INDUSTRIAL ENGINEERING AND MANAGEMENT

BACHELOR THESIS

Forecasting Water Levels in Twente Canal Using Time Series Analysis

G. YURTTAS

August 2024

UNIVERSITY OF TWENTE.

Colophon

FACULTY

Behavioural, Management and Social Sciences

DATE

23/08/2024

AUTHOR

Görkem Yurttaş

SUPERVISORS

Sebastian Piest (first supervisor)

Engin Topan (second supervisor)

EMAIL

g.yurttas@student.utwente.nl

Acknowledgements

Dear Reader,

I would like to dedicate this section to the people who have been part of my journey over the last 3 years. This journey was a challenging one for me on many different ways that I was not prepared for before starting it. However, throughout this journey I was able to meet new people, learn and experience many new things. I was lucky to have these people with me in this journey and receive their support and guidance. I wouldn't have been able to complete it without them, and therefore they all hold a special place in my heart.

Firstly, I would like to thank my supervisor Sebastian Piest for his constant support and guidance throughout this research. This research simply wouldn't exist without him and his guidance. His expertise and knowledge helped me to tackle the challenges faced during this assignment. I am very grateful for all the encouragement he has given me, and all the time he has spent on helping me.

I would like to extend my thanks to the BMS faculty staff as well. It was their commitment to teaching and learning that helped me to uncover my interests in the field and push myself further in the face of challenges.

I also would like to thank my friends who have supported me throughout this journey. It was them who I have shared my lowest moments and helped me to overcome them. It was them who I have shared my fondest moments and created life-long memories. When I came here 3 years ago, I was alone in a foreign country; now I conclude my studies with life-long friendships I have made who make me feel at home. I wouldn't be able to come this far without them.

Finally, my greatest thanks go to my parents. There are no words I can use to express my gratitude to them. Every time I've struggled, they were the ones that filled my heart with joy and helped me overcome it. They were the ones who gave everything they could have to me throughout this journey, with only my happiness in their thoughts. As I said, there are no words I can use to express my gratitude towards them and everything I was able to achieve was thanks to them.

Görkem Yurttaş

Management Summary

Introduction

Water levels in the Twente canal have experienced historical lows in year 2018, which affected the supply chain and the logistic operations of the businesses that use the canal to be hindered. As a result, it has become apparent that predicting these unprecedented changes in water levels is crucial to prevent such negative effects again.

This research aims to provide an answer to how forecast models for the Twente canal could be made using time series analysis. The research is part of a bigger project in which a digital twin of the Twente canal is being made. The forecasts provided by this study serve the purpose of monitoring the water levels in this project and also provide a basis for any further models to be developed for the project.

Theoretical Framework

As the relevant modelling techniques had to be identified as well as a theory needed for the research to be based upon, a theoretical framework was made utilising the existing literature. Several models such as ARMA, ARIMA, SARMA, SARIMA, PARMA, ARIMAX, SARIMAX and MLR were identified as a result of this research. Additionally, several important properties such as seasonality and stationarity were described.

Data Understanding and Transformation

Data used in this report were acquired from Rijkswaterstaat and KNMI, both which are public sources of data. In general, data was of high quality and did not require much cleaning in terms of outliers and measurement types. However, the data gathered from Rijkswaterstaat was measured every 10 minute interval, which had to be transformed into a daily average value. After the data was cleaned and transformed into a daily average value, properties of the dataset were examined. Data exhibited low variance and standard deviation which could be the result of averaging the values for daily measurements. Additionally, data was found to be normally distributed, non-stationary and non-seasonal. Finally, several exogenous variables, for which the data was gathered from KNMI, were investigated for use in modelling. Unfortunately, none of there were deemed suitable for different reasons.

Modelling

After the data was explored and transformed, it was ready to be modelled. Several possible models depending on the properties highlighted before were identified. Mainly ARIMA models were found suitable for modelling time series. Additionally, although expected to not add value, an ARIMAX model using temperature as an exogenous variable was modelled for research purposes. According to the theoretical framework and methods proposed in it, parameters for the models were estimated. After the initial estimations, models were fitted and compared on their information criterion. Based on this comparison, models with the lowest criterions were chosen to be actually modelled for forecasting.

Results and Conclusion

In general, there are mixed results from the modelling phase. In particular, long term forecasts were a failure due to the predicted values converging to the sample mean of the training dataset. Several reasons as to why this behaviour occurs could be unincorporated seasonality and low variance in the dataset. On the other hand, short term predictions were highly accurate and were able to showcase the patterns that the actual values follow. Unfortunately, it is arguable how these short term forecasts could be utilised.

Overall, the research acknowledges the fact that further improvements and development on the models are necessary. The inclusion of an expert on hydrological forecasting and river science would be beneficial for any future research on the topic.

Table of Contents

List of Figures

List of Tables

1. Introduction

This chapter informs the reader with the necessary background information needed to understand the problem which is the unpredicted changes in water level affecting the supply chain of the businesses using the Twente canal negatively. In section 1.1, brief background knowledge regarding the drought which happened in 2018 and its effect on the businesses is discussed. In section 1.2, the core problem is identified. Section 1.3 introduces the research project which this assignment is also a part of. Section 1.4 points out the research sub-questions and knowledge questions that are needed to be answered.

1.1. Background

The year 2018 saw a major drought that affected most of Western Europe, which was thought to be the most extreme one between the years 1980-2020 (Aalbers et al., 2023). One of the countries that was severely affected by the drought was the Netherlands. According to Sluijter et al. (2018), major precipitation deficits and higher levels of evaporation were detected in the Netherlands. Figure 1.1 by Sluijter et al. (2018), shows the maximum observed values of the potential precipitation deficit in the Netherlands and illustrates the severity of the drought.

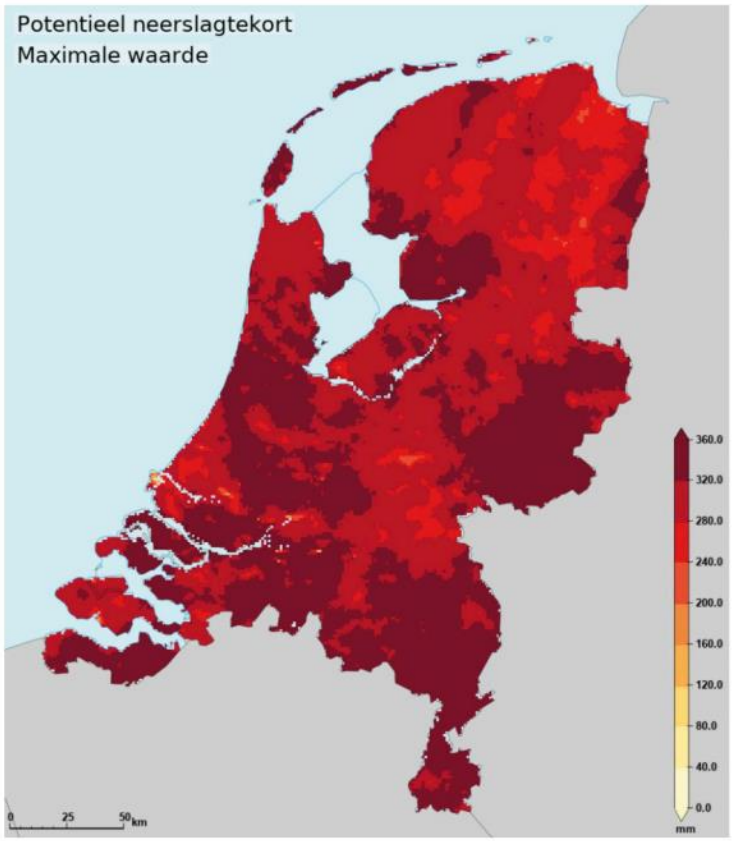

Figure 1.1: Maximum potential precipitation deficit observed (Sluijter et al., 2018)

As a result of the precipitation deficit in the Netherlands and the combined effects of drought in Europe in general, water levels in major waterways dropped severely. To illustrate the effect of the drought on major rivers such as Rhine, an important inland shipping route, discharge of the Rhine River at the town of Lobith, entry point of Rhine into the Netherlands can be seen in Figure 1.2 (Teunis, 2019).

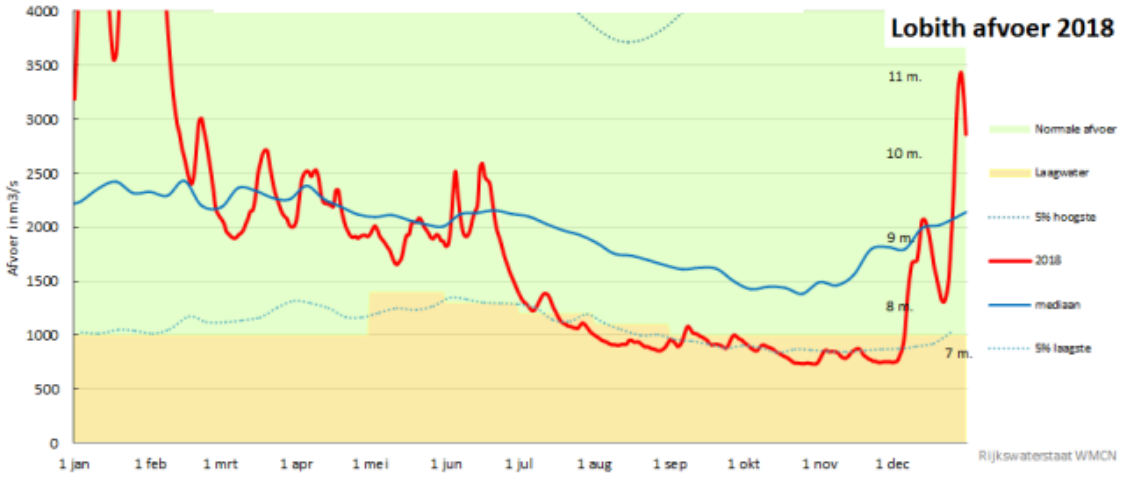

Figure 1.2: Discharge of Rhine River at Lobith, NL (Teunis, 2019)

The Twente canal is a major waterway which connects the cities of Enschede, Almelo and Hengelo to the river IJssel. Many companies in the region use the Twente canal to distribute their goods both domestically and internationally. The Twente canal maintains its water levels by utilizing pumps connected to IJssel. However, this brings up potential issues as the canal depends severely on the situation in Ijssel, especially during times of unexpected water levels. This is mainly due to the existence of a single entry point connecting IJssel to the Twente canal. Due to the existence of a single point, during times of drought when more ships are required due to lower load factors, traffic increases and also makes it hard to pump water from IJssel as only a limited number of pumps can be utilised. The consequences of this were seen during the national drought that affected the Netherlands in 2018. During the drought, the water levels in the canal dropped drastically from the target level of 2.50 meters to a critical level of 1.45 meters, and the target level was not reached for a total of 6 months (van der Kuil et al., 2020). As the water levels dropped to critical levels many businesses had to opt for land transportation and lower load factors on the ships (van der Kuil et al., 2020). Transportation by land was costlier both financially and environmentally, and had a negative impact on the logistic operations as many small-sized enterprises did not have a system set up for it. It is stated by van der Kuil et al. (2020) that, although, it is hard to estimate the exact economic damage the Dutch shipping sector suffered between 65 to 155 million euros.

1.2. Problem Identification

The approach outlined by Heerkens (2017), was used to identify the core problem. The action problem is presented as changing water levels in the canal causing disruptions to the supply

chain. To identify the potential core problems, a why-why analysis is conducted. The method requires to ask why-why questions up to 5 times. The method is applied in Figure 1.3.

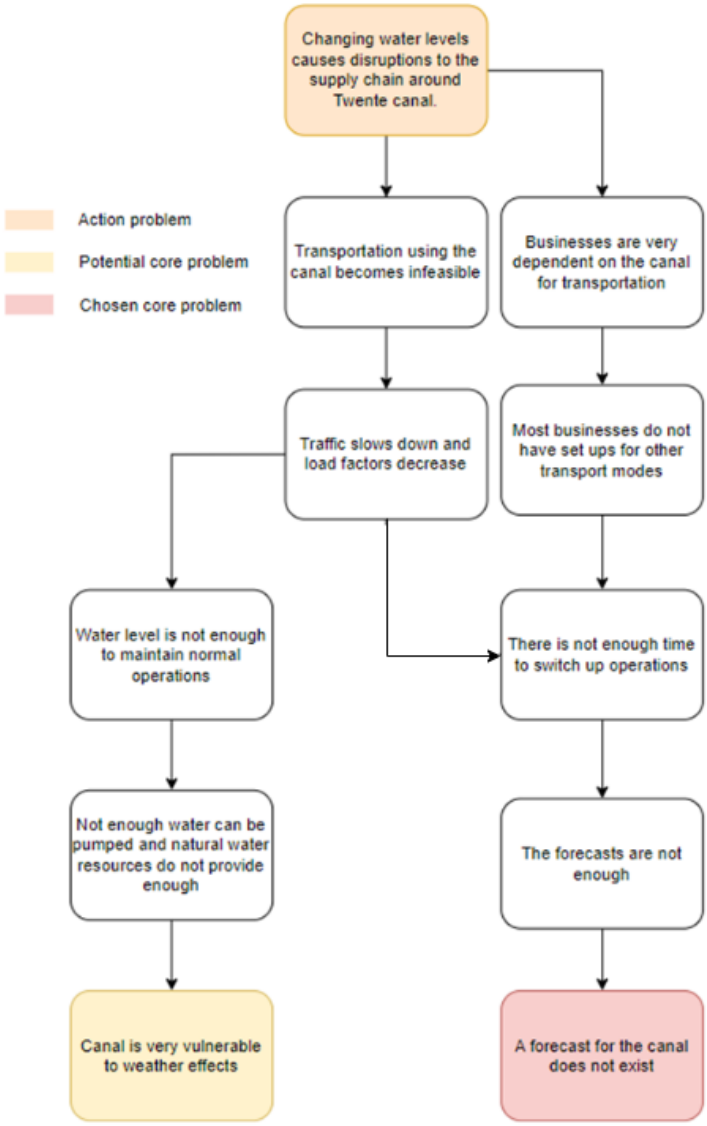

Figure 1.3: Problem cluster using 5 Whys

Following up from the action problem as to the reasons why disruptions to the supply chain occur are identified as transport in the canal being hindered and businesses being too dependent on the canal even during times of crisis. Both of these can be observed during the drought of 2018, in which the transportation in the canal was severely hindered due to lower loading factors with at least one third of the ship's capacity to be utilised (van der Kuil et al., 2020). Additionally, even after the drought started and the canal operations were affected, the businesses were still reliant on the canal and were unable to switch to land transport effectively.

If we follow the line of why transportation in the canal was hindered, it was mainly due to the traffic slowing and load factors decreasing as reported by Van der Kuil et al. (2020). This was due to businesses not having the necessary setups for alternative modes of transportation

and businesses not being able to set them up in time. Another reason is the water level not being enough to maintain normal operations.

If continued on the path of water levels not being enough to maintain normal operations, it comes down to not enough water being pumped and natural water resources being insufficient at maintaining the water level. As the water level declines in the canal, it becomes heavily reliant on water being pumped from the river IJssel. Additionally, the canal is steeper towards the Enschede section as Zupthen, where it connects to IJssel, being at an elevation of 7m while Enschede is around 30 m according to AHN (n.d.). This elevation creates a heavier demand for pumping as it becomes harder to maintain the water level higher up in the canal. Therefore, it is possible to conclude this path with the canal being very vulnerable to weather effects. However, this is not chosen as the topic of the assignment, therefore further elaboration on how this can be tackled will not be presented.

Continuing down on why there is not enough time to set up such systems at least as a temporary solution until the water levels normalise is accounted to forecasts not being enough to determine the operability of the canal. There are currently available forecasts on the water flow of Rhine and Meuse for up to 15 days, which enables companies using these waterways to prepare for severe weather phenomena (Rijkswaterstaat, n.d.-a). However, for the Twente Canal current forecasts are only up to 2 days, with limited statistics and visualization of data compared to the analysis made on Rhine and Meuse (Rijkswaterstaat, n.d.-b). The lack of these makes it hard for companies to determine if the Twente Canal will be affected or not, and how much it will be affected. Therefore, the core problem is identified as the lack of a forecast on the Twente Canal and the assignment aims to provide one.

1.3. Research project

To anticipate and mitigate the risks and costs of extreme natural phenomena such as droughts, a research project was launched to create a digital twin of the supply chain, an advanced computer model of the Twente canal, consisting of companies in the region utilizing the canal as their main transport channel. The digital twin aims to monitor the water levels and provide a toolbox with strategies, techniques and interventions to tackle the problems that occurred during the drought and ultimately build up resilience in the supply chain.

The assignment that is presented in this thesis is related to the above-mentioned research project and aims to develop a forecast model of the water levels in the canal for the digital twin project to determine any unexpected changes beforehand. A forecast model has the ability to act as an early detection system, enabling the companies to prepare for the predicted natural events. By developing the forecast model for the digital twin project, assignment addresses the core problem identified in section 1.2 as well.

The assignment has a set scope and limitations such as the limited time frame of 10 weeks and data that can be acquired. There is no company specific data anticipated to be gathered at the current stage of research, most of data anticipated will be gathered from publicly available data sources. The research will focus on determining the influencing factors, exogenous variables and parameters for the chosen forecast models. The goal of this research is to evaluate different solution alternatives rather than focusing on only one model in-depth. This will be further explained in the problem solving approach in Chapter 2.

1.4. Research question and knowledge questions

This section provides the main research question that is to be answered in this research, which is derived from the core problem identified in section 1.2. Additionally, the research sub-

questions and knowledge questions that have to be answered are formulated. Finally, these questions are fit into the research design.

1.4.1. Research question

The core problem was identified as “A forecast for the canal does not exist”. Based on the core problem, the following main research question was developed to determine the objective and the approach of the assignment.

“How can a forecast model using time-series analysis be developed, to predict the operability of the Twente Canal?”

The assignment requires a forecast model to be developed eventually, which has to be the main focus of the research question. Additionally, a norm and reality must be determined, and the research question must be SMART framework suitable. The following research question is developed according to the requirements described in this paragraph:

The norm-reality can be identified as a forecast model being required and a current model not being available as norm and reality respectively. To elaborate, the current forecasts provided by Rijkswaterstaat are only for the Rhine and Meuse rivers. However, these are not sufficient to determine the operability of the businesses who rely on the Twente canal for their operations mainly, as explained in section 1.1.

The SMART framework requirements proposed by Doran (1981) are met by each requirement as follows:

Specific: The research question specifies the methods used to time series analysis.

Measurable: The research question can be measured using statistical analyses and tests of models created.

Attainable: Historical public data such as precipitation, water level and water flow are readily available online through government agencies such as Rijkswaterstaat or KNMI.

Realistic: The research question aims to tackle the core problem directly by addressing the lack of a forecast model.

Time-bound: The research has to be conducted in 10 weeks as per the requirements of the university.

1.4.2. Research sub-questions and knowledge questions

The following knowledge questions are developed and are required to be answered to conduct the research and answer the research question. The knowledge questions which aim to gather general knowledge about a scientific topic and suitable for SLR are marked as KQ while research questions which are specific to this research are marked as RQ.

I. Which measurement is most relevant for determining the operability of the canal? (RQ)

Many different measurements that can be utilised to determine the operability of the canal. The two most common variables used by Rijkswaterstaat(water management authority in the Netherlands) are water flow and water level. The decision to use which one in the model between these two or another measurement has to be made eventually. Therefore, to make this decision, extensive research into understanding these measurements has to be made.

II. What factors affect the chosen measurement criteria and which ones are most relevant? (RQ)

The chosen measurement criteria, whether it be water flow, water level or another measurement, might be affected by exogenous variables. Precipitation, temperature, ground water and other factors that could affect the chosen measurement must be identified and compared to determine the one/ones to include in the forecast model. If there is enough justification to whether these variables effect the measurement criteria strongly, they should be incorporated into the models proposed.

III. What data must be gathered and where it should be gathered from? (RQ)

After determining the measurement criteria and the variable/variables to be included in the model, the relevant data required must be determined. After determining the required data, the sources which it can be retrieved from must be identified.

