. This could give the models more insight into the variance of the water level. However, given that the previously measured or estimated water level had very high importance for almost all models, this could be unlikely.

Moving on, there are also several potential explanations for the underperformance of the SVR, MLP and LSTM models for surface water. First, the underperformance of the MLP and LSTM might be attributable to the lower amount of training data. Both sewage and groundwater contained more training data because of the presence of more stations. For the sewage water, more training data was also available due to its higher frequency of measurement as well.
Second, Almere's surface water levels are highly regulated, as seen by the oscillation between the minima and maxima in Figure 9, thus making it more difficult to make estimations. Arriagada et al. (2021) mentioned that it is more difficult to infill data when there is a large amount of human interventions, which seems to be confirmed here.

When looking at the practical implementation of the models, two remarks can be made. First, training time for SVR was too long. It took a total of 10 days for the SVR models to complete training, compared to roughly four hours for the other models. This is not surprising, since the water level regime is highly nonlinear. Therefore, the SVR model's kernels are needed to create a nonlinear equation. However, these kernels are very computationally intensive. As a result, the SVR became very computationally intensive.

Second, inference times for the SOM models were also very long. This is not surprising due to its needs to calculate the distance of the input data to all its nodes. However, hyperparameter tuning resulted in a very big map as the best performer. Hence, the map contained a lot of nodes for which the distance had to be calculated, leading to rising inference times.

4.3.4 Implications

These results had several implications for the next case study. Indeed, the models SVR and SOM were dropped since their long training and inference time, respectively, did not lead to superior performance with regards to accuracy. Therefore, sacrificing the user experience of the AI tool is not worth it.

Secondly, a more tailored approach was needed. Although it is not for certain that the infilling performance was hurt by training the models on multiple stations, the exponential rise in training times is not user friendly. Furthermore, when looking at the theory of hydrological regime, it is helpful to take into account the spatiotemporal variance between stations and thus their need for a unique approach.

These findings were taken to the second case study, presented in the next section.

5. Case Study 2 - Inter-Station Infilling

In this chapter, the case study to answer the second research question is presented: *how effective are the models in infilling based on the use of inter-station data?* For this second case study, the first objective was to see whether infilling accuracy could be improved by using the inter-station approach. The second objective was to see whether a general set of steps applied to each station could be established to enhance infilling performance while accounting for spatiotemporal differences between stations. This would help to further shape the underlying method of the AI tool. Third, the goal was to see whether a more transparent method of evaluation could be developed. This is crucial for the user of the AI tool, as users can see how the AI tool performs on their data. At last, the computational efficiency of the process had to be streamlined to further improve the user experience of the tool.

This chapter is structured as follows. First, the methodology section (5.1) explains the steps taken during the implementation of the models in this second case study. Second, the infilling results will be presented and analysed in section 5.2. At last, the results of the case study will be discussed in section 5.3 to provide a clear answer to the research question.

5.1 Data & Methods

5.1.1 Study Area

The study area was expanded as compared to the first case study, focusing on Central Almere as seen in Figure 32. As a result, more stations were taken into account. This study area was chosen for the same reason as the previous study area: good amount of measurement stations for each water type, high quality of data in terms of completeness and different types of land use.

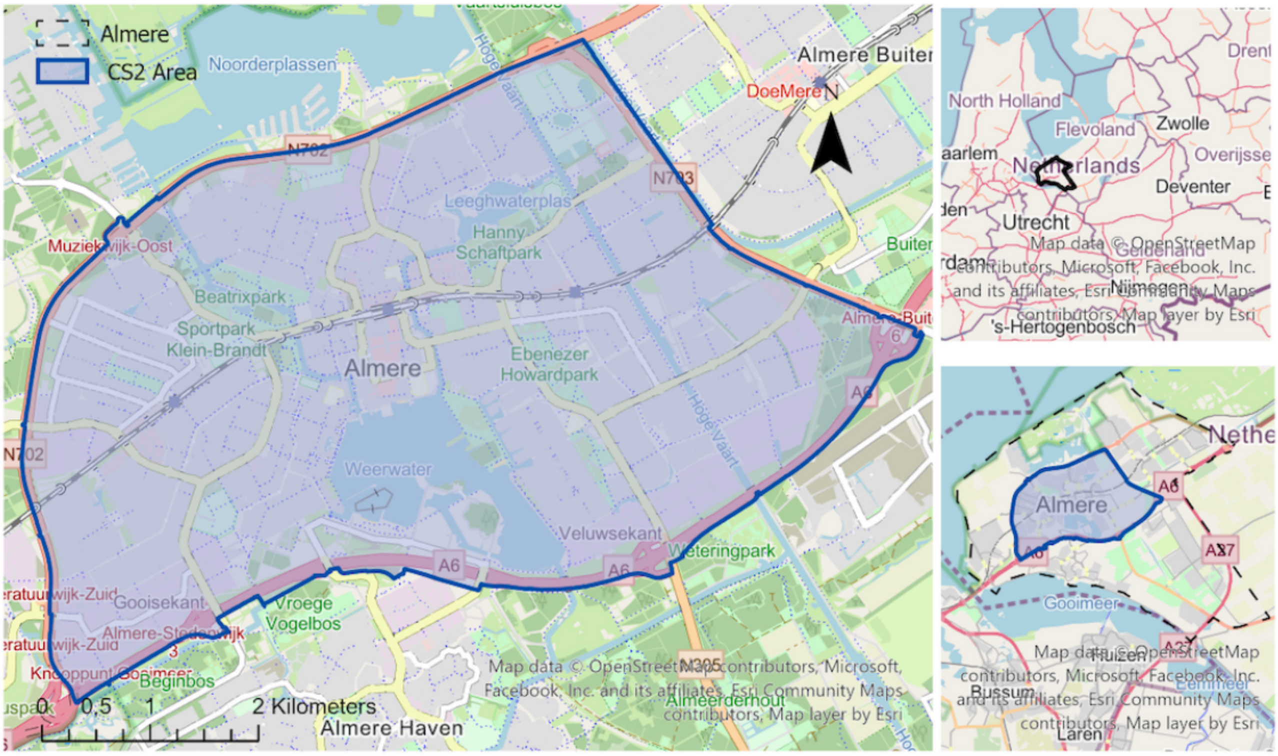


Figure 32, a map of the study area in case study 2.

5.1.2 Data Acquisition & Preprocessing

A total of thirteen groundwater stations, six sewage stations and six surface water stations were taken into account. Furthermore, the same precipitation and evaporation datasets (including RF model estimations) as in the first case study were used. The dataset shares the same characteristics as in Chapter 4, as seen in Table 10. A visualisation of the water level at these stations can be found in Figure 33.

Table 10, an overview of the data in case study 2.

Variable	Unit	Stations	Time interval
Groundwater level	m NAP	3	Hourly
Sewage water level	m NAP	5	10 minutes
Surface water level	m NAP	1	Hourly
Evaporation	0.1 mm	1	Daily
Precipitation	mm	3	1 minute

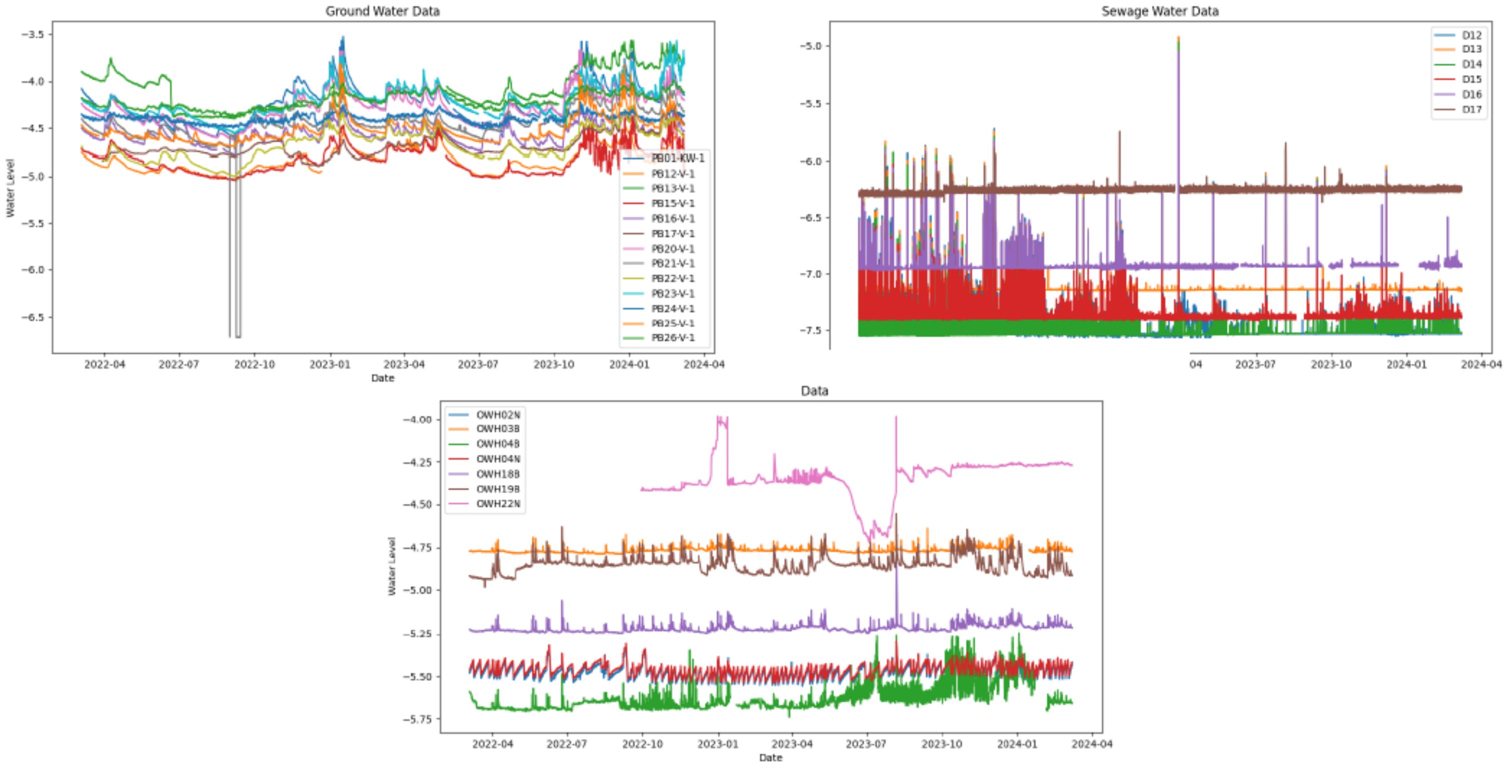


Figure 33, a visualisation of the water levels in the study area.

Next, a wide variety of features related to the meteorological data was created. For both precipitation and evaporation, the total of the last couple of days, weeks and months was calculated, with a maximum of half a year. These features were extracted to provide more context to the hydrological processes of the study area. Additionally, temporal features such as the month and hour of the day were added to keep natural and human cycles into account.

5.1.3 Feature Selection

For feature selection, a different approach was used compared to the first case study. To take into account the spatiotemporal variance between the stations, features were selected per station. For each station, a minimum of three features was selected and a maximum of ten. These features were chosen based on the correlation of the feature and the water level of the specific station. The chosen features had to have a minimal correlation of 0.3, unless the minimum of three was not reached. Similar to the first case study, the data was standardised for the MLP model.

5.1.4 Model Selection

With regards to model selection, SVR and SOM were dropped from the set of models used in the previous case study due to their long training and inference time, respectively. As a result, only RF, GBT, MLP and LSTM were used. For each station, a set of four models was thus trained. Again, hyperparameter tuning using cross validation and grid search was performed for optimal model selection.

5.1.5 Model Evaluation

In this case study, the evaluation approach was different from the previous case study. Since the model does not depend on earlier predictions, the entire seven-month test set was used as one continuous gap to assess infilling accuracy. This method provides a more comprehensive evaluation of the infilling performance, as a longer period contains more variance than a shorter one. Hence, more insight into the strengths and weaknesses of the model could be obtained. Furthermore, this ensures that the models had similar amounts of training data as compared to the first case study. This enabled a fairer comparison of the results. The same metrics as previously were used for evaluation: MSE, NSE and KGE.