IV. What time-series techniques are suitable under which conditions? (KQ)

Extensive research into literature on existing time-series techniques and their applications must be considered to determine the suitable ones for this assignment. Furthermore, they must be compared based on their conditions and limitations to decide on which ones would be suitable for the purposes of this research.

V. How are current forecasts by Rijkswaterstaat made, using which variables and techniques? (RQ)

An opportunity to get information about the current forecasts on Rhine and Meuse would be beneficial to get a better understanding of how modelling on waterways is done in the Netherlands. Additionally, it will also help with understanding the complex relationships between the various variables.

VI. Which indicators can be used to evaluate the model? (RQ)

A selection of indicators is required to evaluate the model in the evaluation phase. Most relevant indicators for the model must be determined using both literature and business objectives. Literature is needed to determine the mathematical quality of the model by looking up the testing norms in academic literature, and business objectives must also be considered to determine the usefulness of the model to the stakeholders.

VII. How can the model be implemented into current operations? (RQ)

A crucial question that has to be answered to conduct the deployment phase, and to give accurate recommendations on implementation of the model into the operations.

1.4.3. Research design

In Table 1.1, the questions proposed in section 1.4.2 are fit into the research design and steps of the CRISP-DM methodology. Type of each question is identified and according to that, suitable methods for answering the questions are proposed.

Knowledge Question	Type of Research	Phase	How to conduct?
I	Exploratory	Business understanding	Through literature research – stakeholder analyses (e.g., skippers) – expert opinions
II	Descriptive and statistical	Data understanding	Through correlation analyses and statistical tests
III	Exploratory	Data understanding	Online databases of government agencies

IV	Exploratory/Experimental	Modelling	Through literature research – characteristics of data – model evaluations
V	Descriptive	Modelling	Through contact with experts
VI	Exploratory	Evaluation	Stakeholder analyses – Business objectives – Literature
VII	Exploratory	Deployment	Literature – Stakeholder analyses – Business objectives

Table 1.1: Research design

1.5. Summary and conclusion

In this chapter relevant background information was given, and additional information on the core problem and important research elements was provided. As stated in section 1.1, the main issues faced by the companies were the lower load factors and costs incurred due to low water levels in the drought periods. Utilising the background information, the core problem was identified through a why-why analysis. Once the core problem was identified, relevant research and knowledge questions were made. Finally, these questions were fit into the research design.

2. Problem-solving approach

The method chosen for the problem-solving approach is the CRISP-DM methodology. CRISP-DM methodology is an abbreviation for Cross Industry Process Data Mining. It is the most commonly used method for projects with heavy focus on data analysis which is also the case for this assignment. The steps of the CRISP-DM methodology are given in Figure 2.1.

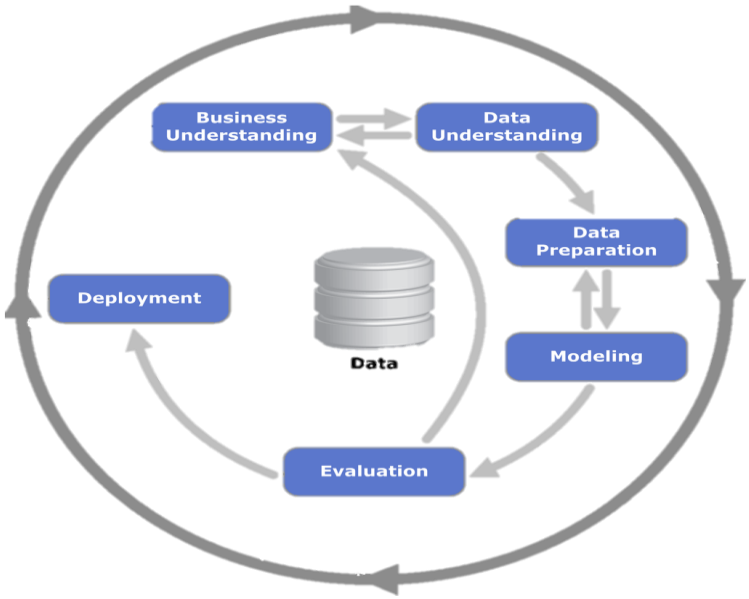

Figure 2.1: CRISP-DM method (Jensen, 2013)

During the assignment, data regarding water levels will be analysed and the factors, variables, and parameters will be determined for the forecast model will be briefly described. Each step of the CRISP-DM methodology will be briefly described below. The methodology will be followed step by step until the deployment phase due to time limitations of the project and the status of the parent research project. A deployment plan will be provided instead.

2.1. Business understanding

According to Schröer et al. (2021), the business understanding phase can be presented in various ways depending on the area of research. However, common approaches are describing the business goals and the use of data mining in the research. A paper by Krishnaswamy et al. (2022) on their research about the application of the CRISP-DM on human-wildlife conflicts, presents a good example regarding the application of the method as a whole and what can be done in each step. In their research, they present relevant stakeholders, business objectives and data mining goals. The assignment will follow a similar course by identifying the main business objectives and expectations through the use of interviews. Additionally, most of business understanding was already covered in chapter 1. The results of analyses of interview summaries are presented in chapter 4.

2.2. Data understanding

In the data understanding phase, it is common to explain how and where data were collected and generate descriptive statistics to learn more about the data at hand (Schröer et al., 2021). Therefore, the main objective in this phase will be to identify the data that are needed for the scope of the assignment. Additionally, variable(s) that will be used in the model will have to be identified in this phase to determine the data required. After the data requirements are determined, the method and the sources of the data will be explained, and the type of data

such as primary or secondary will be identified. Similar to the example case by Krishnaswamy et al. (2022), a description of the attributes will be created, and the quality of the data will be verified by identifying issues related to it. The results of this phase are presented in chapter 5.

2.3. Data preparation

Data preparation is an important phase of CRISP-DM and generally is one of the biggest tasks. The most common tasks in this phase are described as the selection, transformation and cleaning of the data at hand (Schröer et al., 2021). Similar to Krishnaswamy et al. (2022), data will be cleaned to assess the quality issues identified in the data understanding phase and new attributes will be created if deemed necessary. The results are presented in chapter 6.

2.4. Modelling

In this phase, the forecast model is to be made with a number of different modelling techniques to make comparisons. The reason behind comparing different techniques is described by Schröer et al. (2021) as experts finding the comparison results useful during the evaluation of the model in the next phase. Therefore, during this phase, different time-series methods of forecasting will be identified, compared and chosen based on their suitability with the assignment. Finally, a model of each chosen technique will be presented for comparison of results. The results are presented in chapter 7.

2.5. Evaluation

During this phase, the quality of the models developed will be assessed and compared. Several metrics and statistics will be compared across models and explain performance differences between them. Necessary statistical tests will also be conducted during this phase to check the validity of the models. The results are presented in chapter 8.

2.6. Deployment

As the deployment of the model in real life is not possible at the current stage of the research and the scope of this assignment, recommendations on the use and implementation of the model will be given. The recommendations will be on how the model can be used in real life, its limitations and what can be expected from it. The results are presented in chapter 9.

2.7. Summary and conclusions

In this chapter, the CRISP-DM methodology, which will be used in this research was presented. The methodology outlines how the research should be conducted and provides a clear guideline for steps to take to finalise the research. Each step of the methodology was explained with how they will be utilised in this research.

3. Theoretical Framework

In this section, theory which this research is based on is explained and discussed in detail. Firstly, a systematic literature review(SLR) is conducted to identify major concepts, forecast models used in literature and validation methods. After the initial identification of these models and relevant papers, all the relevant concepts, models and validation methods are discussed in detail in their respective sections throughout the chapter. Ultimately, identifies the main concepts that determine the model selection and the limitations of the identified models to determine what models can be used in chapter 7.

3.1. Systematic Literature Review

Through a systematic literature review of various articles and books, many available models have been discovered used in time series data forecasting. Firstly, relevant concepts and databases were identified. After the identification of the databases, the key concepts were searched in these databases and depending on the number of results, all or some of the most relevant papers were inspected. More information regarding the literature search, selection criteria and the concept matrix of the identified papers can be found in Appendix A.

The methods found in the literature include models such as autoregressive moving average(ARMA), autoregressive integrated moving average(ARIMA), Seasonal ARIMA(SARIMA), ARIMA with Exogenous Variable(ARIMAX), SARIMA with Exogenous Variable(SARIMAX), periodic ARMA(PARMA) and Multiple Linear Regression(MLR) are among the methods that were found in the systematic literature review. Other major theoretical aspects are the concepts of stationarity and non-stationarity, as well as seasonality. Finally, various error indicators for model evaluation have been identified which will be discussed in section 3.3.

3.2. Characteristics

This section discusses certain characteristics of the dataset which is important in the context of time series. As time series models assume the data to be stationary, it is important to determine if the used dataset is stationary or non-stationary. Otherwise, it is not possible to model non-stationary datasets using time series models. Therefore, the concept of stationarity is explained. Additionally, some data exhibit seasonal patterns over the course of time. Therefore, the decision to incorporate seasonality in the models or not has to be made for certain dataset. In this section, the seasonality is also discussed to be able to make this decision later on in the modelling phase.

3.2.1. Seasonality

According to Pal & Prakash (2017), seasonality is defined as repetitive and periodic divergences. Examples that are common in real life may be snowboard sales going higher in winter or tourist numbers getting higher in holiday periods. In the case of this assignment, seasonality can be a factor especially if the rainfall, temperature or other weather factors changing between periods of time. To support this claim, Anderson et al. (2012) also suggest that river flows generally contain seasonal shifts in their descriptive statistical values such as mean and standard deviation.

There are various ways to identify seasonality in time series. A common method is to identify seasonality is through conducting an exploratory analysis of the data set through various plots. Pal & Prakash (2017) suggest the use of run sequence plots, seasonal sub series plots and multiple box plots specifically. Through plotting, deviations in the mean and variance can be identified visually. In their research, Narasimha et al. (2017), also utilise a run sequence plot

to identify seasonality, stationarity and any outliers. Differently, another way is to compare the amplitude of fluctuations between different periods (Banas & Utnik-Banas, 2021).

3.2.2. Stationarity

Stationarity is a concept in stochastic models of time series data that heavily influences the models that can be used and parameters that have to be chosen. According to Box et al. (2015), a model can be identified as stationary if it stays in a statistical equilibrium with its probabilistic properties not changing over time, specifically shifting around a constant mean and variance. To elaborate more on it, it implies that statistical properties such as mean, and variance do not change over the course of time. For example, if the GDP of a country has been growing constantly for the past few decades, the time series cannot be considered stationary as the mean GDP of the country has increased over time as well. However, if we were to do a coin toss, and assign 1 and 0 to heads and tails respectively, the process would be stationary as the mean would not change over time because there is equal chance of landing heads and tails.

In the modelling process, stationarity determines both the methods that can be used and what the parameters should be for each method. There are various ways to determine stationarity in data. It is possible to get an understanding of it via utilising visual analysis of a run sequence plot, as Tyagi et al. (2023) did it in their paper before pursuing statistical tests. In their research, it is possible to see from plot that the sugarcane production has been growing continuously implying that there may be non-stationarity (Tyagi et al., 2023).

However, most literature that has been reviewed presents additional tests or analysis in order to determine stationarity. Approaches used by Tyagi et al. (2023), Viccione et al. (2019) and Banas & Utnik-Banas (2021) were Augmented Dickey-Fuller (ADF) unit root test, Phillips-Perron (PP) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Among the methods, ADF seems to be the most prominent one as it used in all 3 papers. Additionally, KPSS test is different compared to the other two in the sense that its null hypothesis is series being stationary whereas the first two have their null hypothesis as the series being non-stationary. Finally, Narasimha Murty et al. (2017) utilises the Autocorrelation (ACF), Partial Autocorrelation (PACF) and Inverse Autocorrelation (IACF) graphs to determine stationarity.

3.3. Modelling

This section discusses the different modelling techniques that are used in the literature. It provides an understanding of these models and explains their properties. It also states the conditions these methods are used as to when they are used in the literature. Additionally, it explains how parameters for the models are chosen and the series is fit into these models.

3.3.1. ARMA, ARIMA and ARIMAX

One of the most common methods in time series forecasting is Autoregressive Moving Average (ARMA). The other 2 methods Autoregressive Integrated Moving Average (ARIMA) and ARIMA with Exogenous Variables (ARIMAX) are expansions of ARMA. Essentially, ARMA can be described as a mixed model of Autoregressive (AR) and Moving Average (MA) models (Box et al., 2015). It is advantageous to use both methods as it allows for greater flexibility to fit the time series (Box et al., 2015). It has two variables which are the order of autoregressive denoted as p, and the order of moving average denoted as q. The paper by Dubey et al. (2021) explains these further as p denoting the model having its own lags and predictors, and q denoting the number of forecast errors lagged.

Although, the ARMA model has its advantage of combining both AR and MA processes, it falls back when compared to its expanded version ARIMA. The main reason behind this is the fact that we can still keep the full advantage of combining the methods, as well as applying the model to non-stationary data. ARIMA is suitable for non-stationary data as well due to its additional parameter *d* which denotes order of differencing. According to Box et al. (2015), it can be described as an integration or summation of the ARMA process on the order of *d*. As described the model is powerful as it allows for the use of non-stationary data as by differencing, the non-stationary data can essentially be made stationary by differencing it *d* times (Tyagi et al., 2023). Additionally, an ARIMA model that is denoted as ARIMA *(p, 0, q*) in which the order of differencing is equal to 0, is in its essence an ARMA (*p, q)* as no differencing implies stationarity (Dubey et al., 2021).

Additionally, there is ARIMAX models which incorporates an exogenous variable into the model. An exogenous variable is defined as a variable that causes the output variable, but the output variable does not cause the exogenous variable (Box et al., 2015). For example, the sunlight causes the grass to grow, but the grass does not influence the sunlight in a meaningful way, thus making sunlight an exogenous variable to the growth of grass. This is relevant for the assignment as well, because there may be outside factors that influence the canal such as water being pumped in, rainfall or ground water levels.

Finally, to determine the parameters *p*, *d*, and *q*, the most common approach employed by Banas & Utnik-Banas (2021), Dubey et al. (2021), De Figueiredo & Cavalcante Blanco (2016) and Narasimha Murty et al. (2017), is the analysis of ACF and PACF. They are determined by identifying the lag number where the values become zero or not significant.

3.3.2. SARIMA and SARIMAX

SARIMA and SARIMAX are closely related to ARIMA and ARIMAX models, and the major difference is the inclusion of seasonality parameters. It is denoted as SARIMA(*p, d, q)(P, D, Q)* in which *P*, *D*, *Q*, represent the order of seasonal autoregression, the order of seasonal differencing and the order of seasonal moving average, additionally s denotes the length of the seasonal period (Banas & Utnik-Banas, 2021).

The method is useful in the presence of seasonality, such as in the implementations of Narasimha Murty et al. (2017) and Cheng et al. (2021). In their research, Narasimha Murty et al. (2021) applies a seasonal period of s=4, due to the monsoon period which lasts 4 months. In another research by Cheng et al. (2021), a daily forecast of emergency room occupancy is modelled, and they use a seasonal period of s=7 to model daily seasonality between weeks. This may also be relevant as well for the assignment due to the possibility that there may be seasonality in the water levels and flow in the canal. Usually, weather factors that affect the water level and flow such as precipitation, temperature or ground water are seasonal. To support this claim, De Figueiredo & Cavalcante Blanco (2016) also claim that SARIMA along with ARIMA is widely used in hydrological time series.

Finally, SARIMAX, similarly to ARIMAX, is a SARIMA model with the inclusion of exogenous variable. It is also important to note that, ARIMA and SARIMA models were the two most common methods that were encountered in the systematic literature review, most likely due to their flexibility and well-established methods.

3.3.3. PARMA

PARMA was encountered in one paper during the systematic literature review and was included due to Anderson et al. (2012) claiming that they provide more accurate models concerning river flows. Anderson et al. (2012) argues that, due to river flows being periodically

stationary and exhibiting seasonality, it is optimal to use PARMA model as an option. The model is developed by minimising the mean squared prediction error (MPSE) by utilising orthogonal prediction.

The model may be useful due to possible seasonality and stationarity in the data set of water levels and flow, however the smaller number of literature available compared to other methods make it hard to find cases where it is applied.

3.3.4. MLR

Multiple Linear Regression (MLR) is a technique that employs regression, a statistical approach that is used for predicting the relationship between variables that have causality, with the use of one dependent and one or more independent variables (Gupta & Agarwal, 2021). Additionally, Birinci & Akay (2010), further define it as being modelled by a least squares function. In their research, Xie et al. (2020) showcases the application of it by demonstrating the formula they use. Xie et al. (2020) determines the exogenous variable they use by checking for the highest correlation coefficient. This method can also be useful for determining the exogenous variables that can be employed in ARIMAX and SARIMAX methods. The formula used by Xie et al. (2020) is presented below in Equation 4.1:

$$R_{t+j} = a * R_j + b * P_j + C$$

Equation 3.1

To explain the Equation 3.1, R_{t+j} is the predicted variable in future t, R_j is the measured value of the predicted variable in the past period j, and P_j is the measured value of the exogenous variable in past period j. The coefficients are a and b, while the C is the constant in the equation. A common method to develop the model used by Birinci & Akay (2010) and Gupta & Agarwal (2021) is to use 80% of the data to train and 20% to test the model by utilising various computer models such as Excel in the case of Xie et al. (2020).

The MLR model can be useful in this assignment due its relative simplicity allowing for more variables to be tested, and for the model to be prepared faster. However, Birinci & Akay (2010) state in their research that ARIMA outperforms MLR when assessment values are compared.

3.4. Validation with error metrics

Most common methods that were observed in the literature during the systematic literature review were: mean absolute error (MAE), mean absolute scaled error (MASE), mean absolute percentage error (MAPE) and mean squared error(MSE). According to Cheng et al. (2021), MAE gives the average difference between the estimated and actual values, MSE provides the squared difference and due to its squared nature it punishes higher differences more, and MAPE providing a measure that can be compared between different models due to it being presented as a percentage and being unitless. The MASE method also provides a measure that can be compared among different forecast methods and is helpful for determining relative accuracy (Dubey et al., 2021).

Due to the assignment containing different methods to assess and compare in between, use of either MAPE or MASE seems to be crucial as it allows for different models and datasets to be compared. Additional measurements can be considered to provide more insights, and the commonly used methods such as MAE and MSE can be considered as well other methods that are less common in the literature.

3.5. Decision tree for model selection

The following decision tree is prepared after the integration of the knowledge presented below in Figure 3.1.