5.1.6 Model Pruning

After evaluating the models, an additional step was introduced to enhance their performance. To potentially increase performance, and reduce the model size, features were recursively pruned based on their relative importance. This pruning stopped when the resulting model was not superior to the existing model, indicate by the evaluation metrics.

Figure 34 provides an overview for the methodology described above. More details about the resources used can be found in Appendix A. Additional resources such as ChatGPT and other AI tools were used as support for bug fixes, refactoring code and explanation of some concepts. Appendix B contains a selection of such AI conversations.

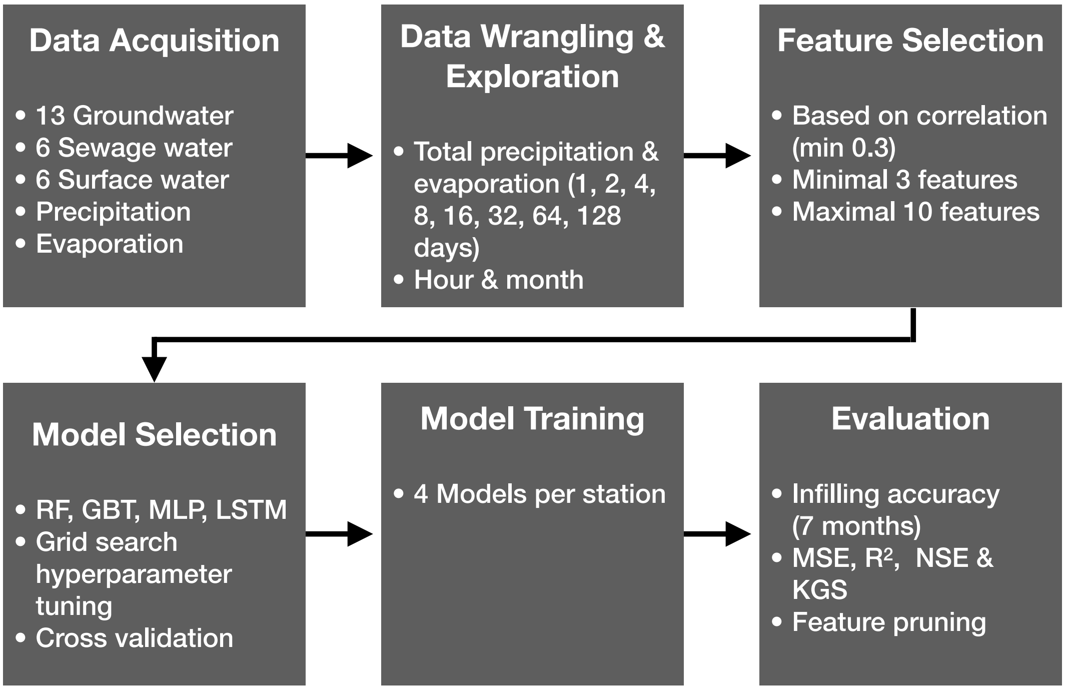


Figure 34, an overview of the methodology of Case Study 2.

5.1.6 Limitations

The robustness of the results were impacted by a few limitations resulting from the methodology described above. First, a limited amount of training data, for which the consequences have been discussed already.

Second, the results could be affected by the amount of correlating stations. For some stations more correlating stations could have been available outside of the study area. This could have improved infilling accuracy for these stations. In practice, there could also be little correlation between stations due to their spatiotemporal variances, limiting the applicability of the inter-station method.

Third, the maximum length available for testing purposes limited the inquiry into the scalability of the infilling accuracy to seven months. However, this length is deemed appropriate enough for decent analysis and robust results, given that the data showed that sensors were offline for only a few weeks at most. Hence, seven months provides more than enough insights into the scalability of the models.

Furthermore, the model selection procedure through hyperparameter tuning again relied upon MSE scores. As a result, the results are biassed towards this metric and not to the NSE and KGE metrics.

At last, the long length of the gap affected the infilling results due to increased uncertainty and thus a straight comparison to the previous case study needs to be made with caution. However, it is safer to extrapolate inter-station results because of their reliance upon ground truths. This makes their behaviour more predictable.

5.2 Results

In this section, the results of the case study will be presented. First, the infilling results in general are presented. Thereafter, a closer look at the individual models will be taken, to gain further insight into their characteristics.

5.2.1 General Infilling Results

Infilling results were acceptable to good as implied by all three metrics. As can be seen in Figure 35, the models have relatively equal performance when looking at the median KGE and NSE. However, it can be seen that for LSTM and MLP, the range of accuracy is much higher than for RF and GBT, indicating lower reliability. Furthermore, all models score higher on the KGE metric compared to the NSE metric. This implies that the models are able to take into account the variance, bias and correlation of the water levels, but the exact predictions can be off sometimes.

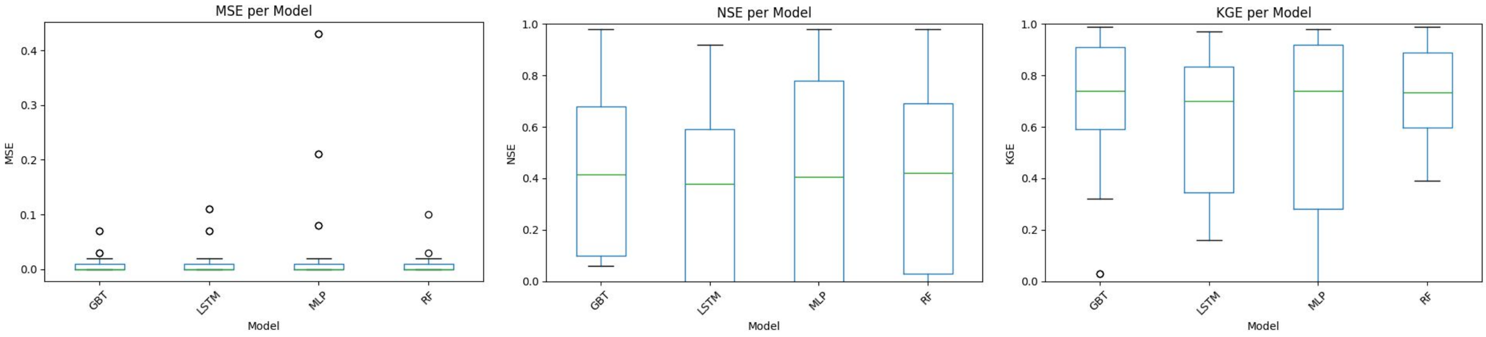


Figure 35, box plots of the obtained infilling results per model (left: MSE, center: NSE, right: KGE).

When looking at the results per water type in Figure 36, it can be seen that the models perform best on the sewage water data, and worst on the surface water data. Again, scores for KGE are higher than NSE.

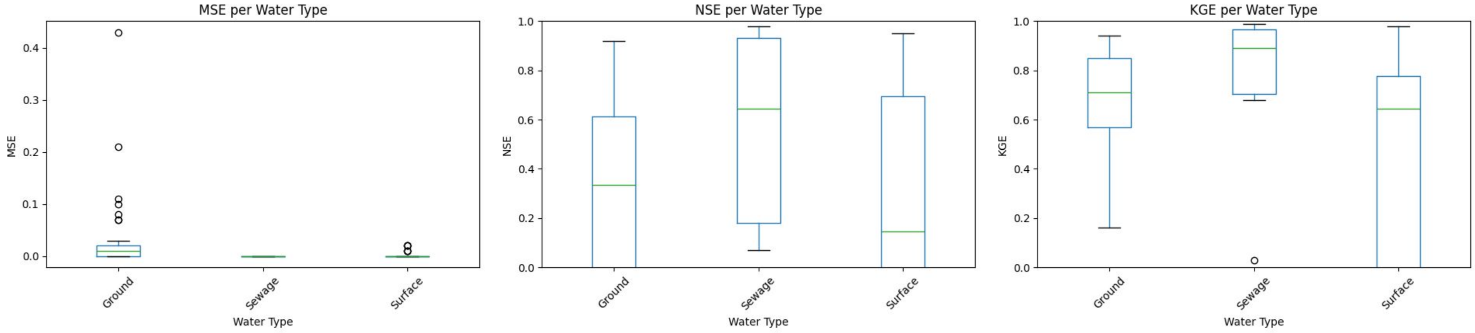


Figure 36, box plots of the obtained results per water type (left: MSE, center: NSE, right: KGE).

Having had a look into the infilling results on a higher level, it is also helpful to dive deeper into the models themselves. This will further shape our understanding of the behaviour of these models.

5.2.2 Random Forest (RF) Results

As can be seen in Figure 37, infilling performance was mostly acceptable to good for the RF models. However, for some stations, infilling performance was very bad. What can also be observed is that, generally, KGE scores are higher than NSE scores.

Taking a closer look into the RF's infilling performance in Figure 38, it can be observed that the RF models still lack some flexibility. Again, the model only uses a few outputs for its estimations and thus horizontal lines can be observed when plotting the observed values to the estimations. This is also seen by the less flexible infilling, where there are some vertical lines, implying a non-continuous function. Despite this inflexibility, the model's accuracy is generally good. Even with low KGE and NSE scores, it is only off by a few decimeters.

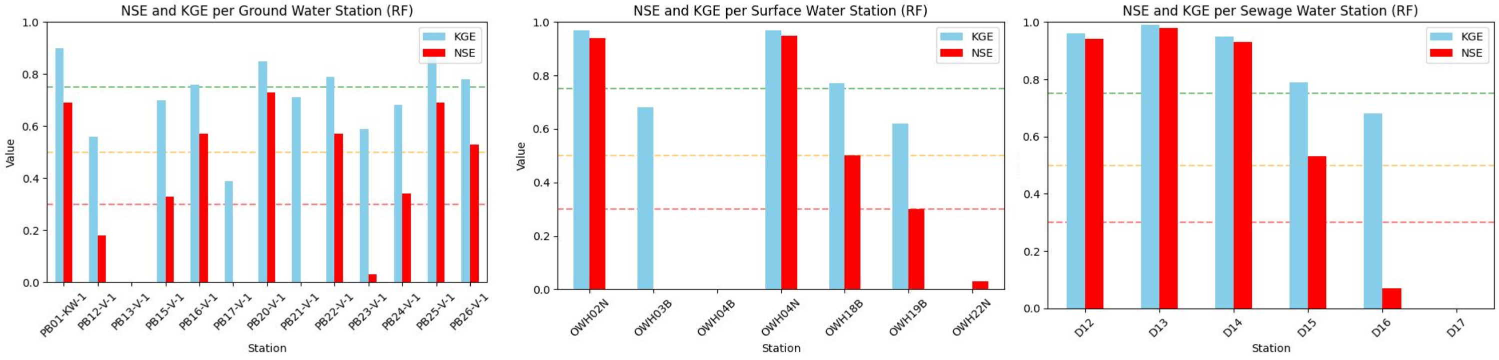


Figure 37, an overview of the obtained infilling results of the RF models per station (left: groundwater, center: surface water, right: sewage water).