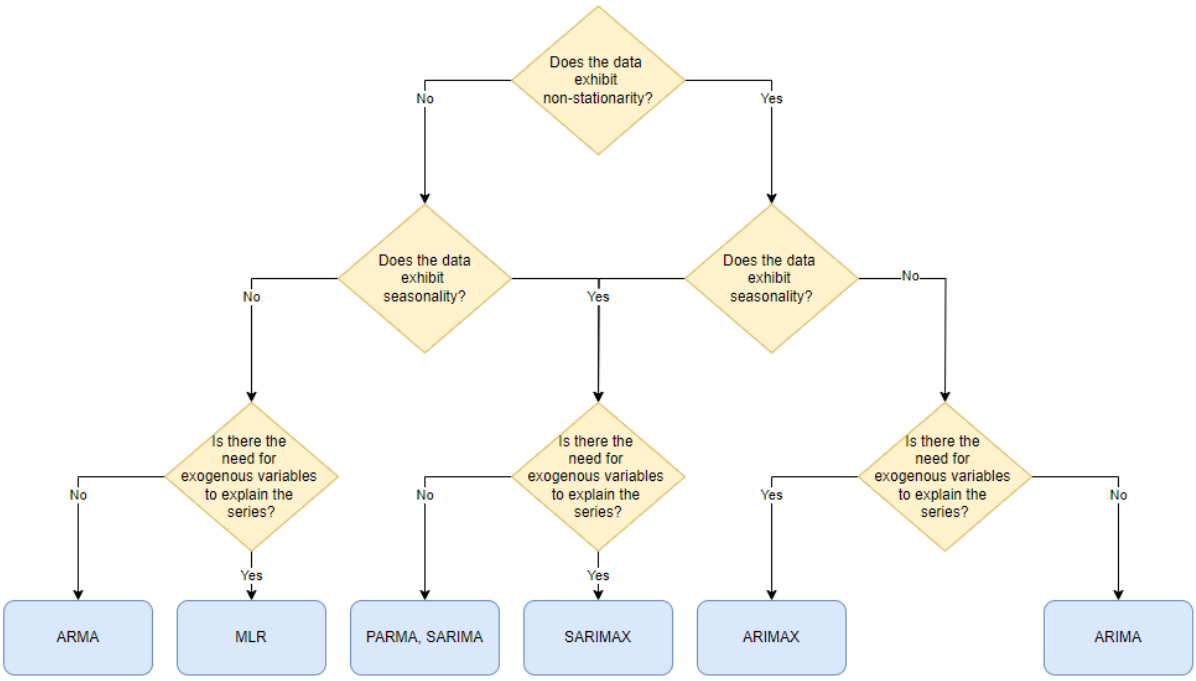

Figure 3.1: Decision tree for choosing a model

To briefly describe the logic behind the decision tree, the data are first assessed on the basis of stationarity as this allows for quick elimination of some models as for example if the data are non-stationary, the ARMA model cannot be used. Then it is assessed based on the seasonality. Here two paths converge in the middle path, and it may raise the question to if it does not matter, if their origin is stationary or not. The key information that should be taken into account is the ARIMA, SARIMA and their exogenous variations can be made into a stationary model very simply by just setting the differencing parameter to 0. So therefore, for these methods it does not explicitly matter if the data are stationary or not. Additionally, for PARMA periodic stationarity and seasonality matter therefore it can be used in many cases as well. Adding on to why ARIMA and ARIMAX models are not included as alternatives to ARMA and MLR if they can be made stationary, is the fact that both the ARMA and MLR can do what ARIMA and ARIMAX does in a much simpler way if they are stationary nonseasonal. Finally, the variables are checked on the basis of exogenous variables being required or not in the mode. The decision to include exogenous variables should be done on a separate basis outside of the scope of this decision tree. If we assume that decision has been made with the help of interviews and data understanding, this last decision points whether we need them or not puts the final mark on which model is suitable. However, it must be kept in mind the aim of the project is to produce multiple models to be compared to decide on the most suitable one and this decision tree serves as a hypothesis to the optimal model based on the theoretical framework. Therefore, even if we end up using SARIMAX for example, we must test other models that are closest to it such as SARIMA as only one decision point differentiates them.

3.6. Summary and conclusion

In this chapter, theory which will be the backbone of this research was provided. The theory was supported with applications from cases in relevant literature. An SLR was conducted to acquire the necessary knowledge for the theory. Each model found in the literature and was

relevant according to the terms of SLR were described briefly. Some models such as SARIMA and SARIMAX were found to be more suitable with data exhibiting seasonal properties. Models such as ARIMAX and SARIMAX were found to be useful when an exogenous variable was incorporated into the models. Additionally, several error metrics for validation were described as seen in the literature. Finally, a decision tree which provides the justification for choosing a model later was made in accordance with the theory.

4. Business Understanding

This chapter briefly discusses the results and findings from several interviews from Anonymous(a) (personal communication, 2024), Anonymous(b) (personal communication, 2024) and Anonymous(c) (personal communication, 2024). The interviews were conducted as part of the greater digital twin project. Additionally, all interviews were anonymised and any data that could be either directly or indirectly linked to the companies were redacted. It is important to note that however, redacted data were irrelevant for this research and therefore o not affect the contribution of the interviews to the research process. The summaries of each interview are presented in Appendix B, C and D respectively.

4.1. Interview findings

In all of the interviews, water levels were an important point discussed due to their effect on operations. Anonymous(a) (personal communication, 2024) and Anonymous(b) (personal communication, 2024) have both stated that water levels cause lower load factors, where the former also underlines this increases the costs per ton. Additionally, Anonymous(b) (personal communication, 2024), also states that during low water government sometimes takes measures such allowing only one-way traffic and during high water levels, the passages under bridges are affected. The locks are also affected by the low water levels due to decreases in frequency of lock operations affect the waiting times for the ships (Anonymous(b), personal communication, 2024). Additionally, Anonymous(a) (personal communication, 2024) also states that the change of transportation mode from shipping through the canal to transport and rail increases both costs and operations difficulty. All interviews highlight the need of forecasting for water levels to lower the negative impacts caused by changing water levels.

4.2. Summary and conclusion

The interview results compliment the findings in chapter 1 regarding the problems the companies face during low water levels. The core problem identification is also justified through the interviews as all state the need of a prediction for water levels to be made. One important points is the emphasis on water levels instead of other measurement criteria such as water flow. This is most likely due to water levels being easily understandable and the main determining factor for the businesses. This concludes that the water levels should be used in the model as well instead of water flow for predictions.

5. Data Understanding

In this chapter, the raw data gathered will be analysed in order to get an understanding of it. This will help with identifying which data will be used or dismissed. It is also crucial in order to understand the quality of the data and what has to be cleaned. Additionally, the properties of the data such as stationarity and seasonality are examined in this chapter which will determine the models to be used later on. As a side note, an initial cleaning has been done in this chapter as well. Normally any cleaning is the part of data transformation phase of CRISP-DM methodology, but for purposes explained later in this chapter and in chapter 6, it was crucial to conduct an initial cleaning at this point of time in the research.

5.1. Gathering initial data

To understand the data, it is important to know about how the initial data were acquired. The water level is chosen as the variable for which the forecast will be conducted on. Historical data for water levels is available online to the public through Rijkswaterstaat (Rijkswaterstaat, n.d.-b). The data are measured as the surface level of the water in reference to the Amsterdam Ordnance Datum, also known as Normaal Amsterdaamse Peil(NAP). In the Netherlands, NAP is the standard measurement for water levels. According to Rijkswaterstaat, it is approximately equal to the average sea level at 0 (Rijkswaterstaat, n.d.-a). Therefore, one must be careful while using the measurement as the elevation of the canal floor is added up to the measurement. To elaborate on this, if an assumption of 1000cm measured on a day, one can incorrectly point out that the canal is not 1000cm deep and assume that the data are wrong. Indeed, the canal is not 1000cm deep and 500cm at the deepest point, however, as the surface height of the water is being measured this actually means that a point on the surface is 1000cm above the NAP not above the canal floor. Therefore, it actually indicates that we are on an elevated position 1000cm compared to the sea level, and the canal floor is also at a higher elevation compared to the NAP. In conclusion, the measurement of 1000cm includes the elevation of the canal floor in addition to the water level at that moment. A visual represantation is presented in Figure 5.1, where A represents the surface water height compared to NAP which is the measurement form in data, B represents the height of the canal floor compared to NAP and C representing the actual water level from the canal floor.

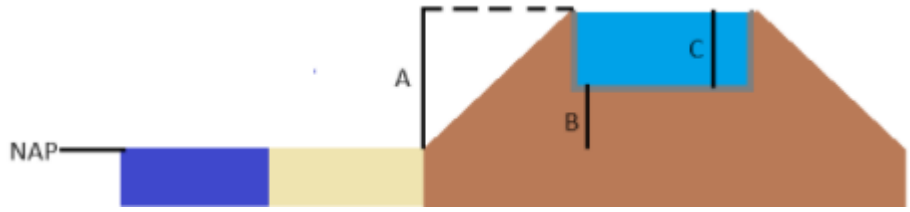

Figure 5.1: Visual representation of the difference between measurement and the actual water height in canal (own work)

As mentioned above, Rijkswaterstaat has a platform where the desired data can be downloaded after choosing a date interval and a measurement point. An interval of 55 years from 01/01/1969 to 01/01/2024 was chosen. The period was chosen in accordance with similar ranges from literature such as De Figueiredo and Cavalcante Blanco (2016), and Narasimha Murthy et al. (2017) which include ranges of 33 and 63 years respectively. It is important to note that 01/01/1969 was the oldest available date available and a longer period was not possible to choose unless the end date was moved, which was not done in order to have the same start and end date.

The data collection point was chosen as “Eefde boven”, which is the measurement point just above the Eefde lock that connects the canal to the river IJssel. There are three other measurement points between the Eefde lock and the Delden lock but the differences between the measurements at these points are negligible. Additionally, this section covers the largest proportion of the canal from Eefde to Delden and Almelo. Finally, due to time constraints and the scope of the research additional measurement points could not be considered, and more on this limitation will be discussed in the chapter 10. Therefore, the “Eefde boven” measurement points were considered the most relevant measurement point.

Figure 5.2: Map of the section between the locks Eefde, and Delden and Almelo in orange. Measurement point “Eefde boven” is identified with red and the green points are other measurement points. (Rijkswaterstaat, 2024) (Edited by own)

5.2. Description of data

For the initial analysis of the dataset, a Python script was used made by Patience (2018). The script analyses the data and returns properties such as number of columns, missing columns, null values and data type. According to the analysis, there were 50 columns out of which 25 were empty. The full list of columns is available in the Appendix E.

The script also concluded that there are a total of 690715 rows. Although this number is unexpectedly high, considering the days between the dates chosen, it is largely due to data containing different time intervals for measurements. This is explained more in detail and transformed to make the intervals uniform in section 5.2.1.

The relevant columns were identified as date, time, measurement method and the numerical measurement. The date is essential for time-series analysis and time is necessary in order to transform the data in later stages of CRISP-DM methodology as there are different measurement intervals. Numerical measurements are all in centimetres compared to NAP, which is explained previously. There are three different measurement methods used at different dates, which are visual readings, water height over previous and next 5 minutes, and arithmetic average over previous and next 5 minutes. It is important to keep these in mind as to determine if different measurement make a difference in forecast quality at later stages. Finally, the numerical value is necessary as it is the value that is being forecasted. The attributes of the columns chosen are below.

Column name	Description	Format

WAARDEBEPALINGSMETHODE_OMSCHRIJVING	Measurement method	object
WAARNEMINGDATUM	Date of measurement	object
WAARNEMINGTIJD (MET/CET)	Time of measurement	object
NUMERIEKEWAARDE	Measurement in cm	int64

Table 5.1: Attributes of relevant data columns from initial dataset

It is important to note two things looking at the attributes. Firstly, the format of the data is not of original Python data types but the NumPy module data types. Secondly, the numerical measurements are all integers as there are no decimal measurement. However, this will change after the initial cleaning done in section 5.2.1 as the daily average of measurements will be taken which will yield decimal values.

5.2.1. Initial cleaning of the dataset

As mentioned, the data downloaded had different intervals between measurements at different dates. From 01/01/1969 to 03/11/1996, the intervals were daily. From 03/11/1996 to 26/11/2013, the intervals were hourly. Finally, from 26/11/2013 to 01/01/2024 the intervals were 10 minutes. Due to non-uniform intervals in the dataset, an initial cleaning that has to take place before moving on to the data preparation step of the CRISP-DM methodology was necessary. Otherwise, the properties of the data such as the total number of entries were misleading.

For this initial cleaning, occurrence of each date in the dataset was counted and the sum of the values in these dates were divided by this count. This results in a daily average value for the measurements. Although, it can be argued that the difference between the number of observations in the earlier and later dates might cause differences in data quality, this was ultimately the only way to achieve a uniform dataset. The only other method would have been to create new data for every 10 minutes in the earlier stages, which would have been less reliable and efficient due to creation way too many non-existing data entries. The formulas used during cleaning are available in Figure 5.3.

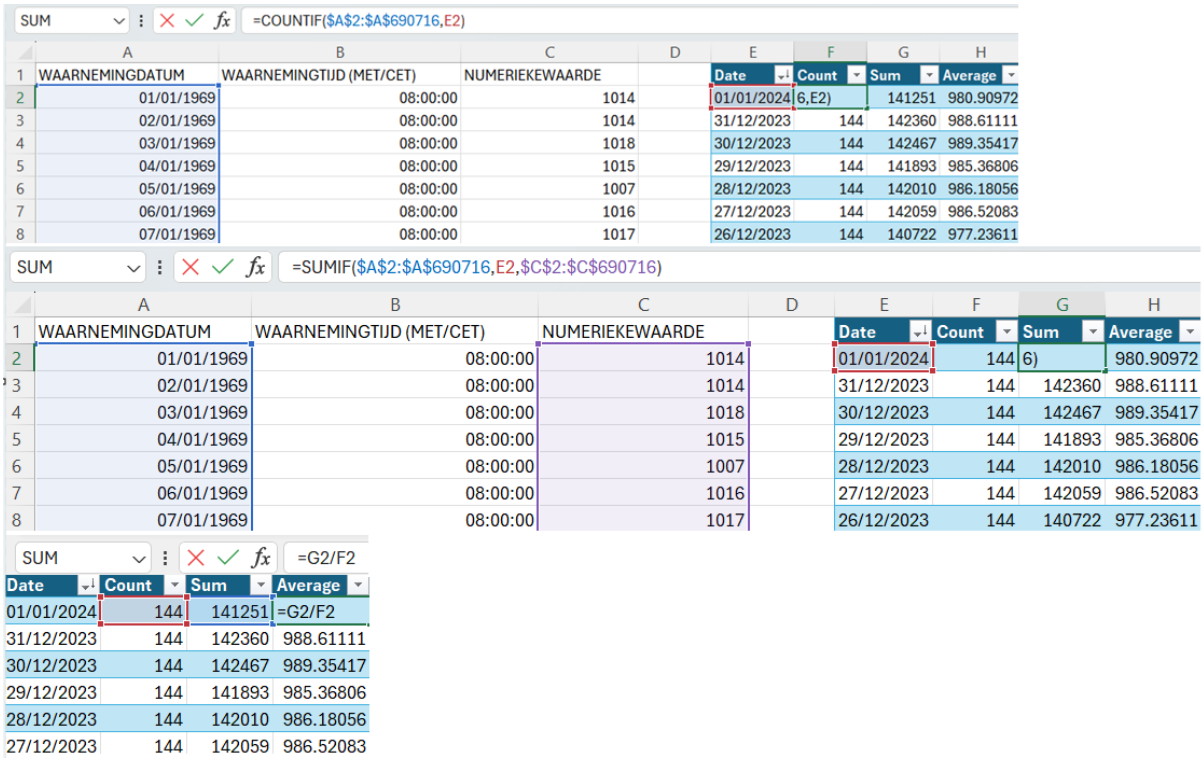

Figure 5.3: Formulas used to convert different intervals of measurement into daily average

During this cleaning, three missing dates were added into the dataset using the value “-1” to identify them easier at later stages where these missing values will be dealt with in a proper way. At this stage, this value is assigned just as a placeholder.

After this initial cleaning the properties were measured once again, and it was apparent that the number of rows were drastically reduced. This is mainly due to having 24 and 144 observations per day at hourly and 10-minute intervals respectively. Once these were converted into daily observations, there was only one value per day. In total the number of rows were measured as 20089, including 3 additional data entries that were added as placeholders for the missing dates. It is important to note that, a full cleaning was not conducted at this point and the data still contains many outliers and mismeasurements which will be dealt with in later stages.

Date	Average
01/01/1969	1014
02/01/1969	1014
03/01/1969	1018
04/01/1969	1015
05/01/1969	1007
06/01/1969	1016
07/01/1969	1017
08/01/1969	1013

Table 5.2: Result of initial cleaning (only the first 8 entries are shown)

The new attributes of the cleaned dataset are presented in Table 5.3.

Column name	Description	Format
Date	Date of measurement	object
Average	Average of daily measurements	float64

Table 5.3: Attributes of the dataset after initial cleaning

It can be noticed that the “time” data are not needed anymore and is omitted from this point on. This is due to taking the average of daily measurements, which means that the time in the sense of hour and minutes of each individual measurement is unnecessary to be known. Additionally, it is noticeable that the data type of measurements has changed from int64 to float64. Due to taking the average value, as a division operation yield results with decimal values. This makes the data type of the measurements change from integer to float, as the latter can contain decimals.

5.2.2. Data quality

In general, the data are of high quality in terms of number of observations available and number of valid observations. There were only 3 missing values without accounting the outliers and mismeasurements. What is identified as mismeasurements are values that are clearly erroneous which are above 100,000. It is impossible for the measurements to be 100,000cm above NAP as the canal floor is 5m above NAP and the depth of the canal is 5m at the deepest point. There were 472 values that were identified as mismeasurements in total which were above 100,000cm. The next highest value after the mismeasurements was 1359cm, which was 300cm higher than the next highest measurement but was left in the dataset as it had to be checked as an outlier despite the fact that it most likely is.

Another issue with the data, as mentioned in previous sections, was the difference between time intervals but this has already been tackled by converting the measurements into daily average. Additionally, the measurements were not observations of the water level in the canal, but the surface water level compared to NAP. To tackle this, the height of the canal floor compared to NAP had to be known as mentioned previously. This was requested from the Rijkswaterstaat, and as a result a map which shows to ground heights in waterways compared to NAP was provided. However, this will not be applied onto data at this stage as it is a part of data transformation step.

5.3. Data exploration

After the initial cleaning was done and the dataset was reduced to a number which is easier to analyse but also a uniform one, additional analysis was carried out. Firstly, the outliers in the data had to be removed before continuing on with further analysis. To do this, the methodology presented by National Institute of Standards and Technology(NIST) (2012) was used. It was decided to remove the values that lie outside of lower outer and upper outer fences. The outer fences were chosen over the inner fences, as choosing the inner fences would result in much of the data being lost. These fences are described as below:

$$Lower\ outer\ fence = Q_1 - 3 * IQR$$

$$Upper\ outer\ fence = Q_3 + 3 * IQR$$

The Q_1 and Q_3 represent the first and third quartiles respectively. IQR denotes the interquartile range which is calculated as:

$$IQR = Q_3 - Q_1$$

To calculate the quartiles and the median value, Excel formulas were used. Below are the results of the calculations:

Median	Q1	Q3	IQR	Lower outer fence	Upper outer fence
1003	1000	1007	7	979	1028

Figure 5.4: Median and the outer fences for the outliers

In total there were 530 values which were identified as outliers in the dataset. This corresponds to approximately 2.64 % of the whole dataset. It is important to note that a lot of the outliers are erroneous values, which account for 472 of the 530 outliers identified. Therefore, only 58 of the valid measurements were outliers in the data.

As the data are cleared of outliers at this point, it is possible to conduct exploratory analysis of the dataset. Below are the descriptive statistics gathered using Excel.

Daily Average	
Mean	1003.433
Standard Error	0.041792
Median	1003
Mode	1000
Standard Deviation	5.844711
Sample Variance	34.16064
Kurtosis	0.908137
Skewness	0.19352
Range	47
Minimum	979
Maximum	1026
Sum	19626156
Count	19559

Table 5.4: Descriptive statistics of the dataset

The mean is around 1003cm, which checks out considering the approximate height of the canal floor of 5m. Therefore, the depth on average is around 5m as well. The variance and standard deviation are relatively low, which are around 6 and 34 cm respectively. This also makes sense considering the relative stability of the water levels in the canal observed in the dataset. The skewness value is close to 0, indicating that the data are relatively symmetrical. Finally, kurtosis is close to 3.91, which means that the data are slightly more peaked than normal distribution as the kurtosis for normal distribution is 3. Note that, the excel calculates relative kurtosis, which subtracts 3 from the actual value, hence why it shows 0.91.

Next, histograms and a QQ-plot are made to better understand the distribution of data.