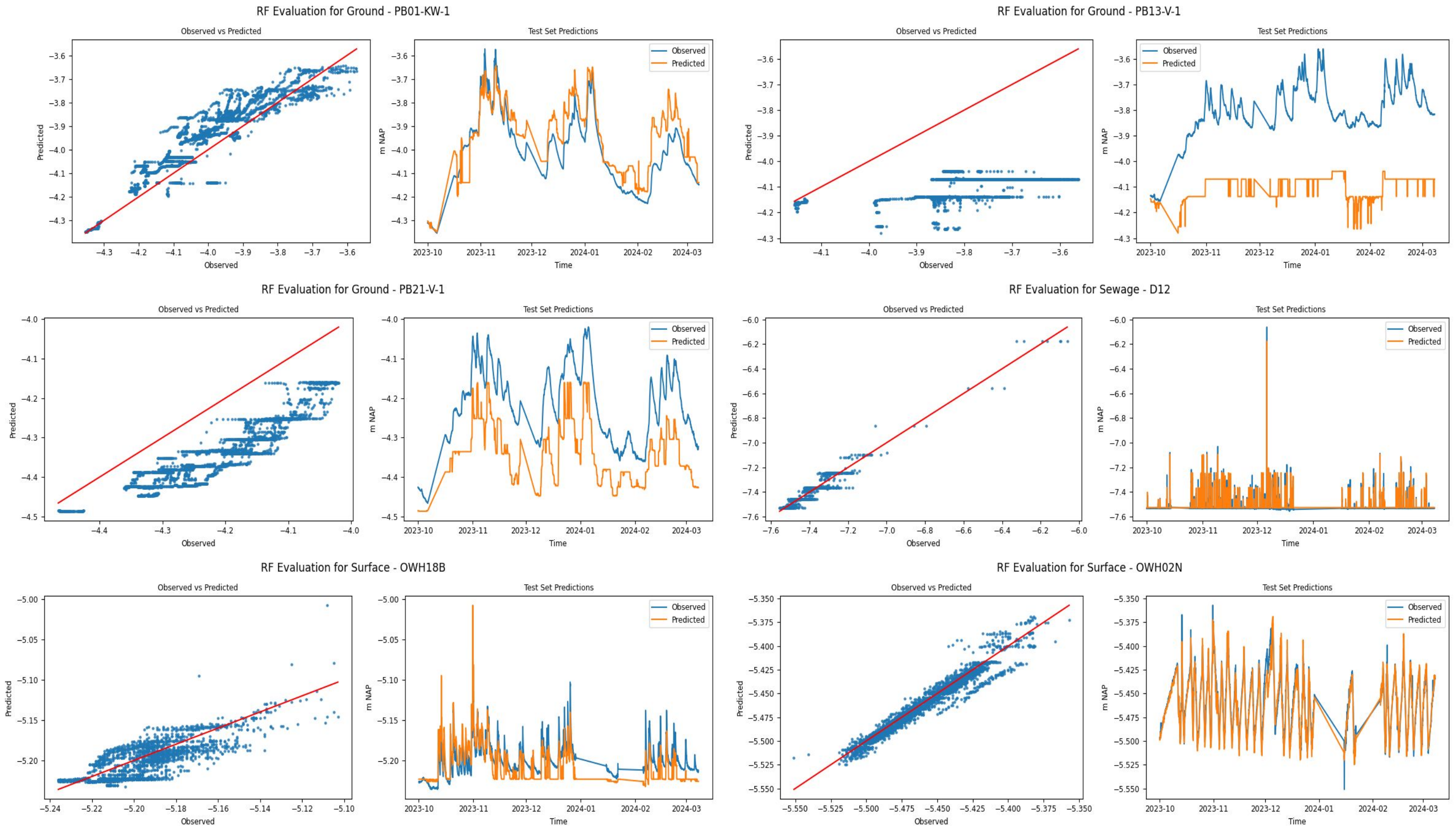


Figure 38, a selection of infilling results obtained by some RF models.

5.2.3 Gradient Boosting Trees (GBT) Results

Figure 39 shows the infilling performance of the GBT models. Again, infilling was mostly good, apart from a few underperforming stations. Similarly to RF, GBT tends to score higher on KGE than NSE. When comparing the GBT results to RF, it can be seen that for some stations GBT outperforms, while for other RF performs better. This is also evident from their similar average scores as seen in Figure 35 and Figure 36.

Zooming in at the GBT infilling's performance, it can be seen in Figure 40 that the model is able to handle the variance of the observed water levels, showing good flexibility. This is also confirmed with good KGE scores. Additionally, even with poor NSE scores (PB21-V-1, center left in Figure 38), it still achieved reasonable infilling performance as its estimates are of only a few decimeters at most.

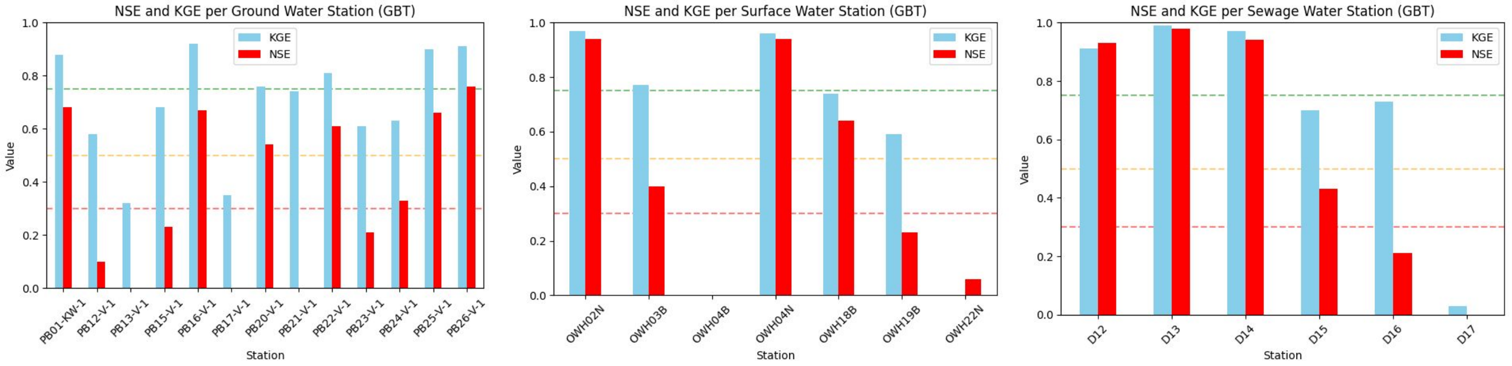


Figure 39, an overview of the obtained infilling results of the GBT models per station (left: groundwater, center: surface water, right: sewage water).

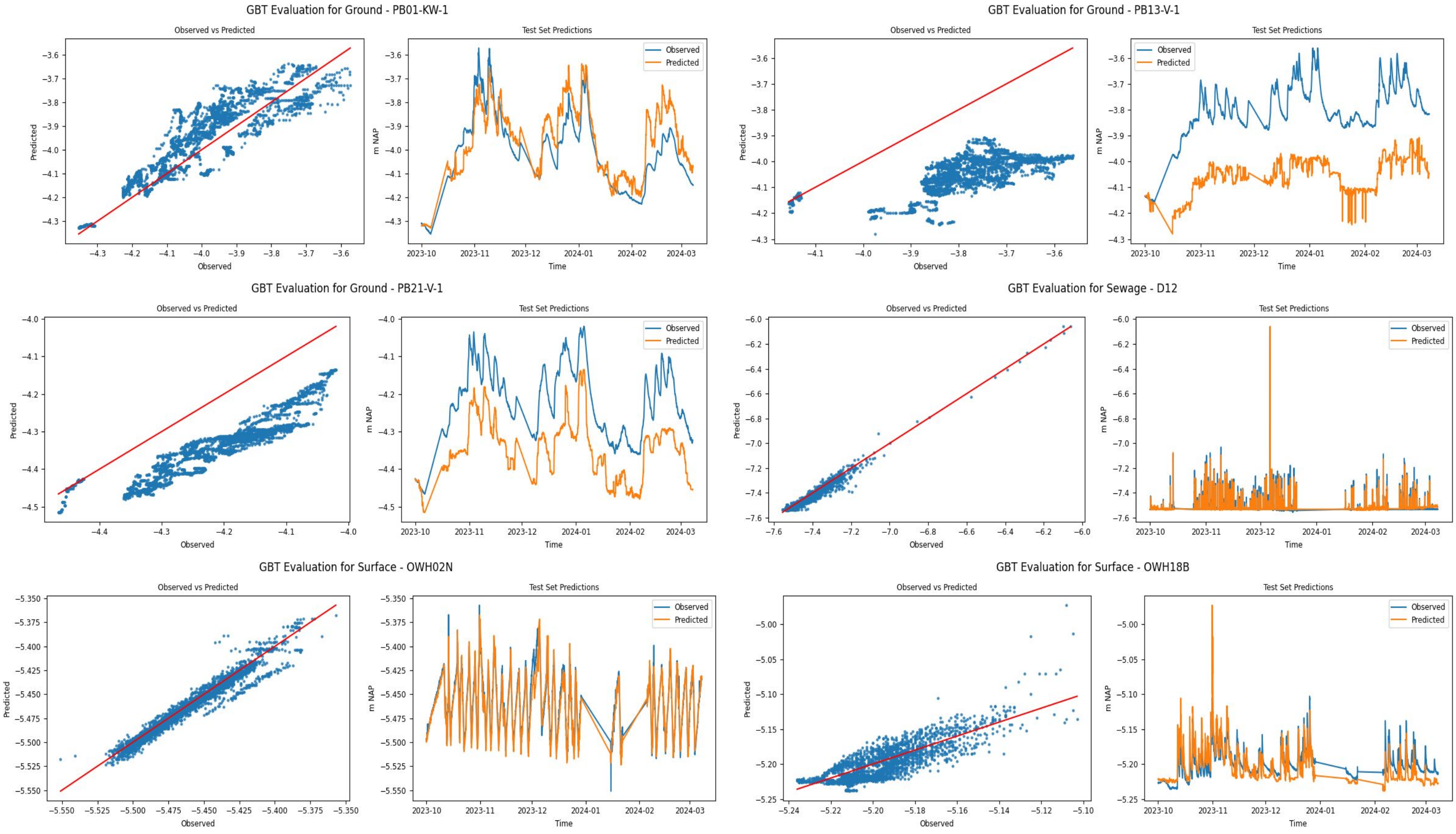


Figure 40, a selection of results obtained by some GBT models.

5.2.4 Multi-Layer Perceptron (MLP) Results

In Figure 41, it can be seen that the MLP models are a lot more inconsistent than both RF and GBT. The MLP models have bad infilling performance for more stations, which was also evident from Figure 36, given the bigger range for KGE and NSE, as well as more outliers when looking at MSE. However, it does also perform very well for other stations.

Figure 42 shows a set of infilling performances of the MLP models. It can be seen that the model produces very flexible time series. What also stands out is that the model seems to be range bound (PB21-V-1), and thus potentially has limited ability to extrapolate. Furthermore, the MLP model seems to be less accurate at times, with residuals occasionally rising above half a metre.

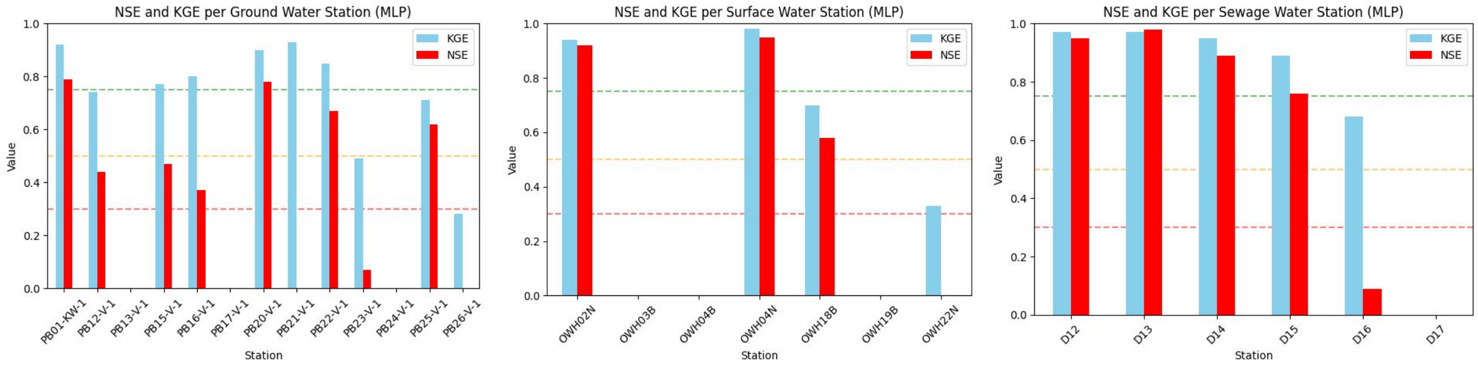


Figure 41, an overview of the obtained infilling results of the MLP models per station (left: groundwater, center: surface water, right: sewage water).