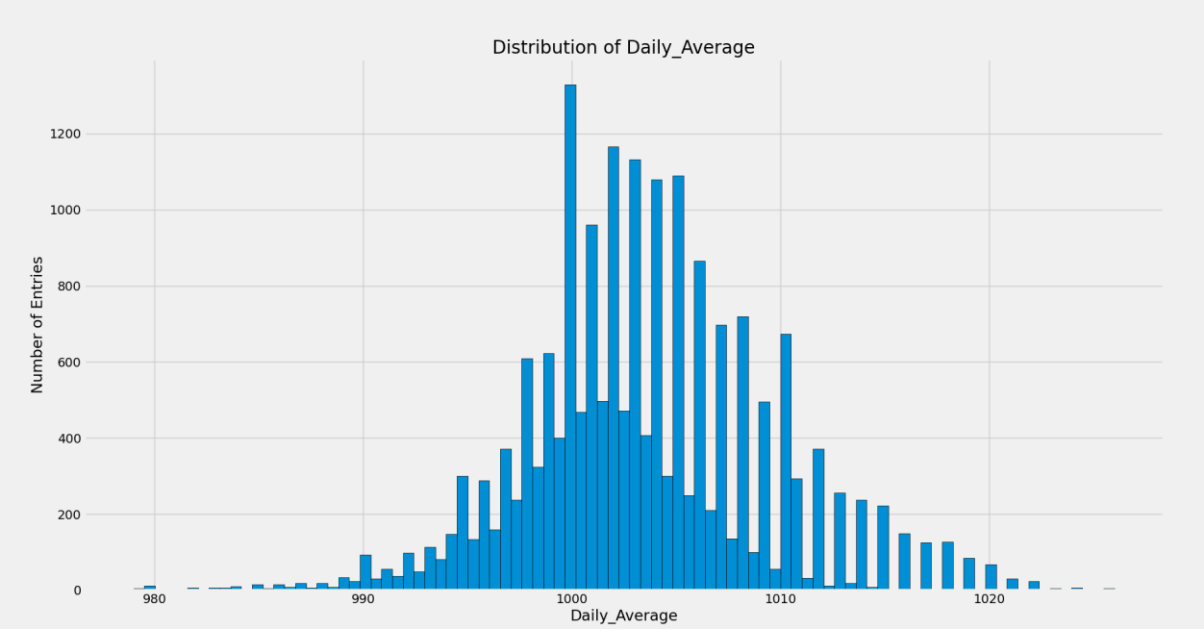

Figure 5.5: Histogram made by using Freedman-Diaconis rule

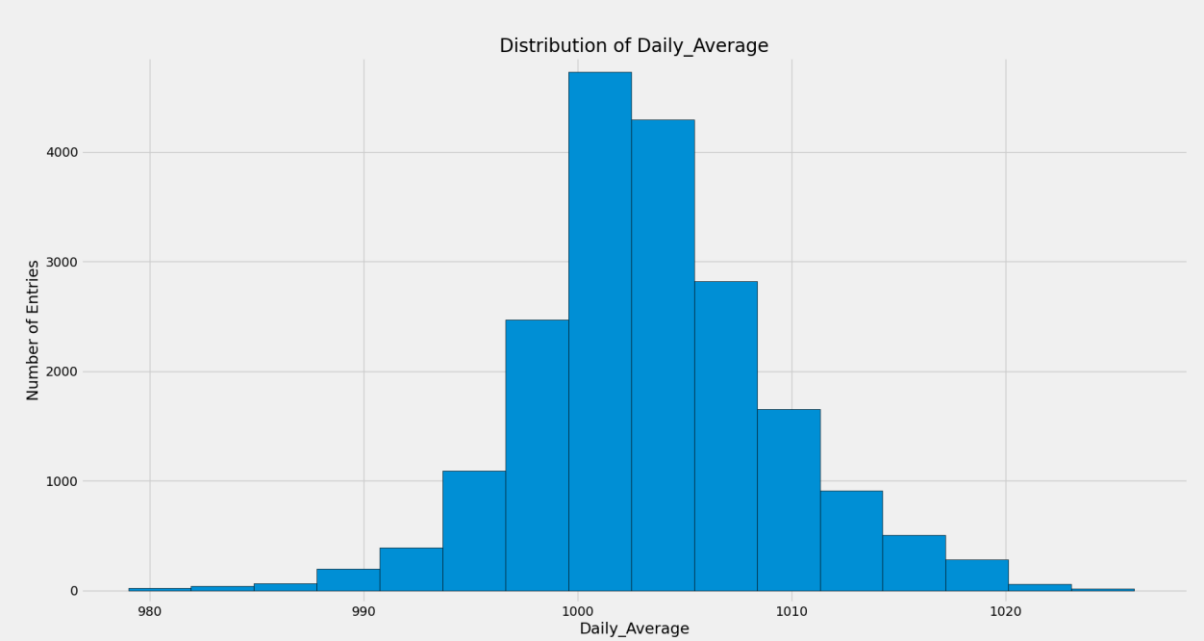

Figure 5.6: Histogram made by using Sturges rule

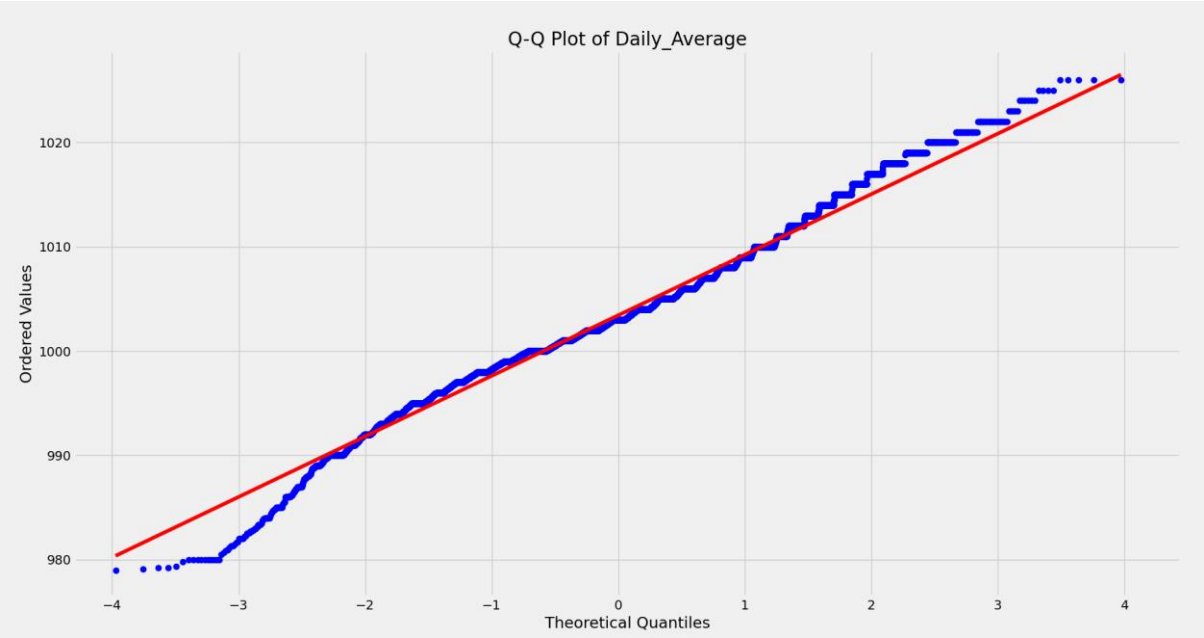

Figure 5.7: QQ-plot of normal distribution and the dataset

Two histograms were made with two different rules for bin sizes. This was done in order to better showcase the distribution. In Figure 5.5, there is the resemblance of a normal distribution, however, the histogram is not smooth and has many spikes at certain ranges. Despite this fact, the spikes still follow a pattern that indicates normal distribution. In Figure 5.6 however, its much clearer that the distribution resembles normal distribution. This due to the fact that Sturges rule smooths the histogram as it has fewer total bins compared to Freedman-Diaconis rule. Nevertheless, we see a pattern of normal distribution which makes sense considering the nature of the data. Since, outliers are already eliminated in the previous steps, there are no outliers present in the histograms. Finally, once the QQ-plot in Figure 5.7 is investigated, it is safe to conclude that the data at hand is normally distributed.

5.3.1. Seasonality

To identify seasonality, the run sequence plot analysis will be used, similar to both Pal & Prakash (2017) and Narasimha et al. (2017). Due to the quantity of the data, a run sequence plot showcasing all of the data were not possible plot, because of readability concerns. Therefore, three different run sequence plots have been plotted, that cover 3 random 2 year periods.

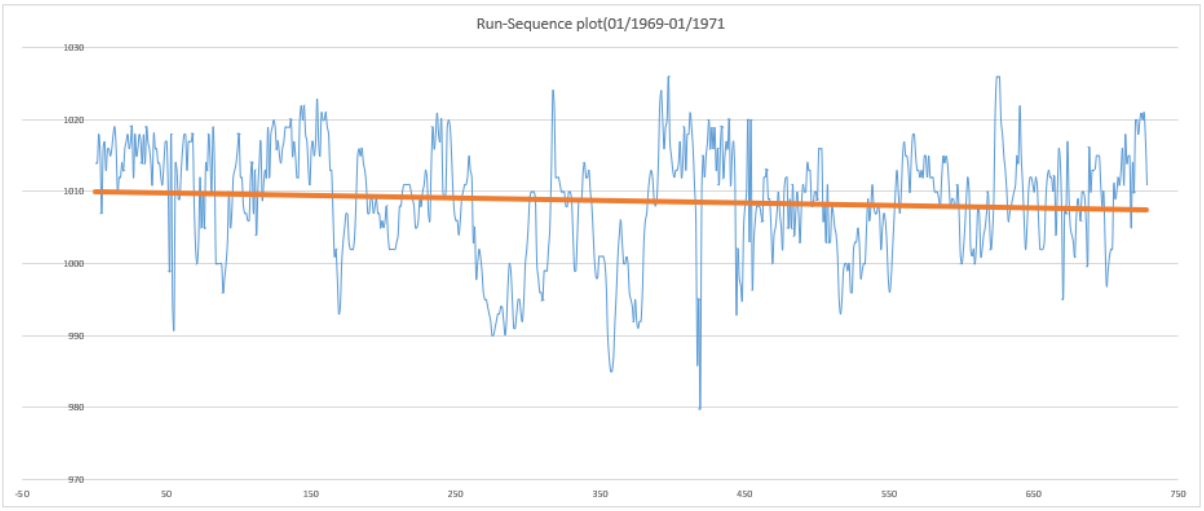

Figure 5.8: Run sequence plot from 01/1969 to 01/1972

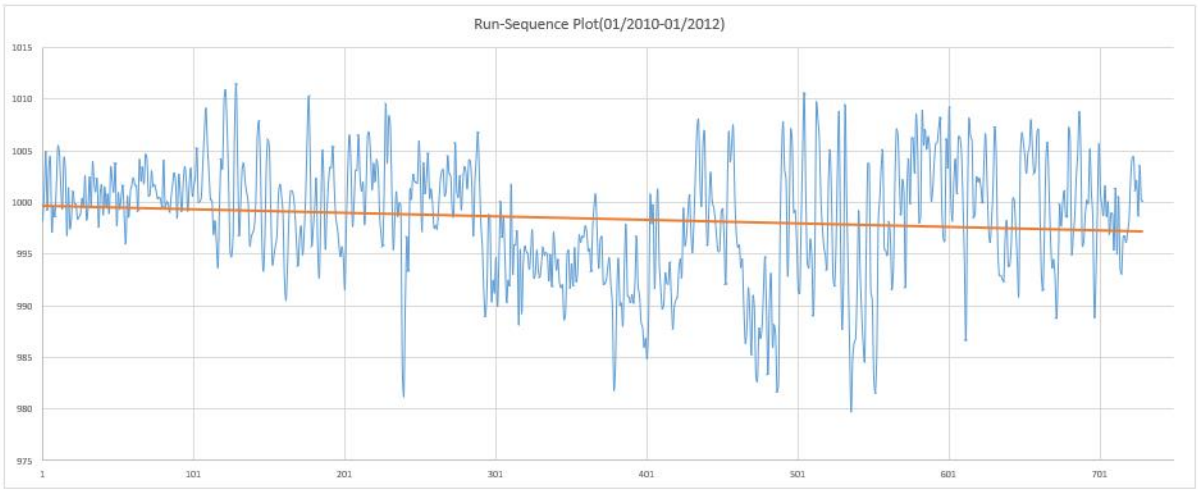

Figure 5.9: Run sequence plot from 01/2010 to 01/2012

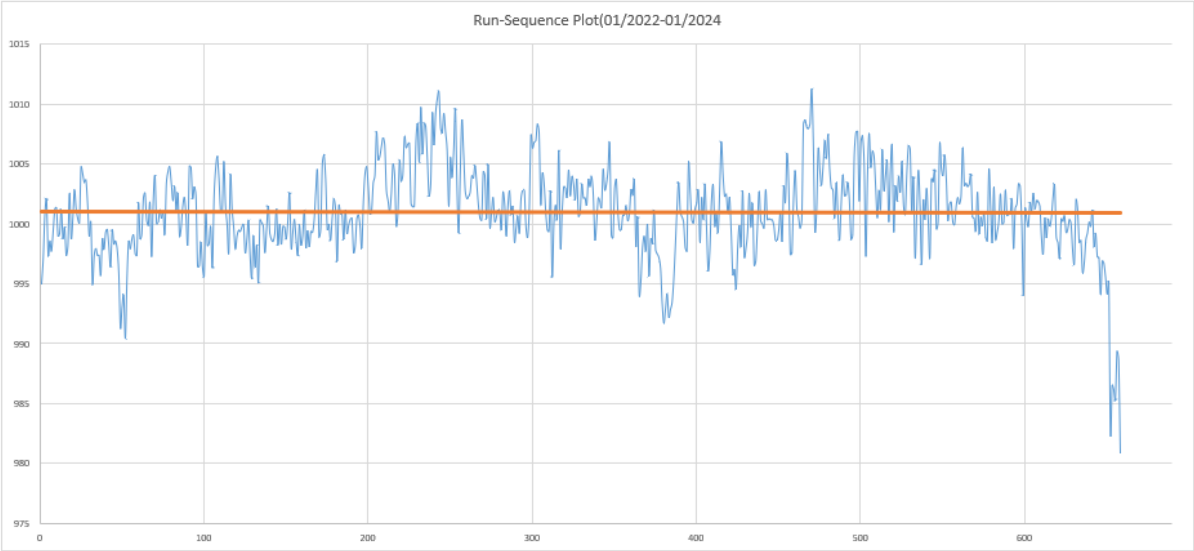

Figure 5.10: Run sequence plot from 01/2022 to 01/2024

In all three figures, various ups and downs as time progresses can be observed. These ups and downs seem to continue repeating as the time progresses, which is observed in seasonality. This corresponds with the definition given by Pal & Prakash (2017), which defines seasonality as repetitive and periodic divergences. However, these ups and downs are not strong enough to confirm seasonality exists for certain. Considering the nature of the data, it makes sense for various seasonal factors such as precipitation and temperature to influence the data. On the other hand, one must consider that Twente canal is a man-made canal which has pumping stations that may be used in the time of unexpected seasonal events. Unfortunately, it is not possible to say what is the specific case is for Twente canal due to lack of domain knowledge in hydrology. Nevertheless, according to California Water Boards (2018), it is possible for channelization activities to alter water temperature and flow.

In addition to the fact that seasonal factors could influence the canal, the evidence provided in section 5.4, was not able to find any significant correlation with the only qualified variable that exhibited seasonality, as the other variables considered were not eligible (see section 5.4 for further explanation). Additionally, when the data were aggregated into monthly average water levels, there were no significant deviations from the mean value of the new series as seen in Table 5.5. Therefore, it can be concluded that seasonality does not exist, or it is not strong enough to warrant the use of a model with seasonality incorporated into it.

Month	Average	Difference
January	1003.0599	-0.70
February	1002.6486	-1.11
March	1003.3345	-0.42
April	1003.9964	0.24
May	1004.7481	0.99
June	1003.7406	-0.02
July	1004.4946	0.74
August	1004.5865	0.83
September	1003.6578	-0.10
October	1003.5798	-0.18
November	1003.7674	0.01
December	1003.2851	-0.47

Table 5.5: Monthly average water level and difference from its mean value

5.3.2. Stationarity

When the run sequence plots in section 5.3.1 are checked, it is possible to see that the values linger around the mean value of 1003cm, but the trend line decreases slowly. However, run sequence plots are not enough to prove stationarity exists. Therefore, the statistical tests mention in section 3.2.2 are utilised, which are ADF, PP and KPSS tests. Both tests are run in Python, using the arch library, see Appendix F for the code. See Table 5.6 for the results.

	ADF	PP	KPSS
Test Statistic	-9.753	-80.048	12.591
p-value	<0.01	<0.01	<0.01
Null Hypothesis	Process contains unit root	Process contains unit root	Process is stationary

Table 5.6: Statistical tests for stationarity (series is not differenced)

As seen in Table 5.6, in all the tests we reject the null hypothesis at a significance level of 1%. Due to rejecting the null hypothesis in all the tests, the ADF and PP results contradict the KPSS result as KPSS indicates that there is a unit root while ADF and PP indicate the series is stationary. This implies that the series is difference stationary, and it should be differenced in order to achieve stationarity, which will be done in section 6.

5.4. Exogenous variables

Before modelling is conducted, three different datasets of exogenous variables have to be examined which are daily temperature, daily precipitation and amount of pumped water to the canal. Unfortunately, the dataset containing the pumped water measurements gathered from Rijkswaterstaat was too large to conduct a full analysis of it. Therefore, the summer of the drought period in 2018 was examined to see if there was significant amount of pumped water to be considered as an exogenous variable. The period consists of days between 01/08/2018 and 01/12/2018, with an observation being made every 10 minutes. Out of 53136 observations only 48 of them were non-zero values. Additionally, all of these observations were unrealistically high values exceeding the measurement method which is the amount of time the pump runs over the previous 5 and next 5 minutes. Therefore, it was concluded that the pump data could not be used in the models.

Secondly, the precipitation data were examined gathered from KNMI at the station Deelen was examined. Unfortunately, the data exhibited many outliers observed using the methods shown in the section 5.3. More than 46% of the data were observed was identified as outliers, even when the outer fences were used. This is most likely due to precipitation not being continuous over days resulting in many days with no precipitation and many days with very high amount of precipitation, causing the many values to be away from the median value of 0.2mm. Therefore, the high amount of data that has to be excluded prevents this dataset being used as an exogenous variable as well.

Finally, the daily mean temperature data from KNMI observed at the station Deelen was examined for its suitability. After the data were cleaned, the quality of was sufficient level with a relatively low number of values being excluded as erroneous values or outliers. Therefore, the correlation between the temperature and daily average water levels was calculated. The resulting correlation was 0.0681. The level of correlation was not sufficient enough to justify that the inclusion of temperature data as an exogenous variable would add valuable information to the model.

In conclusion, the three datasets containing variables that could be added as exogenous variables to the models was examined. Unfortunately, there were no variables that could either be used due to quality of data or could add meaningful value to the training phase of the model. However, despite these results an exception made later in chapter 7 for the purposes of this research and is explained in that chapter.

5.5. Summary and conclusion

This chapter provides both crucial insights and an initial transformation of data that are going to be determining factors for decision taken in the later stages of research. At the start of the chapter, data was gathered from Rijkswaterstaat. The data gathered was the water level data for the reasons mentioned in chapter 4. A description of the data was provided with relevant columns for the research. An initial cleaning of the data was made to remove the unnecessary columns and erroneous values in the dataset. After that, the properties of the data were explored mainly on the concepts of distribution, stationarity and seasonality. The data was deemed, normally distributed, non-stationary and non-seasonal. Finally, several exogenous variables were considered for modelling and were briefly explained.

6. Data Preparation

In this chapter, the datasets will be transformed in order to make them suitable for the models and also easier to interpret the results of the models. This includes achieving stationarity in the datasets and also getting rid of the unusually high water level observation. The latter issue was explained in detail in section 5.1 and is caused by the bottom of the canal being higher than the sea level. The methods and statistical tests used are shown in this chapter with their respective results. Additionally, the reasoning behind not imputing new values instead of missing ones is explained in this chapter. As a side note, the cleaning done in chapter 5 is usually the part of this phase. However, to get a better understanding of the data, some initial cleaning had to be done as the data were in 10 minute intervals which was not going to be useful for the purposes of this research. Therefore, the data were converted to daily average values in chapter 5 instead of this one to understand the properties of the dataset in a way which could help the purposes of the research. See Appendix F for the code.

6.1. Division of the dataset

As mentioned in section 5.2, the dataset contains three different measurement methods. Two of which are readings over 10 minutes, and another one which is a visual reading. The visual readings are present in the data when the data were still being measured once a day. Therefore, they do not represent an average water level per day, but a single reading at any time during that specific day.

Due to different measurement methods present in the dataset, it was decided to create an additional dataset which does not contain the visual measurements but only the measurements taken over the duration of 10 minutes. Specifically, the dataset is from 3/11/1996 to 01/01/2024, which is the period where the measurements are measured over the course of 10 minutes in total. Before the descriptive statistics were analysed, outliers were excluded from the dataset using the methods presented in section 5.3.

Daily Average	
Mean	1001.389
Standard Error	0.04611
Median	1001.5
Mode	1001.042
Standard Deviation	4.459359
Sample Variance	19.88588
Kurtosis	1.136725
Skewness	-0.34864
Range	37
Minimum	982
Maximum	1019
Sum	9365993
Count	9353

Table 6.1: Descriptive statistics of dataset excluding visual readings

As seen in the descriptive statistics, the mean value has decreased to 1001.389 cm from 1003.433 cm. This may be due to factors such as climate change or the visual readings being less accurate. Another important point is that the standard deviation and the variance has

been decreased. Again, this may be the result of more accurate readings or that the water levels in the canal getting more stable due to additional pumping stations being added.

One thing to underline in the newly created dataset is the number of outliers. Due to observations being closer to each other compared to the original dataset, which reflects to the decrease in standard deviation, the upper and outer fences for excluding outliers have become closer. This results in a higher number of outliers in the dataset. In the newly created dataset, there are 568 outliers identified, which corresponds to 5.72% of the dataset. There are no established criteria regarding how much missing data are acceptable to produce statistical analyses. There are arguments that less than 5% is acceptable, whereas some argue that the data are biased when more than 10% data are missing (Dong & Peng, 2013). In the context of this report, the objective is to test different models to give a comparison and see if a time series analysis made using these models would give accurate results for predicting the water levels in the canal. Therefore, considering the percentage of the data that has been excluding is less than 10% and close to 5%, it is reasonable to test this new dataset as well to see if it would give more accurate results in the same models.

6.2. Differencing

As discussed in section 5.3.2, the data has to be differenced in order to achieve stationarity in the series. Achieving stationarity is important as many models assume that the data are stationary when processing the data.

	ADF	**PP**	**KPSS**
Test Statistic	-30.362	-361.113	0.012
p-value	<0.01	<0.01	0.997
Null Hypothesis	Process contains unit root	Process contains unit root	Process is stationary

Table 6.2: Statistical tests for stationarity (series is differenced once)

As seen after the series is differenced once in Table 6.2, the results of the tests are complimentary. The null hypothesis of ADF and PP are still rejected, implying that there is stationarity. Moreover, the contradictory result of the KPSS test while the series was not differenced is now supporting the ADF and PP results as we are not able to reject the null hypothesis of stationarity.

6.3. Subtracting the floor height

As mentioned in section 5.1, the measurements at hand are unusually high due to the floor heigh being higher than the sea level. To address this issue, Rijkswaterstaat was inquired to learn what is the floor height at Eefde. A map was provided by Rijkswaterstaat, which shows the floor height of various waterways in the Netherlands. The map provided is available as Rijkswaterstaat Bathymetrie Nederland.

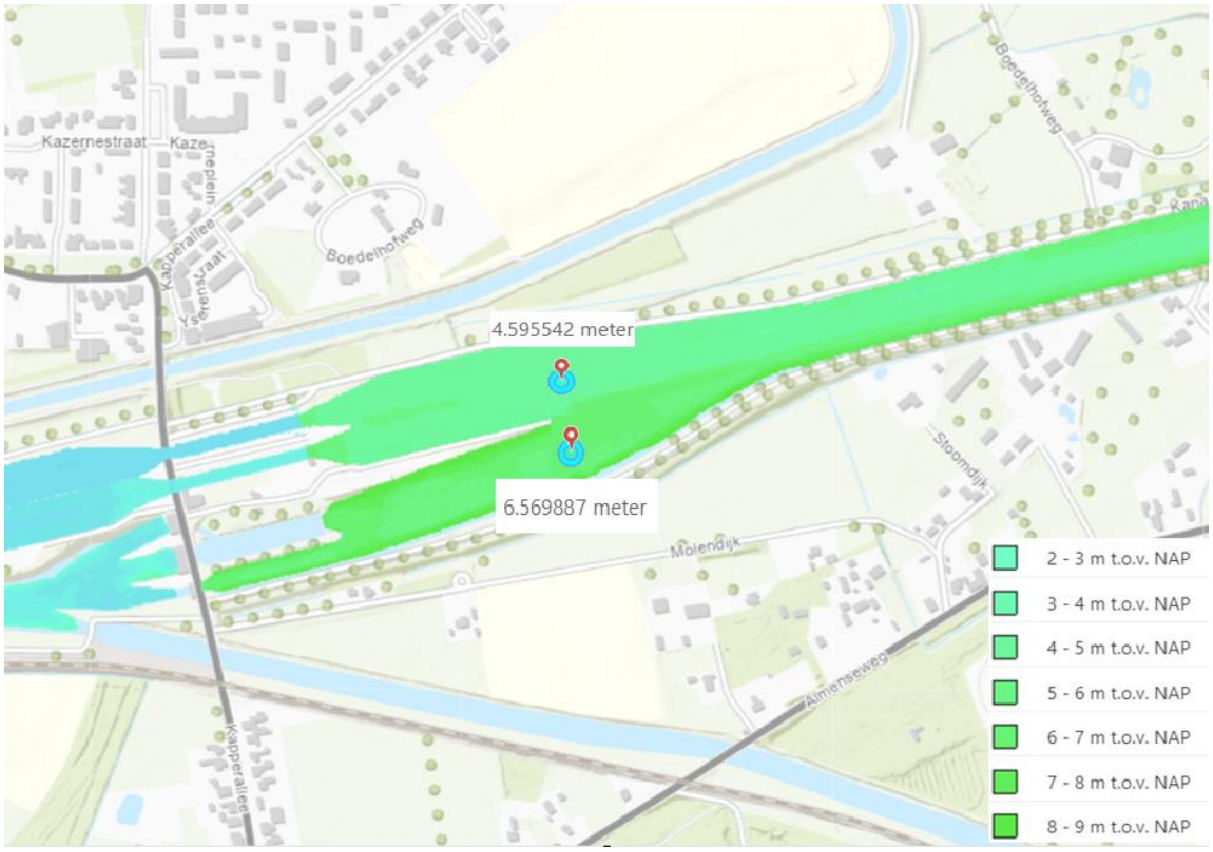

Figure 6.1: Bahthymetrie Nederland map, points are not exact and visual is adjusted for readability (Rijkswaterstaat, 2024.)

As seen in the map, more clearly in the online tool, the floor height near the “Eefde boven” measurement point ranges between 4.5m to 6.5m. Unfortunately, it is not possible to determine the exact point where the measurement tool is, therefore an assumption of 5.5m will be taken. Therefore, a flat 550cm is subtracted from all the data points. Through this operation, the results of the models will be much clearer to read to the end user.

6.4. Data imputation

As mentioned in section 5.3, there were in total of 530 data points missing in the original dataset, 568 in the second dataset excluding the visual readings. There are various methods to impute data such as filling with median, mean, last observation carried forward, linear interpolation etc. However, another option is to drop the missing data from the dataset. The main issue with dropping the values is loss of valuable information that could further improve and help training the model. However, imputing data also has its cons as it could introduce bias to the dataset and create biased predictions that would have been otherwise not present in the dataset.

In the current datasets, considering that they contain a relatively low percentage of missing values, it was decided to not impute new data. It was decided that the possibility of introducing bias to the dataset was not desired and that the loss of information would be negligible. As the percentage of missing data were 2.64% and 5.72%, in the original and the secondary dataset excluding the visual readings respectively, it was concluded that the loss of important information would be covered up by the large quantity of available data. Therefore, it was decided to drop the missing data in favour of not introducing bias to the dataset.

6.5. Summary and conclusion

In this chapter, different data transformations took place to make the dataset ready for modelling in chapter 7. Firstly, the dataset was divided into two depending on the measurement methods used. One dataset contained all the observations whereas the other one excluded the visual reading method. Moreover, the dataset was differenced to transform the data into stationary as explained in section 3.2.2. Afterwards, the floor height which inflated the measurements has been subtracted from the data to make it easier to read for the end user. Finally, the decision regarding whether to impute or no impute data into the dataset was explained in detail in section 6.4.

7. Modelling

In this section, decisions behind choosing models are discussed. Mainly, how models are chosen based on the previously discussed points in this report and the decision tree is discussed. Additionally, model parameters are estimated based on the theory provided in chapter 3. Finally, the estimated parameters are fir onto the models and evaluated using information criterion for comparison and validation. See Appendix F for the code.

7.1. Selected models

Following the theory provided in chapter 3 and the decision tree in section 3.5, two models has been chosen for modelling. As tested in section 5.3.2, the dataset exhibited non-stationarity, thus it was differenced in section 6.2. Therefore, models chosen must be able to incorporate differencing into them. Additionally, it was concluded that the data does not show seasonal properties, or these properties are not strong enough in section 5.3.1. Thus, the use of models such as SARIMA and SARIMAX would not add any value to the accuracy of the forecast compared to models such as ARIMA and ARIMAX. Finally, as explained in section 5.4 it is expected that the use of exogenous variables would not add any value, but for the purposes of this research, only one model of ARIMAX using temperature as exogenous variable will be compared to the best performing ARIMA model to see if it does not add value as expected. This is done in order to test as many models as possible and since seasonal models are not possible to be tested, it is the only other model possible for testing.

Following this information and the decision tree, the ARIMA was chosen for forecasting the time series. The ARIMA model allows for differencing the model thus getting rid of the non-stationarity within the data. Moreover, since SARIMA and SARIMAX models are not needed to be used due to seasonality not being a factor, it is the only valid model according to the decision tree proposed. As a side note, due to the data being daily, a seasonality component of 365 would most likely be problematic if used. These models were generally used combined with monthly data instead of daily data, as seen in the case of Narasimha Murty et al. (2017). The model will be tested for 2 different datasets, with different parameter settings.

7.2. Choosing parameters

To choose the parameters, ACF and PACF plots will be used as explained in chapter 3. The differencing order of 1 was already decided in section 6.2. Due to the p-values calculated as a result of the tests performed in section 6.2, which were very low for ADF and PP tests, and high for KPSS tests, any additional differencing may result in over differencing the series. To choose the order of autoregressive(AR) and moving average(MA) for ARIMA, the ACF and PACF plots of the series was examined after differencing by the order of 1. In Figure 7.1 and Figure 7.2, ACF and PACF plots are shown respectively.

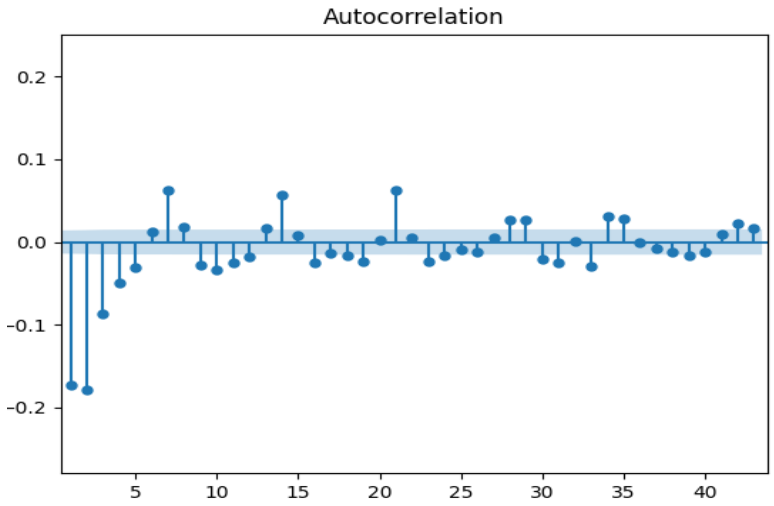

Figure 7.1: ACF plot of the data from 1969-2024 (d=1)

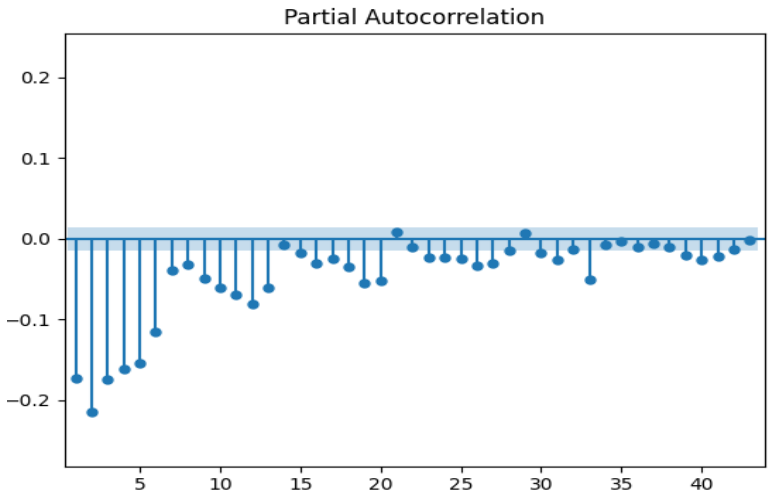

Figure 7.2: PACF plot of the data from 1969-2024 (d=1)

In the figures, the first lag is ignored, and they start from the second lag as the first lag is the correlation of the series with itself, resulting in a measurement of 1. When the ACF plot is examined, it can be noticed that the autocorrelation starts to be insignificant after lag 2, and another significant drop occurs after lag 3, 5 and 7. The PACF plot shows that after lag 6 the correlation drops significantly as well. Therefore, models with the settings ARIMA(6,1,2), ARIMA(6,1,3) and ARIMA(6,1,3) will be tested. Similarly, the other dataset which excludes the visually read observations is also examined. The ACF and PACF plots of this dataset are shown in Figure 7.3 and Figure 7.4.

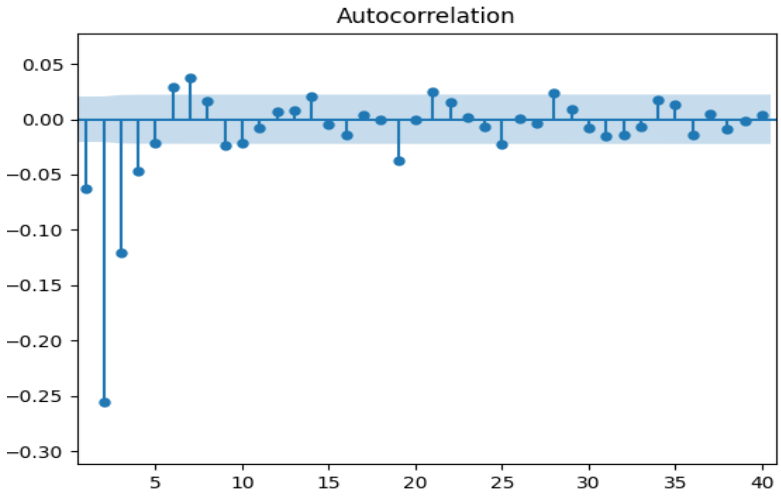

Figure 7.3: ACF plot of the data from 1997-2024 (d=1)

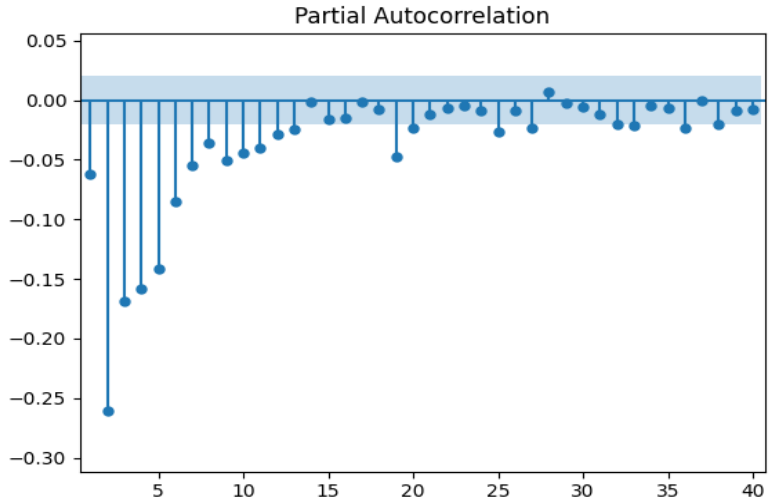

Figure 7.4: PACF plot of the data from 1997-2024 (d=1)

As seen in the Figure 7.3, the autocorrelation becomes dropped significantly after lags 2, 3 and 4. In a similar way, PACF autocorrelation becomes less significant after lags 2, 5, 6 and 7. Therefore, ARIMA models ARIMA(2,1,2), ARIMA(2,1,3), ARIMA(5,1,2), ARIMA(5,1,3) ARIMA(6,1,2), ARIMA(6,1,3), ARIMA(7,1,2) and ARIMA(7,1,3) will be considered.

7.3. Fitting models

Due to the number of models proposed in section 7.2, models will be compared based on their information criterion such as Akaike Information Criterion(AIC) and Schwartz's Bayesian Information Criterion(BIC). The model with the smaller information criterion will be preferred as it is thought be a better fit for the data (Narasimha Murty et al., 2017). The results of the models are presented in Table 7.1 and Table 7.2.

MODEL	AIC	BIC
ARIMA (6, 1, 2)	*105826.061*	*105897.232*
ARIMA (6, 1, 3)	105984.553	106063.632
ARIMA (6, 1, 5)	106068.165	106163.060
ARIMA (6, 1, 7)	**105771.111**	**105881.821**

Table 7.1: Comparison of information criterions (1969-2024)

MODEL	AIC	BIC
ARIMA (2,1,2)	48228.025	48264.037
ARIMA (2,1,3)	48227.790	48271.004
ARIMA (5,1,2)	48218.596	48276.215
ARIMA (5,1,3)	48213.973	48278.794
ARIMA (6,1,2)	**48120.209**	**48185.029**
ARIMA (6,1,3)	48190.166	48262.189
ARIMA (7,1,2)	*48128.387*	*48200.410*
ARIMA (7,1,3)	48167.303	48246.528

Table 7.2: Comparison of information criterions (1997-2024)

From the AIC and BIC, the models that exhibit the two lowest values will be chosen for forecasting the series. For the original dataset, these models are ARIMA(6,1,2) and ARIMA(6,1,7) as seen in Table 7.1. For the dataset that excludes the visual readings, the chosen models are ARIMA(6,1,2) and ARIMA(7,1,2). Additionally, due to choosing the parameters (6,1,2) in both datasets, will allow for a comparison to be made for the difference between the measurement methods, although at a limited level as ARIMA(6,1,2) is not lowest AIC and BIC value in the original dataset, thus is not the most optimal one by the criterion comparisons.

7.4. Summary and conclusion

At the end of this chapter, several models which will be tested are identified successfully. These models were identified using the theory presented in chapter 3 and information criterions. The models with the lowest information criterions were chosen. From each dataset, two models were chosen for the purposes of this research and testing.

8. Evaluation

In this section, the models that were chosen in section 7.3 will be evaluated and compared based on their performance on various error metrics. According to their results, it will be determined if the models are suitable for use in business operations or not. Three different ways of forecasting were done. Firstly, a forecast of the last 100 observations of the timeseries was made using all the previous datapoints up to last 100 observations. Secondly, a forecast of last 4 weeks was made using all the previous data up to last 4 weeks. Finally, a rolling forecast was made using all the data up to last 100 observations and adding each new observation after that to update forecast. Reasonings behind choosing these types of forecasts are explained further in each section.

8.1. Long periods of forecasting

Firstly, the longer periods of 4 weeks and last 100 observations were examined. During the testing process, some values at the end of the dataset were removed due to being missing values which make it harder to visualise the plots. The results are displayed in the figures below.