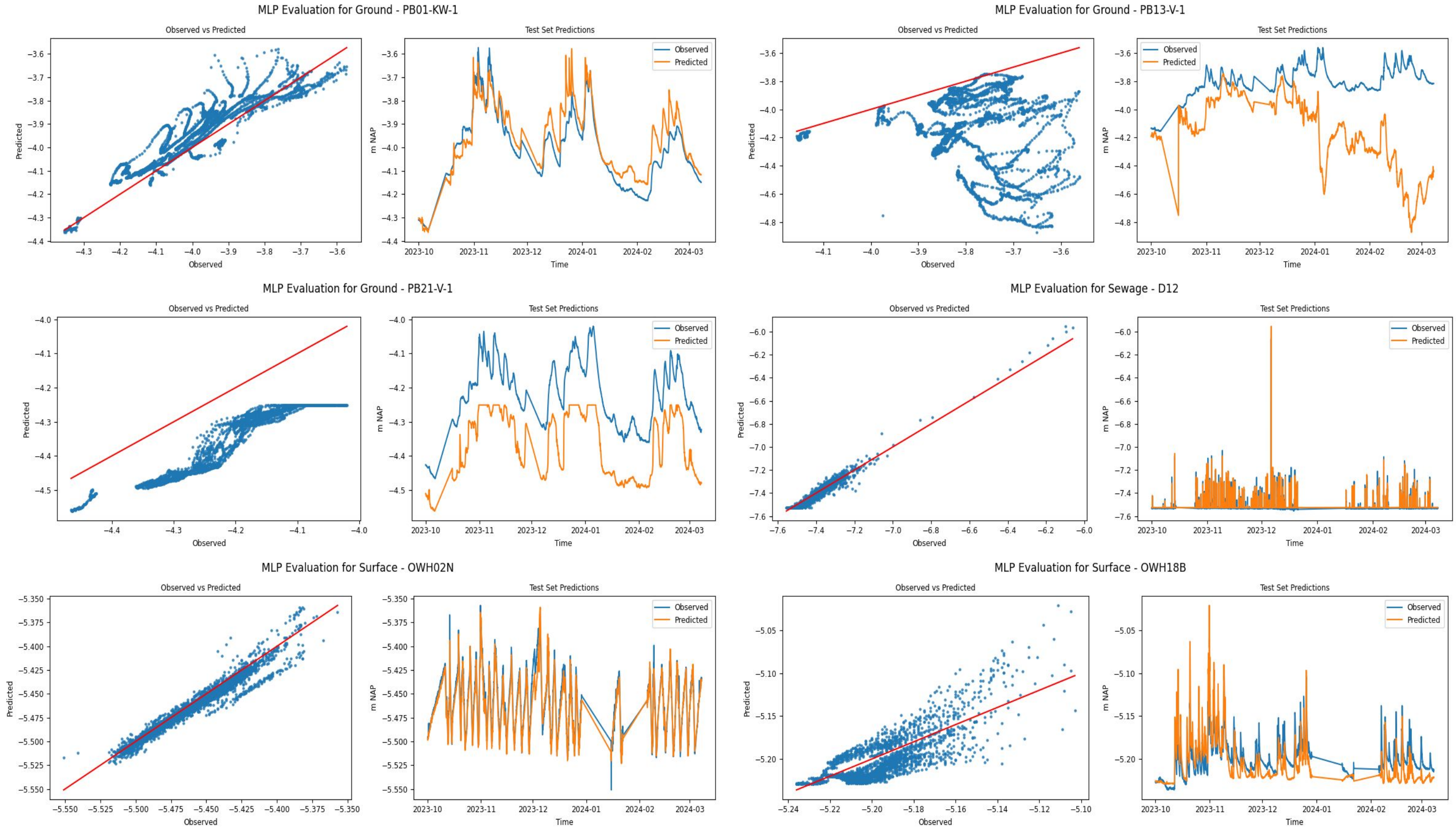


Figure 42, a selection of results obtained by some MLP models.

5.2.5 Long Short Term Memory (LSTM) Results

The LSTM results paint a similar picture as the MLP results as seen in Figure 43, although the LSTMs' performances are slightly more robust. This also explains the smaller range for KGE in Figure 35.

Figure 44 shows that the LSTM models also show great flexibility. The LSTM model for PB21-V-1 even manages to do a great job, where the other models were struggling with accurate estimates. However, similar to the MLP model for PB13-V-1, its estimates can be off by more than half a metre.

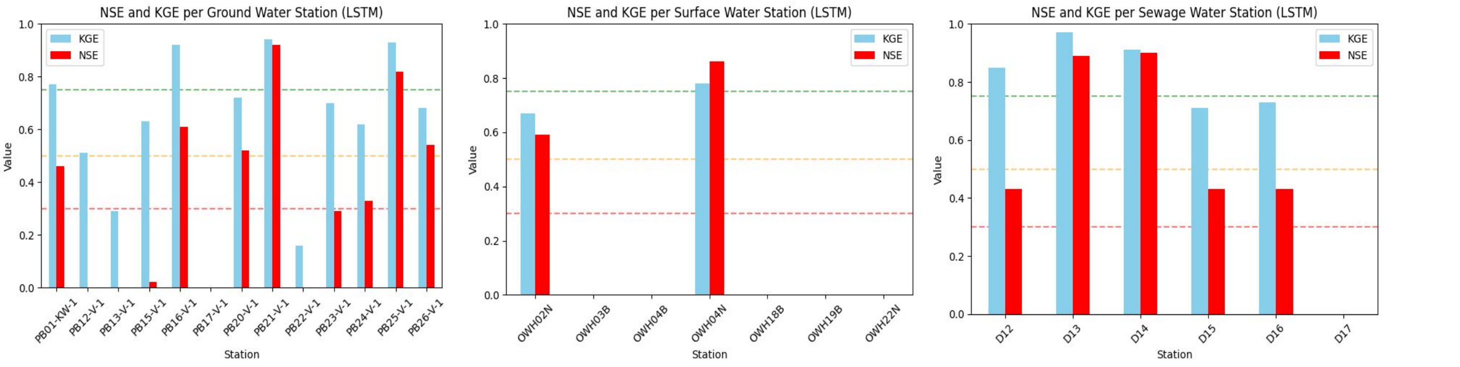


Figure 43, an overview of the obtained infilling results of the LSTM models per station (left: groundwater, center: surface water, right: sewage water).

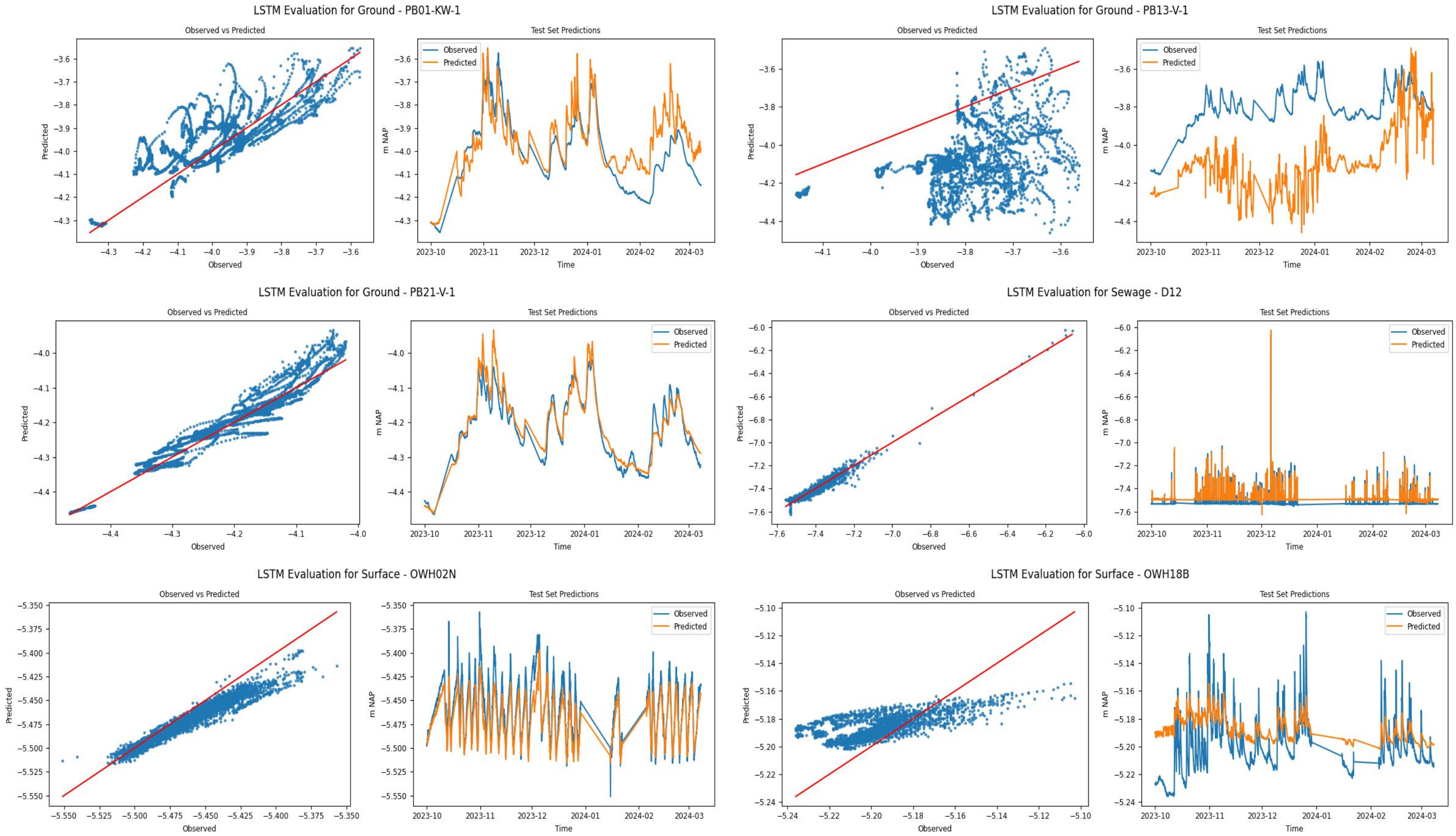


Figure 44, a selection of results obtained by some LSTM models.

5.3 Discussion

The aim of this second case study was to improve infilling performance using the inter-station method. The results were mixed but promising nevertheless. In general, acceptable scores were obtained for MSE, NSE and KGE. Furthermore, the station-specific approach seemed to be working well as the models took into account different variables for most stations. However, infilling for some stations still proves to be challenging. This section will elaborate on these three findings.

5.3.1 Inter-Station Infilling

Starting with the inter-station method, several remarks can be made. First, the ability to estimate variance of the water level is dramatically improved compared to the use of the intra-station approach. For the majority of stations, a positive NSE or KGE is achieved. This improvement is not very surprising because the estimations are now based solely on ground truths. Hence, the models are not vulnerable to propagating earlier mistakes.

Second, pure accuracy suffers when compared to the intra-station approach. Although NSE and KGE are improved, for some models MSE suffers when compared to the intra-station method. This is also not very surprising since there will always be differences between water bodies when it comes to factors such as geology, land use and human interventions. Hence, the meteorological conditions do not have exactly similar effects and thus there are differences in water level which are difficult to estimate. Also, this reduction in accuracy could also be explained by the increase in gap length, which would be in line with the theory as discussed in section 2.2.

5.3.2 Station-Specific Approach

Next, the implementation of a station-specific feature selection procedure also gave some insights. First, this approach works very well with the inter-station method. This is unsurprising given that each station has its own unique environment and thus has different features (i.e. correlating stations, meteorological relations). As a result, the models could fully focus on the dynamics unique to that station.

Contrarily, this approach might also have affected the results. For some stations, the amount of data was presumably not enough for the weights of the neural networks (MLP and LSTM) to converge. This is especially true for the underperforming stations which, due to their complexity, might have required more training data for these models to fully grasp the dynamics.

5.3.3 Performance Challenges

An investigation into the poor results for some of the stations shows several potential explanations. The bad performance for the three groundwater stations has several potential causes. First, for these three stations the dynamics differ significantly from the water levels of the stations used as features, as can be seen in Figure 45 which shows the target variable (green) and its features (red). This difference in hydrological regimes could be caused by a difference in natural and/or human factors.

Second, in the test set, the water level has a more turbulent character than the training set. For PB21-V-1 there are also potential anomalies in the training data as by the dips around September 2022. Third, as seen in the bottom right of Figure 45, the mean water level is lower in the training set, and thus the models are biassed towards lower water level estimations. Again, this implies that the models have limited extrapolation capabilities.

The underperformance of sewage station D17 can be explained by looking at the correlation between D17 and the other sewage stations in Figure 46. It can be seen that there is little correlation. In fact, the total precipitation of the last 128 days has a higher correlation and is thus used as a feature. However, given that Almere has a separated system for sewage water, this is not a good predictor (Almere, n.d.). However, chances are that there are some defects in the sewage system, which is implied by the higher correlation between Station D17 and the precipitation and evaporation variables as compared to the other stations.