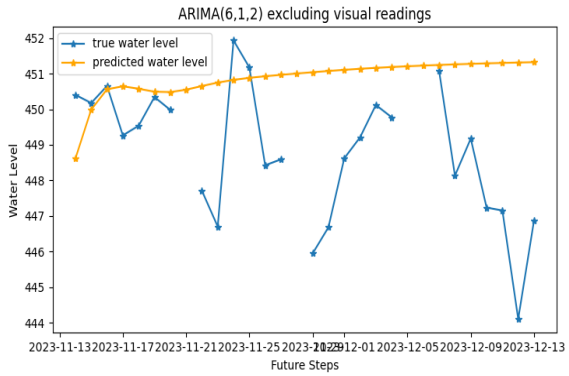

Figure 8.1: ARIMA(6,1,2) without visual readings, period 4 weeks

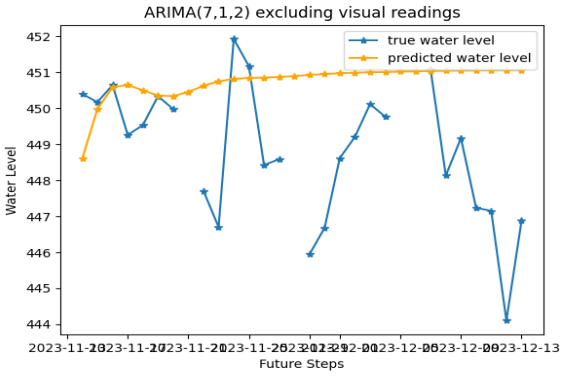

Figure 8.2: ARIMA(7,1,2) without visual readings, period 4 weeks

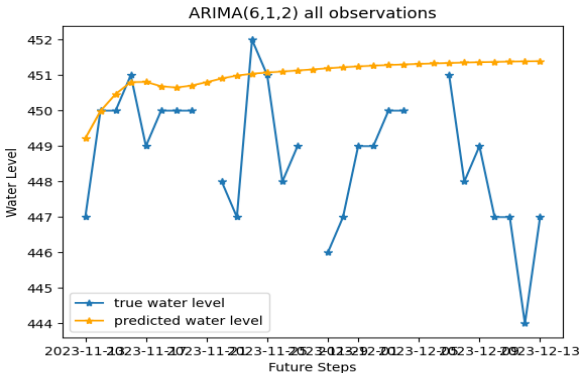

Figure 8.3: ARIMA(6,1,2) all observations, period 4 weeks

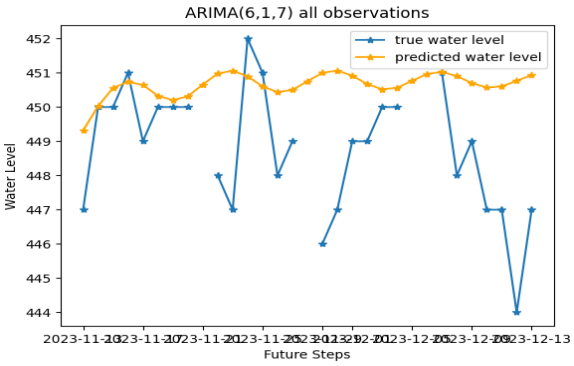

Figure 8.4: ARIMA(6,1,7) all observations, period 4 weeks

As seen in the figures, despite the fact that the first few days are accurate, after a short period of time the values start to become flat towards mean. This may be due to several factors such as low variance in the dataset, or seasonal factors not being incorporated into the models. However, as discussed in section 5.3.1, there was no evidence suggesting there exists seasonality on a statistical basis. The figures for the last 100 observations are presented in the Appendix G, which also produce a similar result. More on the failure of these models in a long period will be discussed in section 8.3.

8.2. Rolling forecasts

The models in this section are dynamic models that include each new observation into the model to predict the next one. This has the benefit of increased accuracy as the real value of every previous observation is known before the next prediction. However, due to recalculating the models at each new real input, the models take much more time compared to static models. The results of the models are shown in the figures below.

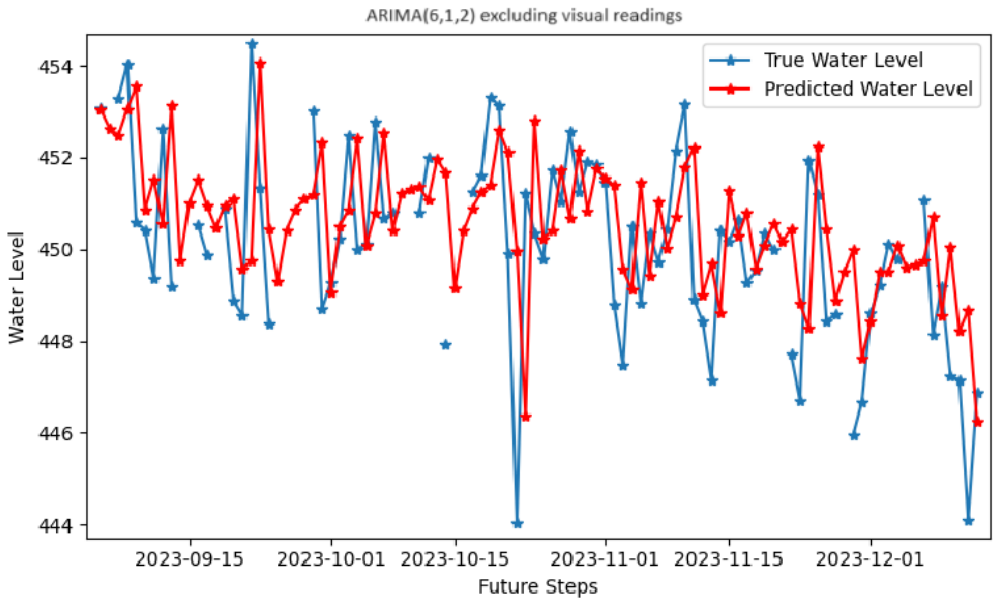

Figure 8.5

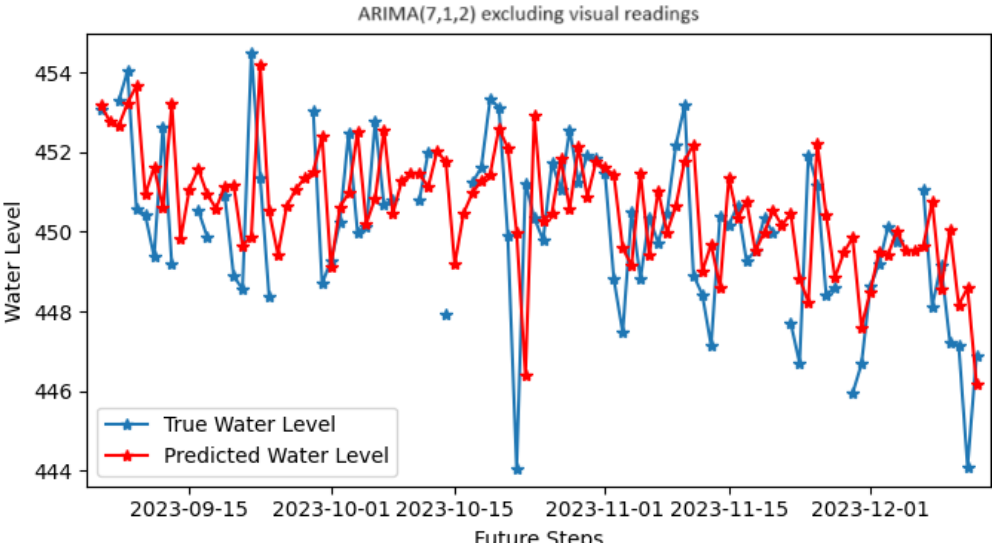

Figure 8.6

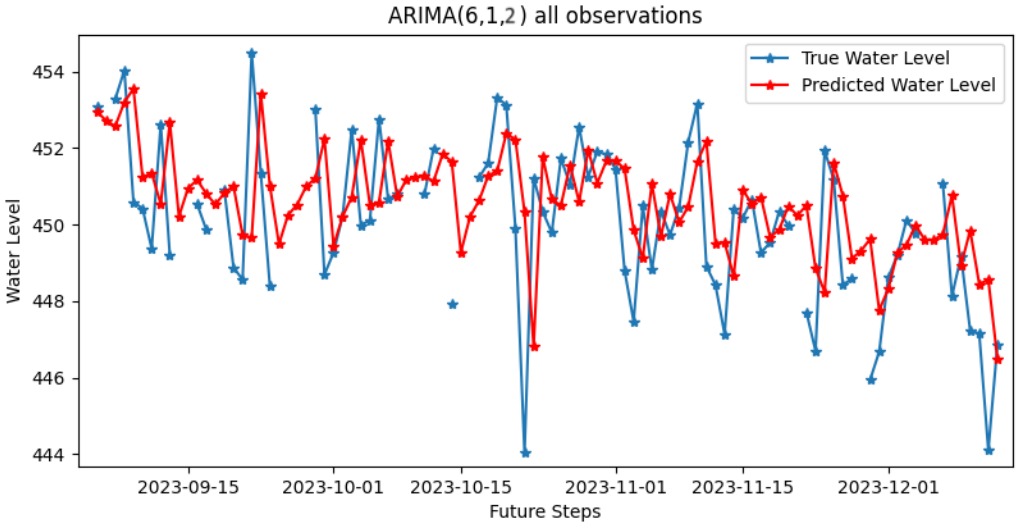

Figure 8.7

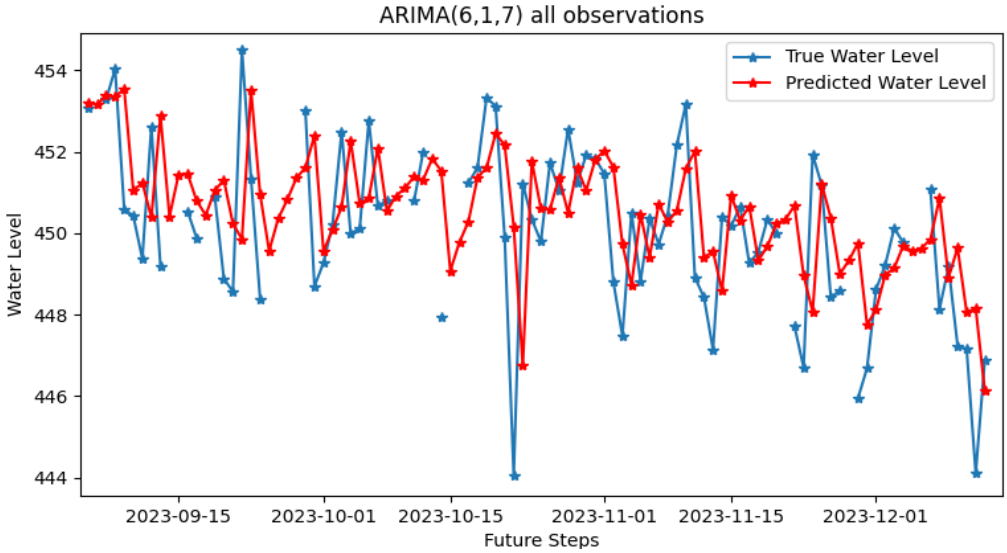

Figure 8.8

	All observations		Excluding visual readings		ARIMAX
Model	ARIMA(6,1,2)	ARIMA(6,1,7)	ARIMA(6,1,2)	ARIMA(7,1,2)	ARIMAX(6,1,7)
RMSE	2.0002	**1.9836**	*2.0390*	2.0451	2.0057
MSE	4.0001	**3.9347**	*4.1574*	4.1827	4.0229
MAPE	**0.0008**	0.0016	*0.0014*	0.015	0.3448
Runtime	*1695 seconds*	2771 seconds	**917 seconds**	1020 seconds	2778 seconds

Table 8.1: Error metrics for the models (bold are the best overall, and italics are best for each dataset)

Overall, the lowest error metrics were achieved in the dataset containing all the observations as seen in Table 8.1. This indicates that the difference between the measurement methods is not significant enough to influence the accuracy of the model, whereas the amount of data used in training the model which is much larger in the original dataset is of significant value. However, this increased accuracy comes with the drawback of increased runtimes as can be seen in Table 8.1. Despite the ARIMA(6,1,7) for all observations performing better than its counterpart ARIMA(6,1,2) in several metrics, the increased runtime of approximately 18 minutes makes the use of ARIMA(6,1,2) justifiable. In addition to this, the MAPE value of

ARIMA(6,1,2) is lower than ARIMA(6,1,7) and it's not too far off from ARIMA(6,1,7) in other metrics as well. Therefore, ARIMA(6,1,2) seems to be the optimal model for achieving both performance and efficiency.

To interpret the error metrics of ARIMA(6,1,2), their implications must be considered. RMSE indicates how many units the actual and predicted values are off from each other. An RMSE of 2 implies that the values predicted are 2cm away from the actual values. Essentially, MSE is RMSE but squared, however, due to its squared nature it punishes the values that are further away from the actual values harder. Therefore, it must be noted that the ARIMA(6,1,7) produced less values that were far from the actual values, which indicates that for the highest accuracy one can use ARIMA(6,1,7) over ARIMA(6,1,2) if efficiency is not a priority. Finally, MAPE is a type of relative error which shows how far away the predictions are from the original values. This metric is very low for all the models, but this most likely due to the fact that the higher the value of the values lower the MAPE. This effect is caused by the division of the errors by the original value, which makes the metric smaller if the actual values are high, which is the case in this dataset as well.

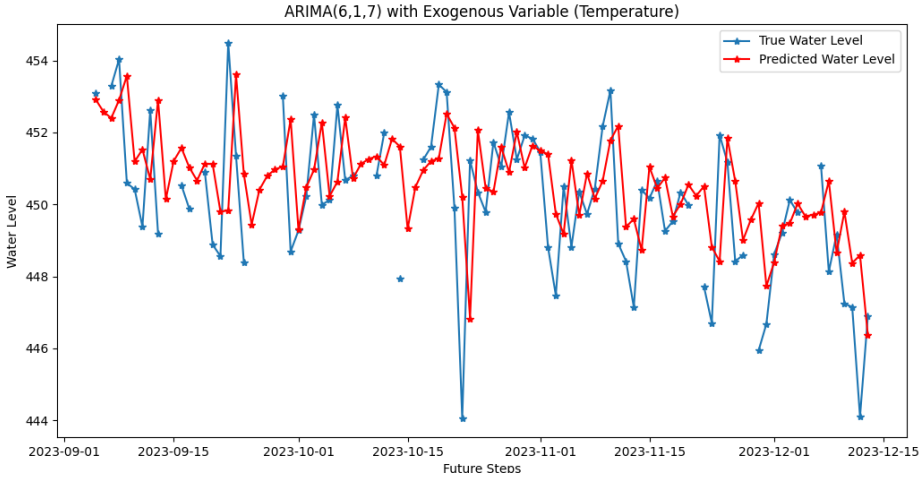

Figure 8.9

Finally, the ARIMAX(6,1,7) model was also fitted to data in Figure 8.9 and its error metrics were noted in Table 8.1. As expected, the addition of the exogenous variable has not improved the model, nor it did add any valuable information to the model. Therefore, it has been confirmed that the correlation between temperature and the water levels in the canal was indeed not high enough for it to influence the results.

8.3. Discussion of the results

As stated in section 8.1, static models that forecast long periods of time do not give accurate results. Therefore, the only way to use the static models are for shorter periods of time where they tend to produce more accurate results that imitate the pattern of actual values.

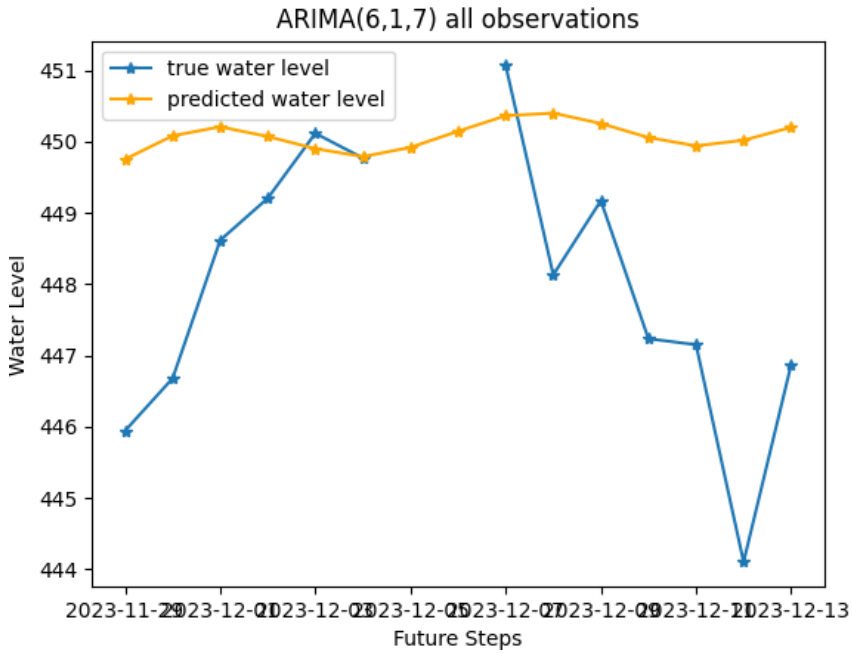

Figure 8.10: ARIMA(6,1,7) all observations for period of 15 days, static model

As seen in Figure 8.10, the ARIMA (6,1,7) model was able to predict the increase and decrease in water levels for approximately 9 days from the start, albeit at a limited accuracy. The error metrics were: 7.55, 2.75 and 0.74% for MSE, RMSE and MAPE respectively. Despite RMSE being not too far from other models' error metrics shown in Table 8.1, the important metric here is the MSE as it punishes the values that are further away harder. As seen in Figure 8.10, despite the unit differences not being too big, due to low variance in the data, the values after the 9th day become far away from the original values. After the 9 day period, the model starts to become stable around the mean. This effect will be further discussed in chapter 10.

Due to these factors, it is not possible to use the models to predict long term water levels. However, it is possible to make short term predictions accurately. In practice, this may allow the companies to start looking out for other means of transport if there is a prediction for water levels to drop. Despite, the period being short, at least emergency measures would have been able to be taken.

On the other hand, the rolling forecasts were very successful and accurate at predicting the water levels. However, the use of these models is also limited due to the fact that they essentially produce the most accurate results for the very next observation. This means that, the businesses would have only 1 day to change their operations if they want to rely on the most accurate prediction. However, it may still be beneficial as it would allow for rescheduling of emergency shipments from the next few days to the current day if a drastic drop in water levels is expected. Additionally, if less accuracy could be tolerated, the rolling forecasts could provide a 3-4 day buffer zone for operational changes if water levels are expected to drop as they are essentially constantly updating static forecasts which provide less but sufficiently accurate results. A 3-4 day buffer zone would allow for additional on-call contractors to be hired and bigger schedule changes to be implemented. More on implementation is discussed in Chapter 9.

8.4. Summary and conclusion

This chapter shows how the models were evaluated using the error metrics which were introduced in the theoretical framework presented in chapter 3. The results are later on discussed extensively. There are two main points in this chapter that should be highlighted, which are: the inaccuracy of the models in the long term forecasting and the highly accurate results in short term forecasting. Additionally, the converging behaviour in long term forecasts is to be noted as well.

9. Deployment

In this chapter, recommendations on how the model can be deployed and used are made. The recommendations are derived from the results discussed in section 8.3. It is important to note that the model cannot be deployed at the time of this research due to time constraints.

9.1. Deployment plan

Figure 9.1: Deployment plan

Figure 9.1 shows the steps to be taken in the proposed deployment plan in this section. Firstly, the planning phase will occur. In this phase it is recommended to allocate a budget for deployments, identify which departments and employees would be responsible for each step of the deployment plan and to determine the specific outcome each stakeholder desires at the end of the end of the deployment plan. The desired outcomes refer to, how each company that is part of the research project wants to utilise the models.

After the planning phase is done, it has to be tested in the real environment to determine if it can be used. The amount of testing required entirely depends on how much the model improves after each configuration. This is cyclical step where multiple cycles of testing and configurations depending on the test results have to be made repeatedly. It is recommended that either a time or budget constraint is implemented in this phase to not waste extensive resources trying to improve a model which does not provide the desired results even after a certain amount of testing and configurations.

As to how the models can be tested is mostly dependent on the objectives determined at the planning phase. However, some general ways to test would be to use the model based on past water levels and crate schedules according to the results. If the newly created schedules based on the models add value over the schedules that were used in the past at the same time period, then the model can be deemed valuable. Unfortunately, during the testing phase it may be required to test the model mostly on the drought periods in the past to see if it adds value as these periods do not occur frequently to be tested in the present. Additionally, model could be tested for its accuracy compared with present data. For example, forecasts could be made for a number of days and see how accurate results they produce over these days. This would help to get a better understanding of how many days can be forecasted accurately during the testing phase. Finally, configurations could be made in order to improve the model depending on the outcomes of the test.

In the analysis phase the results of the testing and configuration would have to be evaluated and a decision to continue with integration or not has to be made. Depending on the objectives set in the planning phase, a set of Key Performance Indicators(KPIs) should be made. Some examples of these KPIs could be related to accuracy of the model, readability of the results, financial resources required, and time required. After the model is analysed regarding the KPIs, a cost-benefit analysis could be conducted to decide whether to integrate the model to the operations or to discard it.