When looking at Figure 47 to see why surface water stations OWH03B, OWH04B, OWH19B are not very predictable, it stands out that the water level at all three stations behave differently compared to its features. Furthermore, for OWH22N (bottom-right) the training data is significantly smaller, and that the dynamics of this station are significantly different than the others.

Most of these performance challenges could thus be attributed to differences in train and test sets, limited correlation and too little data. This implies that the models have limited extrapolation capabilities if they underperform when there is a difference in train and test set.

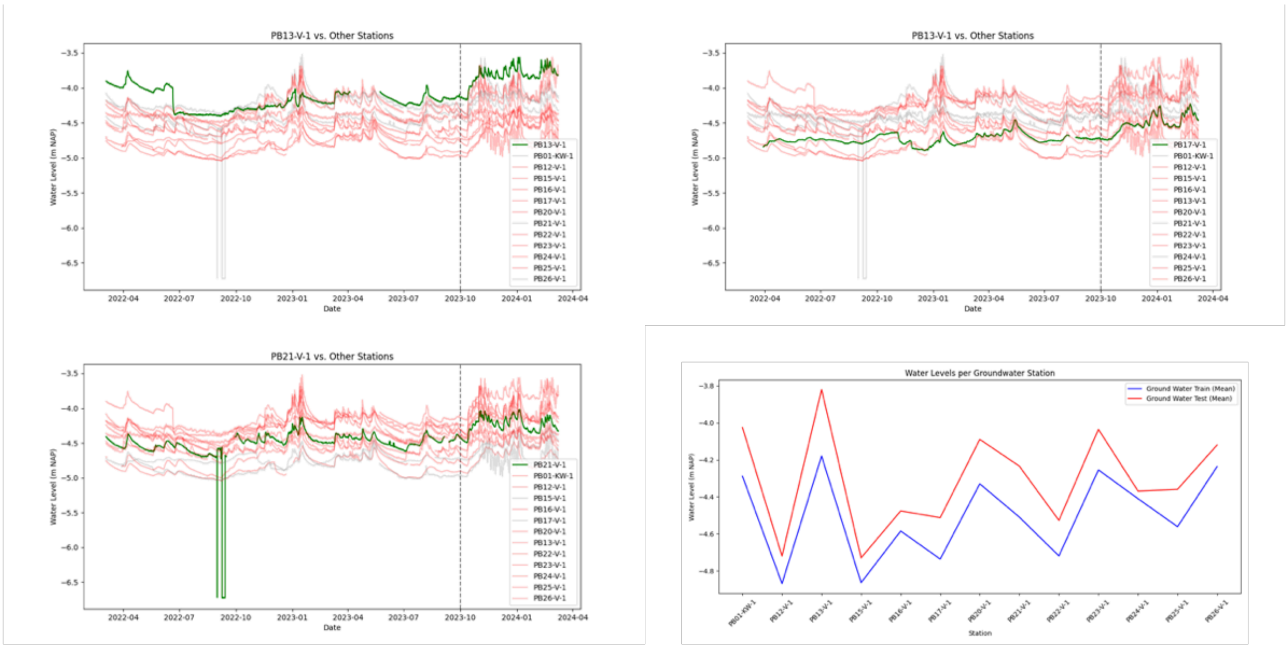


Figure 45, an overview of the underperforming stations' water level (green) compared to their features (red) and the difference in mean water level in the train and test sets (bottom-right).

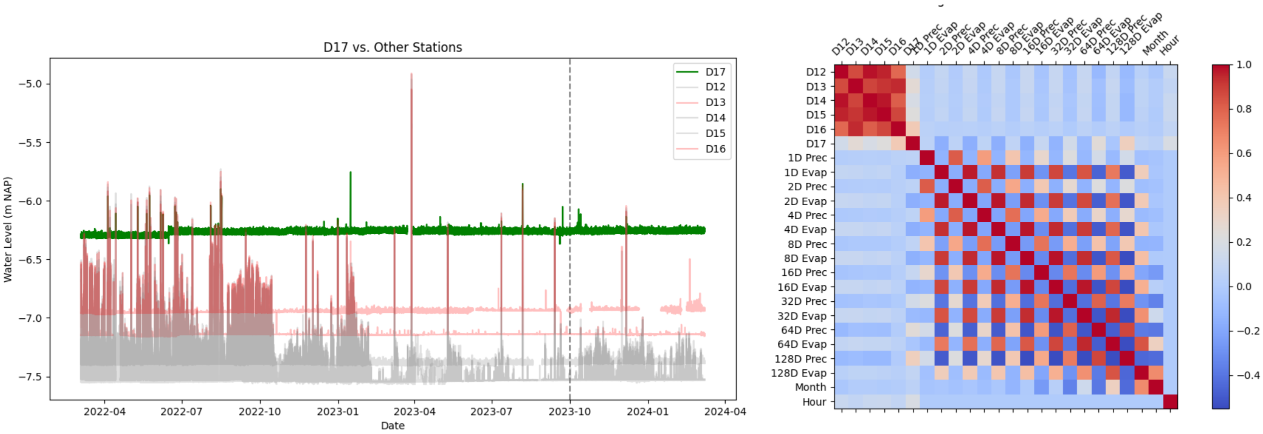
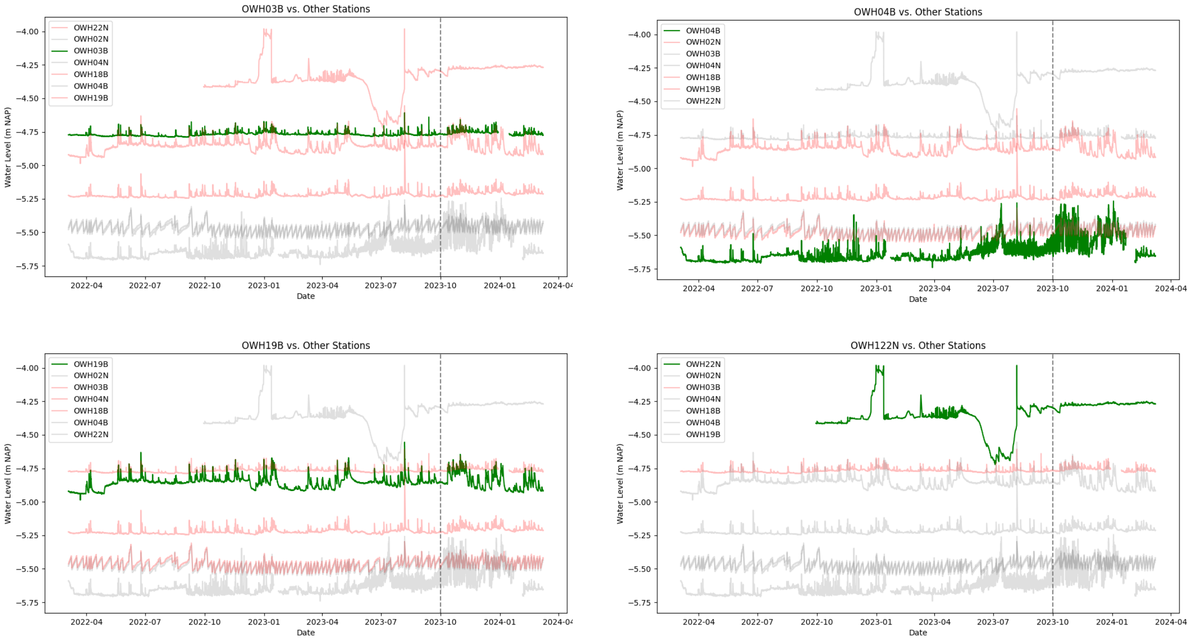


Figure 46, an overview of D17's water level compared to its features (left) and the correlation matrix for sewage water data (right).



OWH03B vs. Other Stations

OWH04B vs. Other Stations

OWH19B vs. Other Stations

OWH122N vs. Other Stations

Figure 47, an overview of the underperforming surface water station compared to their features.

6. General Discussion

Having access to complete hydrological time series data is crucial for correct monitoring and modelling of trends in our water systems as well as the effects of potential measures. However, due to failures, sensors do not record these variables continuously. As a result, hydrological time series can contain gaps, harming the quality of analysis needed for effective water management. During this research project, an investigation into the infilling of gaps in hydrological time series was performed. Several ML models and methods were explored and implemented on different water types. The results of this project enable a comparison between ML models, training methods and performance on different water types to answer the main research question. To come to these answers, several phases were completed.

First, different ML models were explored and assessed to answer the first research question: *which ML techniques exist to fill in data gaps and how do they compare?* This resulted in a long list of ML models that have been researched by other researchers. Via an MCA it was found that MLR and kNN were not sufficiently suitable for infilling hydrological time series, while SVR, RF, GBT, MLP, LSTM and SOM models were.

Next, a simple case study was done to answer the second research question: *how effective is infilling based on the use of intra-station data?* Here it was found that this method provided mixed results. When only looking at MSE for evaluation, some models (RF, GBT and LSTM) did a good job infilling randomised gaps up to a day in length while others did not. However, when looking at NSE and KGE for more information, infilling was bad because all models failed to estimate the observed variance of the water level. Hence, the scalability of this approach was probably limited. One potential explanation for the lack of (correct) variance in the infilling could be attributed to the feature importance: models relied too much upon the last measured or predicted water level. A second potential explanation was the use of a general model: the models likely did not learn the temporal dynamics unique to each station. Contrarily, the use of data of a single station could have led to insufficient amounts of data for optimal model performance, as was shown by the infilling results for surface water.

At last, a second case study was done to answer the third research question: *how effective is infilling based on the use of inter-station data?* Here, the study area was expanded and a different approach was taken. Instead of relying on a set of general models, multiple models were trained for each station and features were added based on the unique needs of each station. It was found that this led to improved NSE and KGE metrics, but a reduction in the quality of MSE scores. Caution is needed here because this reduction in precision could also be caused by the significant increase in gap length. Nevertheless, these results imply an improvement with regards to scalability when compared to the intra-station approach. The station-specific approach was successful as the models used different features for most stations. This implies that the different stations indeed benefit from a specific set of features, instead of a general set of features.

In the next sections, the results described above will be used to provide an answer to the main research question: *What ML approach(es) are most suitable for infilling gaps in hydrological time series in The Netherlands ?* As this is a very holistic question, the answer is divided into multiple sections. First, an elaboration on the different ML models is presented, taking into account findings of the MCA and the practical implementation during the case studies. Second, a comparison between the intra-station and inter-station methods is presented, as well as some commentary on the use of other variables. At last, a discussion on the performance of the models and methods on the different water types is presented.

6.1 ML Model Comparison

Synthesising the results of the MCA and case studies, several remarks can be made. There were several notable differences between the models. First, a difference between parametric and nonparametric models can be observed. Non-parametric models (RF and GBT) were more accurate when looking at MSE only. The infilling of the parametric models worsens as the gap increases when the intra-station infilling method is used. This is especially clear with the exponential behaviour of the MLP models in case study 1. Contrarily, the parametric models do a better job when looking solely at NSE and KGE (Figure 11). This difference could be caused by the more continuous outputs of the parametric models, while the nonparametric models' outputs are less continuous and flexible. As a result, the parametric models were better at estimating variance in water level of only a few centimetres.

Second, the neural network models (MLP & LSTM) perform less consistently than the RF and GBT models. One potential cause of this inconsistency is the absence of sufficient training data. When the dynamics of the water level are too complex, these models need more training data for the weights to convergence for optimal performance. However, these data requirements are not met sometimes. Hence, despite occasionally outperforming RF and GBT, the MLP and LSTM models are not very robust. Given the findings in Chapter 3, where MLP and LSTM scored significantly higher than RF and GBT, this is not in line with other research. However, a lack of data, different training methods and also less complex variations of the models were found to be potential explanations for these inconsistencies (section 4.3.1 and 4.3.2). Also, the infilling performance of the GBT and RF (when using the inter-station method) were in line with other researchers' findings.

Third, there were subtle differences in models that could be grouped together. GBT models were generally more flexible and accurate than RF models, albeit minimal. This is not a very surprising outcome given the subtle difference in methods: RF's ensemble of trees is random while GBT's ensemble is established by aiming to minimise its training error. This was also found during the MCA, strengthening the obtained results. Furthermore, while MLP and LSTM both tended to infill the gaps of Case Study 1 exponentially, the discipline of LSTM to stay within an acceptable range of the observed value could be explained by its capability to remember trends, enabled by its memory cell. However, this capability did not lead to satisfying results, contrary to other researchers as mentioned (section 3.2.5 and section 4.3.3).

There are also various similarities between the models. First, all models experience difficulties when the test set differs significantly from the training set. This suggests that all models have limited extrapolation capabilities. This is somewhat surprising, as a difference between parametric and non-parametric models was expected. It was expected that the RF and GBT models would have trouble extrapolating, since their outputs are solely based on historical data points. Same can be said for the SOM models. Contrarily, since the outputs of SVR, MLP, and LSTM are derived from a mathematical equation, some extrapolation ability was expected for these parametric models. The limited ability to extrapolate suggests that these methods may not be effective in predicting rare extreme weather events, as data on such events is often scarce due to their infrequency or gaps caused by the weather itself (see Textbox C).

Second, the performance of RF and GBT were unsurprisingly similar, despite small differences in training methods as discussed above. As mentioned (section 3.3.2), this is probably due to their reliance upon an ensemble of decision trees. This is inline with the findings in the MCA (section 3.2.4). Furthermore, both MLP and LSTM generally encountered troubles at the same stations. Since both of these models use neural networks as foundation, this is no surprise. This underperformance could have been caused by insufficient amounts of training data.

At last, all models showed similar changes in behaviour considering intra-station and inter-station methods. All four models showed improvement in NSE and KGE scores, while scoring slightly worse for MSE when the switch was made in the second case study. This implies that not only model selection is important, but feature selection is at least as important.

When comparing general and specialised models, several things are worth highlighting. First, a general training method does not work well. When a model is trained on several stations, it fails to learn the temporal variances unique to each station when using the intra-station method. Additionally, for the inter-station method a general training method is not able to take into account the unique set of correlating stations, affecting the (computational) performance of the models.

However, the use of a general model can potentially help to solve problems related to data scarcity. By bundling station data into one dataset, more data can be fed into the models. However, when doing this with the inter-station method, all other stations need to be taken into account and serve as features, instead of optimising for a subset of features. Hence, there are more features for the model to handle, leading to bigger mathematical formulas or decision trees and thus more computational load.

6.2 Systematic Infilling Methods

Making a comparison between the different feature selection methods is an equally important aspect of answering the research question. Some notable differences between the different methods can be detected. First, the inter-station method seems to be more robust when it comes to increasing gap size. There are several potential explanations for this. First, the model is better in estimating variance, implying that it can be more scalable. Furthermore, this method relies upon ground truths recorded at other stations. This method does not rely upon earlier predictions. Hence, it is less prone to propagate earlier mistakes in estimation and thus better at managing the uncertainty that rises as gaps become bigger. Furthermore, this method is able to detect variance in water levels at other (correlating) stations, and is able to translate this to the target station. On the contrary, the intra-station method does not have this information and relies heavily upon the previous water level, making it difficult to estimate variance as previously discussed.

Second, as expected, the intra-station is more robust with multivariate gaps due to its independence. When looking at the infilling performance of the models in case study 2 (Figures 39, 41, 43 and 45), it can be seen that there are still some gaps. These gaps occurred due to the inability of some of the models (MLP and LSTM) to handle missing variables of input data. Hence, the choice was made to delete all missing values as this offered a fairer comparison between the models in case study 2. Otherwise, RF and GBT could have had more room for either over- or underperformance. Hence, this weakness of the inter-station method can be overcome by some models.

Third, the intra-station method is, for some models, more accurate when looking at pure accuracy (MSE) for small gaps, smaller than a day. Despite low NSE and KGE scores, the MSE scores are very good as the estimates are very close to the observed values. This is especially clear when comparing the Q-plots of the different methods. A possible explanation is that water level is continuous and thus generally does not change abruptly in between measurements. Hence, using the last measured (or estimated) water level as the dominant predictor results in a low fault margin, generally, as the next measurement is usually not far off. Since there is always some difference in water level between stations, unless they measure the same water body, it is difficult to achieve this accuracy with the inter-station approach.

Despite these differences, some similarities were also seen. For example, both methods do not compensate for training data shortages. WIth both models, insufficient amounts of data led to suboptimal performance of the neural network models. This comes as no surprise, although it could have been true that some set of features could compensate for a lack of training data by making the problem less complex for the models. Furthermore, both approaches were vulnerable to differences in train and test sets and thus did not enable extrapolation capabilities. This is in line with ML theory, which states that the quality of the input data needs to be high.

At last, when looking at the effectiveness of using meteorological features to take into account some aspects of the hydrological regime, the results are mixed. For some stations and models, a meteorological feature did help in estimations, whereas for others it did not. For the intra-station method, this is not surprising given the heavy reliance upon the last measured or estimated water level. For the inter-station method, this is also not surprising when looking at the method used for feature selection. In Figure 48, it can be seen that, except for ground water, there is little correlation between water level and the meteorological variables used. Hence, most meteorological variables did not make it to the selected feature set. When selected, the models mostly gravitated towards the highest correlating feature since this gives the highest probability of an accurate estimation. Hence, as expected, the effectiveness of these variables as features relies heavily upon the characteristics of the hydrologic regime.

This low level of correlation between the meteorological data and water level has several explanations, which will be elaborated on in the next section regarding different hydrologic regimes (6.3).

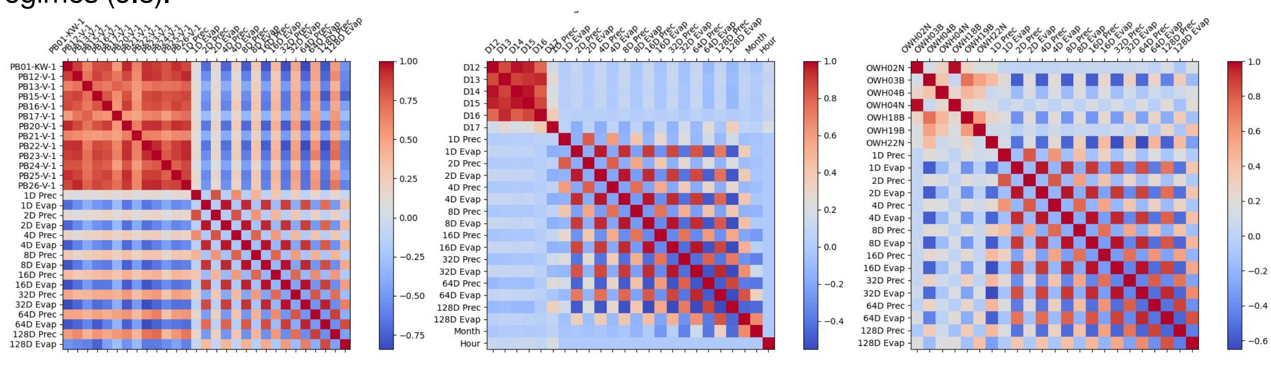


Figure 48, the correlation matrices of groundwater (left), sewage water (middle) and surface water (right).

6.3 Comparison of Infilling Performance for Hydrologic Regimes

When looking at the performance across water types, several observations can be made. First, the best infilling performance was achieved on sewage water data. This could be explained by the fact that Almere has a separated sewage system for wastewaters. Hence, the system is solely influenced by human activity. This makes it more predictable as people tend to be animals of habit. For example, people might shower at approximately the same time everyday. Second, the sewage system is designed by humans. Hence, the dynamics of the sewage water level are somewhat controlled. Furthermore, there is strong correlation between the different sewage stations (Figure 49), possibly also an effect of the design of the sewage system. This strong correlation is especially helpful for inter-station infilling.

Second, the worst infilling performance was achieved on surface water data. A potential cause for this is that the surface water levels of Almere are highly regulated. The city has a total of five sluices (Gemeente Almere, n.d.). Looking at the low correlation between the stations, it is likely that the specifics of this water level regulation differ per station, or water bodies reach local minima and maxima at different times. Hence, it is difficult to use inter-station infilling. However, it is somewhat surprising that the results of using the LSTM in the first case study did not lead to significant infilling accuracy as measured by NSE and KGE. The memory cell of the LSTM should have enabled it to learn the local minima and maxima at which the water level would rise or fall. However, insufficient data was most likely the cause of this underperformance, as discussed.

At last, for groundwater it is no surprise that the meteorological variables are taken into account more often during feature selection in the second case study. These water bodies are simply the only water bodies in Almere in which the hydrological process is mostly influenced by nature itself, although the municipality keeps these water levels at acceptable levels as well (Gemeente Almere, n.d.). There is little spatiotemporal difference in the precipitation and evaporation in the small study area. Additionally, it could be that there is also little difference in geological circumstances as Almere is located in Flevoland, which is a reclaimed piece of land. Hence, differences in land use could be the only factors that influence the difference in hydrologic regimes of the stations. This could explain the high correlations between the groundwater stations (Figure 48).

6.4 Limitations

Some caution is needed with these results and answers to the main research question. The limitations that applied to the results of the individual research questions and methods have already been mentioned in the corresponding chapters. However, some word of caution is also needed for the overall discussion as presented in sections 6.1 to 6.3. First, the validity of the comparison between the intra-station and inter-station method is hurt by the use of general and specialised models. Hence, the comparison between both infilling methods is not one-on-one. As discussed, the intra-station infilling results might have been better if models were trained on a single station. Nevertheless, the A/B test (general models for ground water and sewage water vs specialised model for surface water) in case study 1 did not show much strength for this argument.

Second, caution is needed when comparing both infilling methods' performance when it comes to gap length. Although it seems reasonable to say that the intra-station method will have difficulty with infilling larger gaps due to its failure to estimate variance, this project cannot back it up with proof. Similarly, infilling accuracy using the inter-station method might have different performance on smaller gaps. This is probable because, when looking at the infilling results, in some time intervals the infilled series is more accurate than in other intervals. Hence, the validity and reliability of these results are hurt.

Third, infilling performance of the models could be negatively impacted by the approach of splitting the data into train and test sets, hurting the validity. Here, a training set and a test set were created by artificially splitting the data into two at roughly two thirds of the data. However, in practice, the models could be trained on the whole dataset (except the gaps and a small portion for testing purposes). Hence, more training data is available, potentially enhancing the performance of the models. Furthermore, this improves the quality of the training data because the training data covers the data better since it is less prone to seasonal changes, interventions, etc. Especially the stations where data was insufficient or the test set was considerably different than the training set, performance could be improved dramatically by this change.

At last, infilling results could be improved through the use of more complex architectures of the models. For example, the MLP models could have shown increased performance as more hidden layers and nodes were added. Similarly, the LSTM model could have been more complex by adding more cells, as mentioned in section 4.3.3. However, due to limitations in time and compute resources, such complexity was not pursued as it would probably increase training and inference times significantly.

7. Conclusion

7.1 Conclusion

As pressures on our water systems are rising, effective water management becomes increasingly crucial. Effective water management is achieved by closely monitoring and modelling (potential) measures. Statistical analysis and model calibrations that support decision making are most reliable when data sets are complete. However, due to failing sensors, hydrological time series can contain gaps. Hence, monitoring and modelling becomes unreliable, hindering effective decision making.

Infilling hydrological time series is a complex problem since the hydrological process is highly complex. Additionally, infilling difficulties are proportional to gap length and the number of missing variables. Hence, traditional statistical tools are ineffective. Due to its ability to recognise complex relations and patterns in data, ML has the potential to fill this void. Many ML models and methods have been explored for infilling hydrological variables. However, a comprehensive comparison of these models and methods was lacking. Additionally, little research was available on the infilling performance of such models and methods on different hydrologic regimes.

Therefore, in this research project, ML models and methods were explored on different hydrologic regimes (groundwater, sewage water and surface water) in Almere, The Netherlands. An MCA was executed to answer the first research question: *which ML techniques exist to fill in data gaps and how do they compare?* By taking into account accuracy, scalability, data requirements and computational requirements, it was found that MLR and kNN were insufficiently suitable. Contrarily, the SVR, RF, GBT, MLP, LSTM and SOM models were sufficiently suitable.