Finally, depending on the results of the analysis phase, model should be integrated into the operations systematically. The model should be integrated gradually into the operations to make configurations if necessary. It is important to choose a set of warehouses or businesses

for the model to be integrated and compare this to the set where the model is not integrated in order to analyse the results and not hinder the operations. Additionally, necessary training and information regarding the use of the model and its capabilities should be given to the stakeholders.

Integration could start by providing the model to a set of managers responsible for scheduling in different branches where the model is initially used as a complimentary tool. After this initial phase, it could be observed how did the results of these branches compare to the branches which the model was not given. Afterwards, the set of managers to whom the model was provided should be analysed to see which managers utilised the model best and how they utilised it. Additional feedback regarding the model must be gathered from these managers to determine any configurations to be made. Depending on how managers utilised the tool best, general guidelines and a manual for the model could be designed for further integration. From these points on, it should be business specific case as to whether they would like to integrate the model beyond managerial level or a tool which does not only serve complimentary but as a decision-making tool.

9.2. Monitoring and Maintenance

After the integration phase is done, the model must be monitored for its accuracy and validity periodically. As the water level data is based on dynamic natural factors, it could be the case that the model would have to be configured from time to time. These configurations would have to be tested before replacing the old model to not hinder the operations. These configurations could be related to updating the dataset, addition of variables or change of parameters. Additionally, the tool which model is implemented into should be maintained in terms of its user interface and capabilities.

Another recommendation would be to follow the developments and explore other possibilities of modelling. The ones responsible for the maintenance should look into other possible ways to improve forecast accuracy even if it means to discard the current model. As for the maintenance periods, it is unlikely for frequent maintenances to be needed, especially considering the stable nature of the data. Most of the maintenance required would be the result of feedback given or uncommon natural phenomena influencing the behaviour of the water levels. As a consequence, it might be useful to consider that most maintenance required would be unplanned ones rather than planned maintenance while making a schedule.

9.3. Summary and conclusion

In this chapter, recommendations on deployment were given as general outlines. Specifics were not discussed as many businesses take part in the digital twin project and each has to determine the extent of use for the models. Some businesses might decide to make it a core of their operations while others leave it as a complimentary tool. The chapter highlights the importance of a gradual integration and the importance of testing the model before implementation. Successful testing would ensure a smoother integration later on in the deployment phases.

10. Conclusion

In this chapter, concluding remarks regarding the report are made. In section 10.1, the main results and findings of the report are underlined. Section 10.2 identifies the contributions this report has made to both practice and research. Limitations this report was subject to are

discussed in section 10.3. Finally, in section 10.4, further research and development that can be done in the light of findings and the limitations of this report are presented.

10.1.Main results and findings

There are three main findings to be discussed in this section. They are the inaccuracy of the models with the current state of models proposed in this report, high accuracy of short term forecasts with the current state of models proposed in this report and finally the converging behaviour of the long term models. As shown in section 8.1, the long term models fail to predict values after a certain period and start converging towards the sample mean of the dataset they are trained on.

Secondly, the short term forecasts were highly accurate for predicting values as presented in section 8.2. These models were rolling forecast models which update itself at every new observation. They were successfully validated using error metrics and demonstrated that they have the potential to be implemented.

As for the converging behaviour, due to time constraints and the scope of this research, extensive research into why it happens was not made. However, possible reasons as to why it appears were given as unincorporated seasonality and low variance in the dataset.

The research resulted in mixed outcomes overall. The long term forecasts were inaccurate and not deemed usable, whereas the short term forecasts were highly accurate and usable in further steps. However, the extent of how the short term forecasts was arguable due to time horizon needed to prepare for changes in the water levels and the operations. It may not be possible for the businesses to utilise the short term forecasts effectively without additional improvements.

10.2.Contributions

Throughout this research, existing literature was examined and was utilised to solve a practical problem. There are contributions made to both research and practical problems in this report. Firstly, the theoretical framework provided in chapter 3 outlines the most commonly used models in the field of hydrological forecasting using time series analysis methods. This theoretical framework could be used in further research as a basis. It has also provided insights into difference of seasonal behaviour between man-made canals and natural waterways, as in this research there was no evidence of seasonality encountered in Twente canal. However, it is possible that lack of domain knowledge is the reason for this. Finally, it has shown how datasets with low variance behave under long periods of forecasting by converging to the sample mean.

In practice, the research was successful in providing a short term forecast for water levels to be predicted. The forecast provided allows for water levels to be determined in short term which could be helpful to adjust certain operations according to it. Additionally, it has raised the question whether the canal is affected by seasonal phenomena or not. It was demonstrated that a seasonal variable such as temperature was not valuable to the model successfully. Finally, it has shown that either there exist seasonal factors which were not incorporated into the models in this research or variance in the dataset causing the models to converge to sample mean.

10.3.Limitations

In this section, the limitations of the model and report will be discussed. Additionally, points where improvements could be done are going to be explained. Ultimately, this section aims to provide what can be done in order to improve on from this report and its findings.

One of the main limitations of this report is the lack of domain specific knowledge in terms of hydrology. There is much knowledge that a researcher with knowledge of hydrological forecasting and river science could add to this research. There are many unknown factors which were not considered as a part of this research regarding the behaviour of water levels, seasons and weather patterns.

Unfortunately, there were no exogenous variables suitable to model either due to quality of data issues or low corelation factors. The access to higher quality of data for these variables could have been useful to check if the datasets that were eliminated for quality issues could have been added to the models. These exogenous variables could have modelled the seasonal behaviour of the canal if it exists, mainly because the data considered such as temperature and precipitation already exhibit the seasonal properties of the area considered.

Additionally, again due to time constraints, throughout this research only one measurement points “Eefde boven” was considered. Other points were not considered as the amount of data each point could contain and the time limit of the project made it impossible.

Finally, due to time constraints and additional information to be acquired from contact with Rijkswaterstaat authorities, the model used by them was not used in this research. The model used by Rijkswaterstaat is SOBEK model using ARMA correction. Unfortunately, at the time of this research additional inquiries regarding this model were not answered.

10.4.Future research and development

As mentioned in Section 8.1, the static forecasts were not accurate for long periods of time. This is most likely due to two factors, which are low variance in the original dataset or there being seasonality but not incorporated into the model. As mentioned in section 10.3, the identification of seasonality could be done with knowledge in hydrology. While examining the dataset for seasonality, this was done with the assumption of yearly seasonality. However, this may not be the case and due to lack of domain knowledge additional seasonal periods could not be identified. On the other hand, the data exhibits low standard deviation and variance. This may make the model to stabilise relatively quickly, causing the converging behaviour. To tackle this, additional variance could be introduced to the dataset, however this might cause the models to become less accurate.

More on seasonality, as explained in section 5.3.1, there was no statistical evidence of seasonality that is strong enough to justify the implementation of it in the models. However, seasonality can be established with sufficient domain knowledge regarding hydrology as the researcher would be able to use domain specific knowledge. Additionally, a researcher focused on hydrology could identify additional exogenous variables that could be considered, which might result in different models being tested. Ultimately, it is important for any further research to include an expert on hydrology and hydrological forecasting. This is also important for any desired developments on the model to be made.

As discussed in section 10.3, more research regarding SOBEK model and cooperation with Rijkswaterstaat to develop a forecast using SOBEK could yield beneficial results. It is possible that the SOBEK model incorporates additional factors that were not accounted for in this research. The hydrological knowledge which Rijkswaterstaat possesses could also find a

solution to the previously discussed points of converging behaviour of the long term forecasts and determining if there is seasonality or not.

Additionally, measurement points other than “Eefde boven” should be considered for modelling. As stated in section 10.3, it was no possible to consider all measurement points on the canal. Other measurement points could help validating the models or there may be differences between the measurement points in regard to data properties which would entail different models to be considered.

Finally, there were some models that could not be considered in this research such as PARMA and SARIMA(X) models. Unfortunately, there were not many articles on PARMA available and due to lack of literature it could not be considered on latter phases of this report. On the other hand, despite many articles on SARIMA(X) available, due to computational capabilities a SARIMA(X) model with a seasonal component of 365 could not be modelled. It is possible that the seasonal component of 365, which implies yearly seasonality, is wrong due to the points mentioned previously in this section. However, it might be beneficial to test it for further research purposes.

Bibliography

Aalbers, E. E., van Meijgaard, E., Lenderink, G., de Vries, H., & van den Hurk, B. J. J. M. (2023). The 2018 west-central European drought projected in a warmer climate: how much drier can it get? *Natural Hazards and Earth System Sciences*, *23*(5), 1921–1946. https://doi.org/10.5194/nhess-23-1921-2023

AHN. (n.d.). *Home*. AHN. https://www.ahn.nl/

Anderson, P. L., Meerschaert, M. M., & Zhang, K. (2012). Forecasting with prediction intervals for periodic autoregressive moving average models. *Journal of Time Series Analysis*, *34*(2), 187–193. https://doi.org/10.1111/jtsa.12000

Anonymous(c), (2024). *Interview Summary Company 3*(T. Tao, Interviewer) [Personal communication].

Anonymous(a), (2024). *Interview Summary of Company 1* (T. Tao, Interviewer) [Personal communication].

Anonymous(b), (2024). *Interview Summary Company 2* (T. Tao, Interviewer) [Personal communication].

Banaś, J., & Utnik-Banaś, K. (2021). Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting. *Forest Policy and Economics*, *131*, 102564. https://doi.org/10.1016/j.forpol.2021.102564

Bergen, J. P. (n.d.-a). *Meeting 2: Is science apolitical?* Retrieved May 27, 2024, from https://canvas.utwente.nl/courses/14110/pages/meeting-2-is-science-apolitical?module_item_id=471202

Bergen, J. P. (n.d.-b). *Meeting 3: Is engineering research scientific?* https://canvas.utwente.nl/courses/14110/pages/meeting-3-is-engineering-research-scientific?module_item_id=471203

Bergen, J. P. (2024, March 19). *Meeting 5: Responsibility/ies in research*. https://canvas.utwente.nl/courses/14110/pages/meeting-5-responsibility-slash-ies-in-research?module_item_id=471205

Birinci, V., & Akay, O. (n.d.). A Study on Modeling Daily Mean Flow with MLR, ARIMA and RBFNN . *BALWOIS*. Retrieved May 27, 2024, from https://balwois.com/wp-content/uploads/old_proc/ffp-1948.pdf

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis : forecasting and control* (5th ed.). John Wiley & Sons.

Brownlee, J. (2020, August 20). *Probabilistic Model Selection with AIC, BIC, and MDL*. Machine Learning Mastery. https://machinelearningmastery.com/probabilistic-model-selection-measures/

California Water Boards. (2018). *Nonpoint Source Pollution (NPS) Control Program Encyclopedia | California State Water Resources Control Board*. Www.waterboards.ca.gov. https://www.waterboards.ca.gov/water_issues/programs/nps/encyclopedia/5_1a_chnlmod_c

hnlz.html#:~:text=Channelization%20and%20channel%20modification%20activities%20diminish%20the%20quality%20and%20diversity

Cheng, Q., Argon, N. T., Evans, C. S., Liu, Y., Platts-Mills, T. F., & Ziya, S. (2021). Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, *48*, 177–182. https://doi.org/10.1016/j.ajem.2021.04.075

De Figueiredo, N. M., & Cavalcante Blanco, C. J. (2016). Water level forecasting and navigability conditions of the Tapajós River - Amazon - Brazil. *La Houille Blanche*, *3*, 53–64. https://doi.org/10.1051/lhb/2016031

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1). https://doi.org/10.1186/2193-1801-2-222

Doran, G. T. (1981). There's a SMART Way to Write Management's Goals and Objectives. *Management Review*, *70*(11), 35–36.

Dubey, A. K., Kumar, A., García-Díaz, V., Sharma, A. K., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments*, *47*, 101474. https://doi.org/10.1016/j.seta.2021.101474

Gupta, S. K., & Agarwal, A. (2021, February 19). Predicting Total Sugar Production Using Multivariable Linear Regression. *International Conference on Computing, Communication, and Intelligent Systems*. https://doi.org/10.1109/icccis51004.2021.9397078

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., & Gérard-Marchant, P. (2020). Array Programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Jensen, K. (2013). Process diagram showing the relationship between the different phases of CRISP-DM [Online Image]. In *Wikipedia*. https://upload.wikimedia.org/wikipedia/commons/thumb/b/b9/CRISP-DM_Process_Diagram.png/1024px-CRISP-DM_Process_Diagram.png

Krishnaswamy, V., Singh, N., Sharma, M., Verma, N., & Verma, A. (2022). Application of CRISP-DM methodology for managing human-wildlife conflicts: an empirical case study in India. *Journal of Environmental Planning and Management*, *66*(11), 2247–2273. https://doi.org/10.1080/09640568.2022.2070460

Narasimha Murthy, K. V., Saravana, R., & Vijaya Kumar, K. (2017). Modeling and forecasting rainfall patterns of southwest monsoons in North–East India as a SARIMA process. *Meteorology and Atmospheric Physics*, *130*(1), 99–106. https://doi.org/10.1007/s00703-017-0504-2

National Institute of Standards and Technology. (2012). *What are outliers in the data?* NIST/SEMATECH E-Handbook of Statistical Methods. https://doi.org/10.18434/M32189

NWO. (2018). *Netherlands Code of Conduct for Research Integrity*. https://www.nwo.nl/sites/nwo/files/documents/Netherlands%2BCode%2Bof%2BConduct%2Bfor%2BResearch%2BIntegrity_2018_UK.pdf

Pal, A., & Prakash, P. K. S. (2017). *Practical time series analysis* (1st ed.). Packt Publishing. http://app.knovel.com/hotlink/toc/id:kpPTSA0003/practical-time-series?kpromoter=marc

Patience, G. (2018). *CRISP-DM Analysis Template*. GitHub. https://github.com/patiegm/Datasci_Resources/blob/master/CRISP-DM%20Analysis%20Template.ipynb

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Rijkswaterstaat. (n.d.-a). *Normaal Amsterdams Peil (NAP)*. Www.rijkswaterstaat.nl; Rijkswaterstaat. Retrieved July 10, 2024, from https://www.rijkswaterstaat.nl/zakelijk/open-data/normaal-amsterdams-peil

Rijkswaterstaat. (n.d.-b). *Online Waterberichtgeving | Rijkswaterstaat*. Waterberichtgeving.rws.nl. Retrieved May 27, 2024, from https://waterberichtgeving.rws.nl/owb/droogtemonitor/rijnenmaas

Rijkswaterstaat. (n.d.-c). *Rijkswaterstaat Waterinfo*. Waterinfo.rws.nl; Rijkswaterstaat. Retrieved July 10, 2024, from https://waterinfo.rws.nl/#/nav/publiek

Rijkswaterstaat. (2024). *GeoWeb Catalogus*. Rijkswaterstaat.nl. https://maps.rijkswaterstaat.nl/GeoWebPortaal/

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*(1), 526–534.

Seabold, Skipper, & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*.

Sheppard, K., Khrapov, S., Lipták, G., Hattem, R. van, mikedeltalima, Hammudoglu, J., Capellini, R., alejandro-cermeno, bot, S., Hugle, esvhd, Fortin, A., JPN, Judell, M., Russell, R., Li, W., 645775992, Adams, A., jbrockmendel, & Migrator, L. (2024). *bashtage/arch: Release 7.0.0*. Zenodo. https://zenodo.org/records/10981635

Sluijter, R., Plieger, M., van Oldenborgh, G. J., Beersma, J., & de Vries, H. (2018). *De droogte van 2018*. https://cdn.knmi.nl/system/readmore_links/files/000/001/101/original/droogterapport.pdf?1543246174

Sony, R. K. (2020). *Time Series Forecasting with ARIMA*. Kaggle.com; Kaggle. https://www.kaggle.com/code/redwankarimsony/time-series-forecasting-with-arima

Teunis, B. (2019). *Droogteseizoen 2018*. https://open.rijkswaterstaat.nl/publish/pages/70930/droogteseizoen_2018_lcw_terugblik.pdf

The pandas development team. (2024). pandas-dev/pandas: Pandas (v2.2.2). *Zenodo*. https://doi.org/10.5281/zenodo.10957263

Tyagi, S., Chandra, S., & Tyagi, G. (2023). Climate Change and its Impact on Sugarcane Production and Future Forecast in India: A Comparison Study of Univariate and Multivariate Time Series Models. *Sugar Tech/Sugar Tech*, *25*(5), 1061–1069. https://doi.org/10.1007/s12355-023-01271-2

University of Twente CES. (2023). *Student Charter*. https://www.utwente.nl/en/ces/sacc/regulations/charter2023.pdf

van der Kuil, E., van Bezu, K., & Kloosterman, A. (2020). *Handelingsperspectieven droogte IJssel en Twentekanalen*. https://logisticsoverijssel.nl/wp-content/uploads/202007-Eindrapportage-Handelingsperspectieven-droogte-IJssel-en-Twentekanalen.pdf

van Putten, D., & Rijkswaterstaat. (2024). *Inquiry regarding the models used by Rijkswaterstaat* [Email to Gorkem Yurttas].

Viccione, G., Guarnaccia, C., Mancini, S., & Quartieri, J. (2019). On the use of ARIMA models for short-term water tank levels forecasting. *Water Supply*, *20*(3). https://doi.org/10.2166/ws.2019.190

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., & Carey, C. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xie, S., Zhang, K., Li, Y., & Dai, C. (2020). Establishment of Low Water Runoff Forecast Model in Yichun River Basin by Multiple Linear Regression Method. *IOP Conference Series. Earth and Environmental Science*, *474*(7), 072067–072067. https://doi.org/10.1088/1755-1315/474/7/072067

Appendices

Appendix A – Systematic Literature Review

Key concepts and Sources

Key concepts of the question were identified as follows:

	KEY CONCEPTS
1	Time Series Models
2	Hydrological Forecasting
3	Stochastic Modelling

Table A.0.1: Key concepts

Time series models are the core of this research; therefore, they have to be included as a key concept in the Systematic Literature Review. Additionally, hydrological forecasting is another primary concept as they focus on estimating water flows by utilising meteorological observations which is also the aim of the thesis.

KEY CONCEPTS		RELATED TERMS	NARROWER TERMS	BROADER TERMS
1	Time Series Models	-	Multivariable linear regression ARMA - Autoregressive moving average ARIMA – Autoregressive integrated moving average SARIMA - Seasonal ARIMA ARIMAX – ARIMA with exogenous variable SARIMAX – SARIMA with exogenous variable PARMA – Periodic ARMA	Data Analysis Statistics Time Series Analysis
2	Hydrological Forecasting	-	Drought forecasting Water flow forecasting Flood forecasting Water level forecasting	Forecasting Geoscience
3	Stochastic Modelling	Stationarity	-	Statistics Stochastic modelling Time Series Data

Table A.0.2: Search terms

Databases selected for searching sources are: MathSciNet, Scopus, and Web of Science. MathSciNet database contains many articles related to mathematical models and statistics such as time series models, making them a valuable source. Scopus and Web of Science are large databases that contain a vast library of articles, thus allowing for a wider range of results from various disciplines and different applications. This wide range of results help understanding the models in more detail.