These six models were implemented in two case studies using the intra-station and inter-station methods. The first case study focused on infilling water level data for groundwater, sewage water and surface water using the intra-station method. During this case study the second research question was answered: *how effective is infilling based on the use of intra-station data?* The models were trained on data from multiple stations for groundwater and sewage water, whereas the models were trained on one station's data for surface water. The results were mixed. When looking at MSE scores, RF and GBT models scored very high while the other model obtained insufficient accuracy, especially the MLP. When looking solely at NSE and KGE, the results were bad across the board. Hence, intra-station infilling can lead to precise estimates but fails to predict variance of the water level. This was mainly due to the high feature importance of the last recorded water level measurement. Additionally, the use of a general model presumably prevented the models from learning the temporal dynamics unique to each station. However, caution was needed as the results of the specialised model for surface water showed no difference. Furthermore, the SVR and SOM models were discarded due to their long training and inference times, respectively.

During a second case study, the third research question was answered: *how effective is infilling based on the use of inter-station data?* In this case study a station-specific approach was taken. For each station, a unique set of features was selected. It was found that the pure accuracy of infilling was worse, looking at MSE. However, this method did a better job at predicting variance as was evident from the significant improvement in NSE and KGE scores. Additionally, the inter-station method is likely to be more scalable as gap length increases, given its performance on a test set of seven months. It was also found that RF and GBT models produced the most consistent results, whereas MLP and LSTM showed unreliability. Chances are, this unreliability was caused by insufficient amounts of training data. Acceptable results were not obtained for all stations. The underperformance of these stations could likely be attributable to differences in train and test sets, implying limited extrapolation powers from the models, and insufficiently correlating stations.

These results helped to answer the main research question: *What ML approach(es) are most suitable for infilling gaps in hydrological time series in The Netherlands?* First, the most reliable and robust models are the RF and GBT, with the GBT slightly outperforming due to its greater flexibility. These models are also able to handle multivariate gaps. Despite high accuracy, MLP and LSTM models are too unreliable due to their hunger for data. SVR and SOM models are not user friendly due to their computational needs. At last, MLR and kNN models are not able to find the complexities of the hydrological process.

Second, both intra-station and inter-station infilling methods can lead to success, albeit in different circumstances. Intra-station is most suitable for filling small gaps, which only have little variance. However, inter-station infilling is more robust as gap size increases and there is more variance in the data. The use of meteorological variables is only helpful when the hydrologic regime of the water level is meaningfully affected by it. When water bodies are only limitedly influenced by these natural factors, or are heavily influenced by human activity, these variables are unlikely to move the needle in terms of accuracy.

Third, the development of a set of specialised models generally works better than one generalised model. A general model generalises too much as it fails to take into account the spatiotemporal variances resulting from the different hydrological regimes. A specialised approach, mixed with the inter-station approach, enables models to take into account these different spatiotemporal variances, while also utilising spatiotemporal similarities.

At last, it was concluded that all models and methods suffer from poor extrapolation capabilities. Hence, their applications to, for example, gaps caused by rare extreme weather events is limited, as is illustrated by the example of Sammerspolder.

Sammerspolder Example. To illustrate the limitations of the limited extrapolation powers of ML models, take a look at Sammerspolder again. When using the a GBT model and inter-station infilling, plus randomly splitting the train and test set, infilling the water level or Sammerspolder obtains disappointing results. Hence, it should be noted that ML is not the holy grail of infilling and more manual methods can also work.

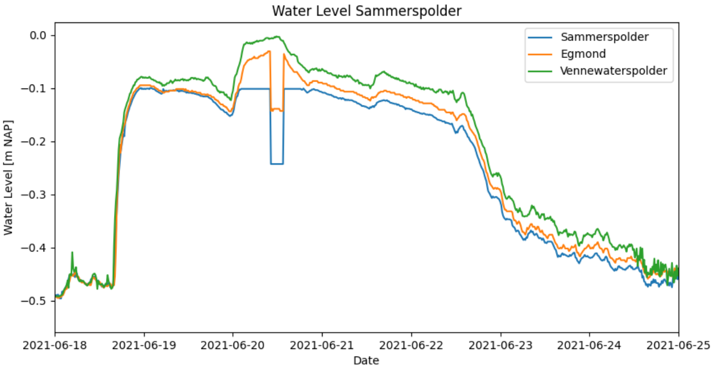


Figure D, result of infilling the gap in Sammerspolder.

7.2 Recommendations

These findings result in the following recommendations with regards to the practical implementation and potential direction for future research.

Implementation

With regards to implementation, it is recommended to use the RF or GBT model. These models have proven to be effective and consistent, while also being able to handle multivariate gaps. Furthermore, it is recommended to use the inter-station infilling method as this approach is more scalable with increasing gap length.

Furthermore, it is recommended to build models that are tailored to the unique circumstances of each station, instead of building a one-size-fits-all model. This enables the models to learn the temporal dynamics unique to each station, solving the problem of spatiotemporal variance between stations. Hence, it is recommended to establish a general method that can be applied in myriad locations.

The intra-station approach could still be valuable. Models based on this approach can help to detect anomalies in real-time. Hence, they can help to reduce downtime by alarming operators when sensor failure is suspected. As a result, the gaps in hydrological data can become smaller and potentially easier to infill.

Future Research Directions

There are several pathways for future research, including both improvements on this project and improvement in infilling accuracy. First, to improve the findings of this project, it should be researched whether it is indeed true that infilling accuracy can be improved by a more realistic method of splitting training and testing data. As mentioned, this could result in a bigger and more representative training set.

Second, for a more robust comparison between intra-station and inter-station infilling, both methods should be tested with similar sets of different gap lengths. During this research, results suggested that intra-infilling was more suitable for small gaps, while inter-station infilling is more suitable for large gaps. Comparing the performance of both methods on different gap lengths would provide better insight into the tradeoff between these models.

Third, an expansion of this comparison could be made by comparing the infilling performance of the different ML models and methods to different sensor failures. As concluded in this thesis, the ML models have limited extrapolation capabilities. Hence, they are likely to be less effective when extreme weather events cause a gap. Contrarily, infilling with ML models might be more effective when a sensor cannot upload its data due to a random failure in uploading its data caused by, for example, an internet outage. It would be helpful to have more insight into these applications of ML models.

There are also promising pathways that could take the whole infilling research domain to a further level. further research is needed to enhance the extrapolative capabilities of the models. While the infilling results obtained in this project were acceptable, the models exhibited limited extrapolation power, which is concerned for monitoring and modelling extreme weather events. Such events can cause inaccuracies in measurement equipment and records, as shown in the Sammerspolder example. Moreover, the infrequency of these events means the models have limited data to learn from, resulting in difficulties handling them. Unfortunately, these extreme events are the most critical to understand and manage, as effective interventions must be capable of addressing these anomalies.

Second, to further investigate infilling accuracy, the potential of mixing the intra-station and inter-station methods should be researched. Intra-station can lead to highly accurate estimations but low variance prediction and scalability. Contrarily, inter-station infilling leads to slightly less accurate predictions but good estimation of variance and scalability. Hence, combining the two methods could cover both methods' weaknesses.

Third, more research into the amount of data needed for the neural network models to perform effectively is needed. Or potential methods to reduce this amount of data. This would greatly help the implementation of neural networks, which despite their unreliability still have enormous potential.

References

Aghelpour, P., & Varshavian, V. (2020). Evaluation of stochastic and artificial intelligence models in modeling and predicting of river daily flow time series. Stochastic Environmental Research and Risk Assessment, 34(1), 33-50.

Aho, T., Sievi-Korte, O., Kilamo, T., Yaman, S., & Mikkonen, T. (2020). Demystifying data science projects: A look on the people and process of data science today. In Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21 (pp. 153-167). Springer International Publishing.

Albuquerque, M. B., de Araújo, A. A., Medina Martinez, C. E. N., Mauad, F. F., & Okawa, C. M. P. (2019). Sustainable Urban Drainage: a brief review of the compensatory techniques of structural and non-structural measures. Revista Eletrônica em Gestão, Educação e Tecnologia Ambiental, 23.

Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications, 71, 257-278

Alley, W. (2009). Encyclopedia of inland waters. Encycl. Inl. Waters, 684-690.

Alsdorf, D. E., Rodríguez, E., & Lettenmaier, D. P. (2007). Measuring surface water from space. Reviews of Geophysics, 45(2).

Andreasen, M. H., Agergaard, J., Møller-Jensen, L., Oteng-Ababio, M., & Yiran, G. A. B. (2022). Mobility disruptions in Accra: Recurrent flooding, fragile infrastructure and climate change. Sustainability, 14(21), 13790.

Arriagada, P., Karelovic, B., & Link, O. (2021). Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. Journal of Hydrology, 598, 126454.

Bartholomeus, R. P., van der Wiel, K., van Loon, A. F., van Huijgevoort, M. H., van Vliet, M. T., Mens, M., ... & Pot, W. (2023). Managing water across the flood–drought spectrum: Experiences from and challenges for the Netherlands. Cambridge Prisms: Water, 1, e2.

Bhatnagar, P. (2019). Data Science for Marketing Analytics | Data | eBook. Packt. https://www.packtpub.com/en-mt/product/data-science-for-marketing-analytics-9781789959413

Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. Wiley Interdisciplinary Reviews: Water, 8(3), e1520.

Chan, N. W., Ghani, A. A., Samat, N., Hasan, N. N. N., & Tan, M. L. (2020). Integrating structural and non-structural flood management measures for greater effectiveness in flood loss reduction in the Kelantan River basin, Malaysia. In Proceedings of AICCE'19: Transforming the Nation for a Sustainable Tomorrow 4 (pp. 1151-1162). Springer International Publishing.

Collins, A. J. (1991). Modern methods of data analysis.

Contractor, S., & Roughan, M. (2021). Efficacy of Feedforward and LSTM Neural Networks at predicting and gap filling coastal ocean timeseries: Oxygen, nutrients, and temperature. Frontiers in Marine Science, 8, 637759.

Coutinho, E. R., Silva, R. M. D., Madeira, J. G. F., Coutinho, P. R. D. O. D. S., Boloy, R. A. M., & Delgado, A. R. S. (2018). Application of artificial neural networks (ANNs) in the gap filling of meteorological time series. Revista Brasileira de Meteorologia, 33, 317-328.

Dąbrowska, J., Orellana, A. E. M., Kilian, W., Moryl, A., Cielecka, N., Michałowska, K., ... & Włóka, A. (2023). Between flood and drought: How cities are facing water surplus and scarcity. Journal of Environmental Management, 345, 118557.

Dahmani, S., & Latif, S. D. (2024). Streamflow Data Infilling Using Machine Learning Techniques with Gamma Test. Water Resources Management, 38(2), 701-716.

Dastorani, M. T., Moghadamnia, A., Piri, J., & Rico-Ramirez, M. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. Environmental monitoring and assessment, 166, 421-434.

De Veaux, R. D., Hoerl, R. W., & Snee, R. D. (2016). Big data and the missing links. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(6), 411-416.

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G., & Schaefli, B. (2019). Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. Journal of Hydrology, 569, 573-586.

Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A review on hydrological models. Aquatic procedia, 4, 1001-1007.

Dolman, N. (2021). Integration of water management and urban design for climate resilient cities. Climate Resilient Urban Areas: Governance, design and development in coastal delta cities, 21-43.

Dong, G., & Liu, H. (Eds.). (2018). Feature engineering for machine learning and data analytics. CRC press.

Fankhauser, S. (2010). The costs of adaptation. Wiley interdisciplinary reviews: climate change, 1(1), 23-30.

Fitts, C. R. (2002). Groundwater science. Elsevier.

Folguera, L., Zupan, J., Cicerone, D., & Magallanes, J. F. (2015). Self-organizing maps for imputation of missing data in incomplete data matrices. Chemometrics and Intelligent Laboratory Systems, 143, 146-151.

Fu, T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1), 164-181.

Gemeente Almere (n.d.). Water. Retrieved on July 7, 2024, from https://www.almere.nl/wonen/beheer-en-onderhoud/water

Gemeente Almere (n.d.). Riolering Retrieved on July 9, 2024 from https://www.almere.nl/wonen/beheer-en-onderhoud/water/riolering#:~:text=In%20Almere%20hebben%20we%20een,oppervlaktewater%2C%20zoals%20meren%20en%20sloten.

Hamilton, H., Henry, R., Rounsevell, M., Moran, D., Cossar, F., Allen, K., ... & Alexander, P. (2020). Exploring global food system shocks, scenarios and outcomes. Futures, 123, 102601.

Hamzah, F. B., Mohd Hamzah, F., Mohd Razali, S. F., Jaafar, O., & Abdul Jamil, N. (2020). Imputation methods for recovering streamflow observation: A methodological review. Cogent Environmental Science, 6(1), 1745133.

Harvey, C. L., Dixon, H., & Hannaford, J. (2012). An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. Hydrology Research, 43(5), 618-636.

He, L., Chen, S., Liang, Y., Hou, M., & Chen, J. (2020). Infilling the missing values of groundwater level using time and space series: case of Nantong City, east coast of China. Earth Science Informatics, 13, 1445-1459.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Janbain, I., Deloffre, J., Jardani, A., Vu, M. T., & Massei, N. (2023). Use of long short-term memory network (LSTM) in the reconstruction of missing water level data in the River Seine. Hydrological Sciences Journal, 68(10), 1372-1390.

Kazijevs, M., & Samad, M. D. (2023). Deep Imputation of Missing Values in Time Series Health Data: A Review with Benchmarking. arXiv preprint arXiv:2302.10902.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing data science: A framework and case study. Harvard Data Science Review, 2(1).

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., & Cho, K. H. (2015). Comparative studies of different imputation methods for recovering streamflow observation. Water, 7(12), 6847-6860.

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), 4323-4331.

Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. Educational and psychological measurement, 68(3), 431-442.

Körner, P., Kronenberg, R., Genzel, S., & Bernhofer, C. (2018). Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. Meteorologische Zeitschrift, 27(5), 369-376.

Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Wongchaisuwat, P., Wimala, S., Sarinnapakorn, K., & Boonya-Aroonnet, S. (2021). Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. Water, 13(13), 1862.

Leira, M., & Cantonati, M. (2008). Effects of water-level fluctuations on lakes: an annotated bibliography. Ecological effects of water-level fluctuations in lakes, 171-184.

Longman, R. J., Newman, A. J., Giambelluca, T. W., & Lucas, M. (2020). Characterizing the uncertainty and assessing the value of gap-filled daily rainfall data in Hawaii. Journal of Applied Meteorology and Climatology, 59(7), 1261-1276.

Luo, L., & Wood, E. F. (2007). Monitoring and predicting the 2007 US drought. Geophysical Research Letters, 34(22).

Mackay, S. J., Arthington, A. H., & James, C. S. (2014). Classification and comparison of natural and altered flow regimes to support an Australian trial of the Ecological Limits of Hydrologic Alteration framework. Ecohydrology, 7(6), 1485-1507.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

McGregor, G. R. (2019). Climate and rivers. River Research and Applications, 35(8), 1119-1140.

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. Wiley Interdisciplinary Reviews: Water, 5(6), e1319.

Moradzadeh, Arash & Mansour Saatloo, Amin & Mohammadi-ivatloo, Behnam & Anvari-Moghaddam, Amjad. (2020). Performance Evaluation of Two Machine Learning Techniques in Heating and Cooling Loads Forecasting of Residential Buildings. Applied Sciences. 10. 10.3390/app10113829.

Mounce, S. R., Mounce, R. B., & Boxall, J. B. (2011). Novelty detection for time series data analysis in water distribution systems using support vector machines. Journal of hydroinformatics, 13(4), 672-686.

Myšiak, J. (2006). Consistency of the results of different MCA methods: a critical review. Environment and Planning C: Government and Policy, 24(2), 257-277.

Nanda, T., Sahoo, B., & Chatterjee, C. (2017). Enhancing the applicability of Kohonen Self-Organizing Map (KSOM) estimator for gap-filling in hydrometeorological timeseries data. Journal of Hydrology, 549, 133-147.
Nkiaka, E., Nawaz, N. R., & Lovett, J. C. (2016). Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental Monitoring and Assessment, 188, 1-12.

Nowak, B. M., & Ptak, M. (2019). Natural and anthropogenic conditions of water level fluctuations in lakes–Lake Powidzkie case study (Central-Western Poland). Journal of Water and Land Development.

Oakes, B. J., Famelis, M., & Sahraoui, H. (2022). Building Domain-Specific Machine Learning Workflows: A Conceptual Framework for the State-of-the-Practice. arXiv preprint arXiv:2203.08638.

Pagano, T., & Sorooshian, S. (2002). Hydrologic cycle. Encyclopedia of Global Environment Change.

Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., ... & Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. Neural Computing and Applications, 35(12), 9071-9091.

Pluntke, T., Pavlik, D., & Bernhofer, C. (2014). Reducing uncertainty in hydrological modelling in a data sparse region. Environmental earth sciences, 72, 4801-4816.

Portuguez-Maurtua, M., Arumi, J. L., Lagos, O., Stehr, A., & Montalvo Arquinigo, N. (2022). Filling gaps in daily precipitation series using regression and machine learning in Inter-Andean Watersheds. Water, 14(11), 1799.

Post, D. A., & Jones, J. A. (2001). Hydrologic regimes of forested, mountainous, headwater basins in New Hampshire, North Carolina, Oregon, and Puerto Rico. Advances in Water Resources, 24(9-10), 1195-1210.

Pykes, K., (July 13, 2023). A Guide to Monitoring Machine Learning Models in Production. Retrieved on February 26, 2024, from https://developer.nvidia.com/blog/a-guide-to-monitoring-machine-learning-models-in-production/

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.

Ren, H., Cromwell, E., Kravitz, B., & Chen, X. (2022). Using long short-term memory models to fill data gaps in hydrological monitoring networks. Hydrology and Earth System Sciences, 26(7), 1727-1743.

Sahoo, A., & Ghose, D. K. (2022). Imputation of missing precipitation data using KNN, SOM, RF, and FNN. Soft Computing, 26(12), 5919-5936.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Sarafanov, M., Nikitin, N. O., & Kalyuzhnaya, A. V. (2022). Automated data-driven approach for gap filling in the time series using evolutionary learning. In 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021) (pp. 633-642). Springer International Publishing.

Sharma, V., & Yuden, K. (2021). Imputing missing data in hydrology using machine learning models. Int. J. Eng. Res. Technol, 10(2021), 78-82.

Song, W., Gao, C., Zhao, Y., & Zhao, Y. (2020). A time series data filling method based on LSTM—Taking the stem moisture as an example. Sensors, 20(18), 5045.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). Using multivariate statistics (Vol. 6, pp. 497-516). Boston, MA: pearson.

Taie Semiromi, M., & Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. Hydrological sciences journal, 64(14), 1711-1726.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10(6), 363-377.

Tatachar, A. V. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. International Journal of Innovative Technology and Exploring Engineering, 8(9), 853-860.

Umar, N., & Gray, A. (2023). Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data. Water, 15(8), 1519.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8), 5929-5955.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. Hydrology and Earth System Sciences, 5(1), 13-26.

Wesström, I., Messing, I., Linner, H., & Lindström, J. (2001). Controlled drainage—effects on drain outflow and water quality. Agricultural water management, 47(2), 85-100.

Wu, R., Hamshaw, S. D., Yang, L., Kincaid, D. W., Etheridge, R., & Ghasemkhani, A. (2022). Data imputation for multivariate time series sensor data with large gaps of missing data. IEEE Sensors Journal, 22(11), 10671-10683.

Yang, T. H., & Liu, W. C. (2020). A general overview of the risk-reduction strategies for floods and droughts. Sustainability, 12(7), 2687.

Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., & Yuan, L. (2023). Llm lies: Hallucinations are not bugs, but features as adversarial examples. arXiv preprint arXiv:2310.01469.

Appendix A - Resources

Water Level Data Almere

View Envision

Coding

Visual Studio Code

Scikit-learn library for SVR, RF, GBT & MLP models, standardising and model selection and MSE scores

Torch for LSTM model (self-made)

Susi for SOM model

Pandas for data processing

Hydroeval for NSE and KGE scores

Random Seeds Case Studies

np.random.seed(0)

torch.manual_seed(1)

Appendix B - ChatGPT Conversations

Learning about Hydrology
https://chatgpt.com/share/0cfece30-26f4-4d91-8da7-d34c9adacf99

Hydrological Evaluation Metrics
https://chatgpt.com/share/e3481aa9-9812-49a1-a428-8f7687bf54c1
https://chatgpt.com/share/87abcb5e-3b48-4340-a9f2-c781a168d046
https://chatgpt.com/share/093d8db5-dd24-4df3-bf00-6818f7571158
https://chatgpt.com/share/449ad371-e80f-4941-9544-0c8e5891c9d7

Learning about ML
https://chatgpt.com/share/dd8ca3f7-5238-4e71-8cfd-23e8bf31b36a

Exploring ML Solutions for Infilling Hydrological Time Series Data
https://chatgpt.com/share/78c4f4b9-b005-41a4-882d-30fe561dbbaa

Learning about LSTMs, MLPs and SOMs
https://chatgpt.com/share/42c65298-a800-4cb7-a67a-e477208bfa24
https://chatgpt.com/share/12bcdb07-809f-4184-96b0-f1924e2e2446
https://chatgpt.com/share/9ef2c11a-53e5-41a2-98ae-d676c85f0f6e
https://chatgpt.com/share/04e7b95a-fe45-4aba-95e2-da8fe074f108

MCA Criteria Validation
https://chatgpt.com/share/42c61867-4b4d-43e5-8c13-9720c43f149b

MCA Score Validation
https://chat.openai.com/share/abe70525-4da1-4548-81bd-f4ec6fba1b2d

Learning to Save ML Models
https://chatgpt.com/share/b7d2cee6-8adb-4156-b4db-99a6c2043e4f

Feature Selection Tips
https://chatgpt.com/share/690bd573-b0bf-4e7d-9f54-c43b2f6d794a

How to Scale Data in Python
https://chatgpt.com/share/66b7e67c-1e72-492c-8a8b-fe4ba7d5ca19

Implementing LSTM with PyTorch
https://chatgpt.com/share/1b8b3fab-eec0-4211-995c-5abf5aca9583
(See for similarities between ChatGPT and other sources: https://medium.com/@mike.roweprediger/using-pytorch-to-train-an-lstm-forecasting-model-e5a04b6e0e67
https://machinelearningmastery.com/lstm-for-time-series-prediction-in-pytorch/)

Example of Learning about Pandas & Debugging Code
https://chatgpt.com/share/c26113d7-ad4d-4c76-bb76-42f30ded8620

Appendix C - Reflection on AI Use

Several AI tools were used during this research as mentioned briefly in the methodology sections, and a elaboration on the use of these tools feels appropriate. First, ChatGPT was used as tutor, validator and discussion partner. I think that using ChatGPT as tool to explain certain concepts improved my learning rate. For example, when learn gin about hydrology or the neural networks, it was sometimes helpful to let ChatGPT explain concepts in simpler terms. Because papers around these concepts are written by experts, for whom the basics are evident, they tend to explain concept in complex ways. When it got too complex, I simply asked ChatGPT to explain it to me like I am 10 years old, slowly making it more complex. Clearly, ChatGPT should not be trusted on its word and therefore was used next to Google and other information sources. However, for some stuff ChatGPT is a better and quicker resource than Google.

Second, when writing code I used ChatGPT and GitHub Copilot for bug fixes and to explain how certain things can be coded. I always made sure to ask the why behind stuff, as this allowed my to learn why certain steps are taken and to replicate it in the future. For coding, it might seem like these AI tools provide answers that are ready to use and thus closer to cheating. However, I would disagree with this. There are always little things in your code (i.e. variable, values, functions, etc.) that are different than the code provided by these AI tools. Therefore, you cannot simply copy and past code without understanding, as this will lead to trouble down the road. Furthermore, using these AI tools is more efficient than looking for bug fixes on Google or Stack Overflow, which basically provides the same answers (see OpenAI's deal with Stack Overflow (OpenAI, 2024)). However, with ChatGPT and Copilot, you can ask follow-up question to better understand the code.

At last, taking into account the financial benefits of productivity gains, knowledge gains, etc., organisations will be forced to use AI due to its benefits. Therefore, it is also essential for people to understand how to work with these tools effectively. Therefore, instead of trying to limit students' AI use, universities should probably see it as a competence to develop, just like coding, interviewing, literature reviews, etc. are tools needed for competent students. It also dramatically lowers the bar for tasks that require expert knowledge or skills.