Date	Source	Search string	# of results	Notes

20/05	Scopus	{time series} AND {river forecast} OR {water level forecast} OR {hydrological forecast} OR {drought forecast} AND NOT {machine learning} AND NOT {deep learning} AND NOT {neural networks} AND PUBYEAR > 2003 AND PUBYEAR < 2025 AND (LIMIT-TO (LANGUAGE , "English"))	312	The first 100 results were investigated ordered on relevance. There were some useful papers, however there were many which did not include any of the methods required.
20/05	Scopus	TITLE-ABS-KEY ("time series") AND TITLE-ABS-KEY (forecast) AND TITLE-ABS-KEY (arimax) OR TITLE-ABS-KEY (sarimax) AND PUBYEAR > 2003 AND PUBYEAR < 2025 AND NOT ALL ("deep learning") AND NOT ALL ("machine learning") AND NOT ALL ("neural network") AND NOT ALL ("neuro fuzzy") AND (LIMIT-TO (LANGUAGE , "English"))	35	This time a search on ARIMAX and SARIMAX was conducted as there were not many results related to this in the previous query results. Additionally, the requirements for being related to hydrological forecast was removed as knowledge regarding the models in general was sought. Also title, abstract and keyword search was conducted instead of all as in the previous query a lot of unrelated results were shown.
20/05	MathSciNet	any{hydrological forecast}	20	Not many papers were found in this query. Not many search queries were included to broaden the research and method terms were not included as well. The main reasoning was that the database being already focused on mathematical sciences.
20/05	Web of Science(WoS)	(((((((ALL=(time series)) AND ALL=(river forecast)) OR ALL=(hydrological forecast)) OR ALL=(drought forecast)) OR ALL=(water level forecast)) NOT ALL=(machine	38	None were relevant as they contained deep learning and machine learning which were exclusion criteria

		learning)) NOT ALL=(deep learning)) NOT ALL=(neural)		
20/05	WoS	((((((((((((AB=(time series)) AND AB=(forecast)) AND AB=(ARMA)) OR AB=(ARIMA)) OR AB=(PARMA)) OR AB=(SARIMA)) OR AB=(moving average)) OR AB=(autoregressive)) OR AB=(linear regression)) NOT AB=(neural network)) NOT AB=(deep learning)) NOT AB=(machine learning)) NOT AB=(fuzzy network)) AND AB=(comparison)	19.588	There were too many results, most likely due to many OR terms being present. Nevertheless, articles were ordered on relevance and the first 100 were inspected due to the sheer number of results.
21/05	WoS	((AB=(time series)) AND AB=(multivariate linear regression)) AND AB=(forecast)	83	This query was done to get more knowledge regarding MLR method. The results were not many but sufficient enough to investigate.
21/05	Backtracking	-	-	References of papers that were relevant were examined

Table A.0.3: Search log

Selection criteria, Sources and

Inclusion Criteria	
Criteria	Reasoning
Must contain time series models	The search must include time-series models specifically to abide by the SMART framework for the reasons mentioned earlier in this paper.
Must contain insights into forecasting environmental factors such as water levels, precipitation or drought or must be a general explanation of one of the forecast models mentioned in Table 7.2	Time series models are used in many branches including business cases to forecast demand, sales etc. However, the topic of this assignment requires a forecast to made into something with little direct human interference such as the water levels and drought which are caused by the environmental factors. Therefore, the paper must include something similar to get an insight on which models might or might not be suitable for the purposes of the assignment. Otherwise, the model must include relevant information on the application of one of the models proposed.
English language	For better understandability and a quality review, the language of the papers that are going to be reviewed must be English.

Table A.0.4: Inclusion Criteria

Exclusion Criteria	
Criteria	Reasoning

Use of any machine learning algorithm	As stated in earlier in this paper, machine learning algorithms are not a part of the scope considered by this assignment. Due to lack of expertly understanding in this area is also another factor why machine learning algorithms cannot be considered in a meaningful way. Therefore, papers that use machine learning algorithms will not be considered.
Articles that are not peer-reviewed	This criterion is put in place in order to maintain higher quality papers that have been approved by experts other than the author.
Articles older than 20 years	The time frame established may come as wider compared to more commonly used 5 year time frame. However, the requirement of time series techniques combined with the second inclusion criteria narrow down the search highly. Therefore, a wider time frame had to be considered in order to yield better results. Additionally, time series techniques are well established methods which allows for older research to be used effectively as well. Finaly, the recent popularisation of machine learning algorithms makes it harder to find papers that include time series models as most of the research is focused on machine learning. Therefore, a wider time frame had to be considered to find relevant results.

Table A.0.5: Exclusion Criteria

Concept matrix

#	Source	Article	Time Series	Hydrological forecasting	Stationarity	Additional notes
1.	Scopus	Establishment of Low Water Runoff Forecast Model in Yichun River Basin by Multiple Linear Regression Method (Xie et al., 2020)	X	X		Case example of runoff depth forecast in a river basin. There is the use of multivariable linear regression, using the exogenous variable total rainfall amount.
2.	Scopus	Water level forecasting and navigability conditions of the Tapajós River - Amazon – Brazil (De Figueiredo &	X	X		Case example and application of a SARIMA model with useful insights into making the model and validating it. Additionally, the forecast provides a link to the economic activities and how the model can be used to

		Cavalcante Blanco, 2016)				estimate costs, which may prove useful in this project as well.
3.	Scopus	Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting (Banas & Utnik-Banas, 2021)	X		X	Insights into development of ARIMA, SARIMA and SARIMAX models. Important information about choosing exogenous variables for SARIMAX
4.	Scopus	Climate Change and its Impact on Sugarcane Production and Future Forecast in India: A Comparison Study of Univariate and Multivariate Time Series Models (Tyagi et al., 2023)	X		X	Paper discusses application of 6 different models including ARIMA and ARIMAX. It makes comparisons among them using different statistical tests and measures. It is important to understand the measures used to compare the two models. However, information regarding the other 4 methods is not relevant for this research.
5.	Scopus	Forecasting emergency department hourly occupancy using time series analysis (Cheng et al., 2021)	X			Paper uses a SARIMAX model to forecast occupancy in an emergency department. The paper uses more than one exogenous variable, which makes it interesting. There are also insights into validation statistics and tests.

6.	Math SciNet	Forecasting with prediction intervals for periodic autoregressive moving average models (Anderson et al., 2012)	X	X		Detailed explanation of application of the PARMA model. PARMA model is a less common method compared to more common techniques such as ARMA, ARIMA or SARIMA. This paper is one of the few examples I found. It also has an example case where the model is applied on river flow.
7.	WoS	Study and analysis of SARIMA and LSTM in forecasting time series data (Dubey et al., 2021)	X			In detail explanation of ARIMA model and its application. SARIMA explanation is brief compared to the ARIMA as it is explained through ARIMA, but it includes an application of a SARIMA model.
8.	WoS	On the use of ARIMA models for short-term water tank levels forecasting (Viccione et al., 2019)	X	X		In detail explanation of the use of ARIMA model for short period (1 day and 3 days), explanation of validating linearity. Also, insights into methods for validating the model.
9.	WoS	Modelling and forecasting rainfall patterns of southwest monsoons in North–East India as a SARIMA process (Narasimha Murty et al., 2017)	X	X	X	Application of SARIMA for forecasting rainfall. Detailed information in regard to model and parameters is present. Additionally, some information regarding properties that indicate non-stationarity are present, and how to handle it by differencing.
10.	WoS	Predicting Total Sugar Production	X			It is one of the few papers found related to MLR and its

		Using Multivariable Linear Regression (Gupta & Agarwal, 2021)				application. Unfortunately, it lacks detailed application process and testing methods.
11.	Backtracking	Time Series Analysis: Forecasting and Control (Box et al., 2015)	X		X	This book is one of the landmark literatures for time series analysis and is referenced a lot in many articles. It defines many of the models investigated. It also touches upon the concepts of stationarity, describing it.
12.	Backtracking	A Study on Modelling Daily Mean Flow with MLR, ARIMA and RBFNN (Birinci & Akay, 2010)	X	X	X	A study of various methods such as MLR and ARIMA and their application on streamflow forecast

Figure A.6: Concept matrix

Appendix B – Interview Summary Company 1

Interview Summary

Interviewee

- **Interviewee**: Employee of Company 1, board of the Twente Canal in X.
- **Date:** 05 Jun,2024

Company Background

- The company is located by the Twente Canal in X.
- **Raw materials** are primarily sourced from Asia, Turkey, and locally in Europe, transported by large vessels to ports in the Netherlands and Belgium, and then transported to the company by barge, rail, and truck.

Operational Challenges

1.Water Level Issues

- Predictability of water levels is crucial for the company's transportation, especially during summer when levels are low.

- Low water levels increase transportation costs as load capacity decreases, raising the cost per ton.

2.Logistics

- The company receives raw materials by barge, but relies more on **truck and rail** transportation when water levels are low, increasing operational difficulty and costs.
- Rail transportation is sometimes unreliable, causing disruptions in operational plans.

3.Cost Control

- The company needs to balance the cost of imported materials with the high prices of local European purchases while facing import restrictions and tariff adjustments.
- Market competition and rising costs in 2022 and 2023 have impacted the company's profitability.

4.Inventory Management

- The company needs to **ensure sufficient inventory** to meet market demand, but excessive inventory increases storage costs and cash flow pressure.
- Inventory levels need to be flexibly adjusted based on water levels and market conditions.

Solutions and Recommendations

1. Predicting Water Level Changes

- **Importance**: Predictability of water levels is crucial for transportation, especially during summer when levels are low.
- **Method**: Use historical data and weather forecasts to **simulate water level changes** over the coming weeks through a digital twin platform. This can help the company plan transportation routes and storage strategies in advance.

2. Optimizing Transportation Methods

- **Diversified Transportation**: Flexibly use various transportation methods such as barge, rail, and truck. Choose the **optimal transportation method** based on water levels and road conditions to reduce costs and improve efficiency.
- **Real-time Monitoring**: Establish a real-time monitoring system to track the status of each shipment and adjust transportation plans promptly.

3. Data Sharing and Collaboration

- **Data Sharing with Partners**: Share real-time data with logistics companies, ports, and other supply chain partners to ensure **transparency and efficient** operation of the entire supply chain.
- **Establish Collaboration Mechanisms**: Establish collaboration mechanisms with relevant stakeholders (such as port authorities, logistics companies, etc.) to communicate and resolve transportation issues in a timely manner.

Next Steps

- Plan to contact relevant logistics department to gather more information on waiting times and transportation costs.
- Continue communication with the project team to ensure that the development and application of digital twin technology meet the company's needs.

All names, direct and indirect identifiers that could be linked to persons and companies were redacted from the interview. The redacted data have been replaced with the letter X. Additionally, parts which were both irrelevant for the purposes of this research and contained extensive information that could be used for identification were removed.

Appendix C – Interview Summary Company 2

Interview Summary

Interviewee

- **Interviewee**: Operations Manager of Company 2, focus on Supply Chain Optimization, Trucking, Shipping, Logistics, International Logistics and more.
- **Date:** 20 Jun,2024

Company Background

Company 2 is a **logistics and shipping company** based in the Netherlands, specializing in inland barge transportation and terminal operations. Company 2 operates multiple terminals in the Netherlands and abroad, with a total of X terminals. They transport containers by barge to and from X.

1. Company 2's business scope includes:

- **Barge Transportation**: Company 2 connects major ports and inland terminals in the Netherlands through inland barge transportation, ensuring efficient cargo movement.
- **Terminal Operations**: Company 2 manages multiple terminals, providing loading and unloading, warehousing, and logistics services.
- **Multimodal Transport**: Combining barge and road transport, Company 2 offers integrated logistics solutions.

2. X Terminal

- **Key Operations Hub**: The X terminal is a crucial operational hub for Company 2, located in eastern Netherlands, serving as a vital link between the Netherlands and Germany.
- **Facilities and Services**: X terminal is equipped with advanced loading and unloading equipment, handling primarily containers and bulk cargo, and providing warehousing and value-added logistics services.
- **Transport Connections**: The terminal is connected to major sea ports such as Rotterdam by inland barges and is linked to important logistics networks in the Netherlands and neighboring countries via road connections.

Operational Challenges

1. Water Level Management

The water level fluctuations of the Ijssle River are the main constraint on navigation in the Twente region.

a. **Low Water Levels:**

- In extreme cases, the government mandates that the Ijssle Canal allows only one-way traffic, affecting the navigation plans of vessels heading to the Twente region.
- Low water levels require vessels to adjust their cargo loads, leading to increased transit time and effort.

b. High Water Levels:

- High water levels may restrict passage under bridges, especially for multi-layer container ships.

2. Bridge and Lock Management

a. Bridge Passage:

- When water levels are too high, vessels may be unable to pass under certain bridges, necessitating route replanning or waiting for water levels to change.
- Some bridges, such as railway bridges, are difficult to open, adding to the complexity of navigation.

b. Lock Usage:

- The operation of locks is significantly impacted by both high and low water levels. In low water conditions, lock operation frequency may be limited, increasing wait times for vessels entering the Twente Canal from the Ijssle.
- Effective emergency management and communication are needed for lock malfunctions or maintenance periods. There is such a channel, this is the mailing service from “De Waterkamer” But in the past we were more directly informed from the people of Rijkswaterstaat in the field. Due to reasons there is not anymore a person who set this as his job.

3. Vessel Scheduling and Queue Management

a. Queue Management:

- Queue management for vessels at locks needs optimization to avoid long waiting times.
- Priority queueing strategies have been discussed to improve overall passage efficiency. X is working on a waiting model for lock passage.

b. Scheduling Optimization:

- Scheduling needs to consider various factors such as different vessel loads, water level changes, and other operational factors.
- Optimized scheduling can reduce fuel consumption and operational costs, improving efficiency.
- By understanding the specific lock passage times in advance, vessel speeds can be optimized, affecting fuel consumption rates and indirectly impacting the company's operating costs. There is a rule that the ship which arrives first at the lock will be the

first to pass the lock. It is therefore that it is not so easy to tell skippers to slow down in speed, because if another ship passes him, he will be first in line to pass the lock.

Solutions and Recommendations

a. Real-Time Monitoring and Prediction

- **Water Level Management**: Continuous monitoring of water levels and using historical and real-time data to predict changes allows for proactive adjustments to navigation routes and schedules, mitigating the impact of both low and high water levels.
- **Bridge and Lock Status**: Integrating real-time data on the status of bridges and locks provides instant alerts and rapid response to issues.

b. Information Share System

- Developing a centralized platform where all stakeholders (including vessel operators, port authorities, and maintenance teams) can access real-time data and updates.
- **Queue Management**: Providing real-time information on queue lengths and waiting times at locks and bridges, allowing vessel operators to adjust speeds and schedules accordingly.
- **Real-Time Updates**: Offering updates on lock and bridge maintenance status, completion times, and changes from two-way to one-way navigation.

c. Vessel Scheduling Management

- **Speed Optimization**: Providing recommendations for optimal sailing speeds to reduce fuel consumption and adjust for expected wait times at locks and bridges.

All names, direct and indirect identifiers that could be linked to persons and companies were redacted from the interview. The redacted data have been replaced with the letter X. Additionally, parts which were both irrelevant for the purposes of this research and contained extensive information that could be used for identification were removed.

Appendix D – Interview Summary Company 3

Interview Summary

Interviewee

- **Interviewee**: Project manager at Company 3, focuses on recycling complex materials, such as e-waste, construction waste, and other metals.
- **Date:** 13Jun,2024

Company Background

Company 3 is a leading company in the recycling industry, becoming part of the X Group in X. The company specializes in processing and recycling valuable materials from complex waste streams, such as electronic waste ,construction waste, and other metals. Company 3 is committed to sustainability, environmental compliance, and efficient supply chain management.

3. Material Recycling:

- Company 3 recycles a wide range of materials, including ferrous and non-ferrous metals, plastics, and complex electronic waste. These materials are sourced from construction sites, public sectors like the X, and other industrial clients.
- The company employs advanced technologies such as infrared, X-ray, and magnetic separation to effectively sort and process materials.

4. Supply Chain and Logistics:

- Company 3 strategically utilizes both sea and inland shipping to transport materials, optimizing logistics efficiency and reducing costs.
- The company has a robust logistics network near major transportation hubs, such as the German border, ensuring efficient material flow and export capabilities.
- The recycled materials are primarily sold to markets in Germany and within the Netherlands.

5. Sustainability Initiatives:

- Company 3 is dedicated to reducing CO2 emissions and adheres to strict environmental regulations and certifications, including the CO2 performance ladder certification in the Netherlands.

Operational Challenges

4. Water Level Fluctuations

- Inland water level fluctuations can affect the navigability of vessels, especially during dry seasons when low water levels make it difficult for large vessels to pass. This results in transportation delays and increased costs, necessitating the use of more trucks to replace waterway transport.

5. Weather Conditions

- Adverse weather conditions can impact transportation and dock unloading operations, including strong winds, heavy rain, and extreme temperatures, leading to instability during loading and unloading, causing vessel rocking.

6. Market Price Fluctuations

- The market prices of recyclable materials fluctuate significantly, requiring sales at the optimal time to maximize profits. Flexible inventory management and market monitoring systems are needed to sell promptly when prices are favorable. High demands on transportation and delivery timing mean any transportation tool failure or partner mishap can lead to delays and additional costs.

7. Vessel Monitoring and Scheduling

- In order to calculate and control carbon emissions during transportation, it is necessary to understand the engine specifications/emissions from the overview of available ships
- Vessel transport often encounters situations where return trips are empty. To save costs, companies will contact others to see if they need to use the vessel. Currently, communication is primarily done via phone inquiries, leading to inefficient information flow between companies.

- Recreational activities on the canal, such as rowing and other water sports, pose a risk to navigation safety.

Solutions and Recommendations

1. Real-Time Monitoring and Prediction

- **Water Level Prediction**: Use a digital twin model to monitor canal water levels in real-time, predict future changes, and optimize vessel scheduling and navigation plans.
- **Weather Monitoring**: Integrate real-time weather monitoring systems to provide early warnings of adverse weather conditions, allowing adjustments to loading/unloading operations and transportation plans.

2. Vessel Scheduling Management

- **Multimodal Transport**: Flexibly use various transportation methods such as barge, rail, and truck. Choose the optimal transportation method based on water levels and road conditions to reduce costs and improve efficiency.
- **Communication Platform**: Establish a digital communication platform to share real-time information on vessel availability based on their location and capacity, optimizing the utilization of return trips, enhancing information flow and efficiency between companies.
- **Navigation Safety Monitoring**: share the information of recreational activities on the canal, such as rowing and other water sports, to ensure navigation safety.

All names, direct and indirect identifiers that could be linked to persons and companies were redacted from the interview. The redacted data have been replaced with the letter X. Additionally, parts which were both irrelevant for the purposes of this research and contained extensive information that could be used for identification were removed.

Appendix E – Column list

#	List of columns in the dataset
1	**MONSTER_IDENTIFICATIE**
2	**MEETPUNT_IDENTIFICATIE**
3	**LOCATIE_CODE**
4	**TYPERING_OMSCHRIJVING**
5	**TYPERING_CODE**
6	**GROOTHEID_OMSCHRIJVING**
7	**GROOTHEID_ CODE**
8	**PARAMETER_OMSCHRIJVING**
9	**PARAMETER_ CODE**
10	**CAS_NR**
11	**EENHEID_CODE**
12	**HOEDANIGHEID_OMSCHRIJVING**
13	**HOEDANIGHEID_CODE**
14	**COMPARTIMENT_OMSCHRIJVING**
15	**COMPARTIMENT_CODE**
16	**WAARDEBEWERKINGSMETHODE_OMSCHRIJVING**
17	**WAARDEBEWERKINGSMETHODE_CODE**
18	**WAARDEBEPALINGSMETHODE_OMSCHRIJVING**
19	**WAARDEBEPALINGSMETHODE_CODE**

20	**BEMONSTERINGSSOORT_OMSCHRIJVING**
21	**BEMONSTERINGSSOORT_CODE**
22	**WAARNEMINGDATUM**
23	**WAARNEMINGTIJD (MET/CET)**
24	**LIMIETSYMBOOL**
25	**NUMERIEKEWAARDE**
26	**ALFANUMERIEKEWAARDE**
27	**KWALITEITSOORDEEL_CODE**
28	**REFERENTIE**
29	**NOTITIE_CODE**
30	**NOTITIE_OMSCHRIJVING**
31	**STATUSWAARDE**
32	**OPDRACHTGEVENDE_INSTANTIE**
33	**MEETAPPARAAT_OMSCHRIJVING**
34	**MEETAPPARAAT_CODE**
35	**BEMONSTERINGSAPPARAAT_OMSCHRIJVING**
36	**BEMONSTERINGSAPPARAAT_CODE**
37	**PLAATSBEPALINGSAPPARAAT_OMSCHRIJVING**
38	**PLAATSBEPALINGSAPPARAAT_CODE**
39	**BEMONSTERINGSHOOGTE**
40	**REFERENTIEVLAK**
41	**EPSG**
42	**X**
43	**Y**
44	**ORGAAN_OMSCHRIJVING**
45	**ORGAAN_CODE**
46	**TAXON_NAME**
47	**GROEPERING_OMSCHRIJVING**
48	**GROEPERING_CODE**
49	**GROEPERING_KANAAL**
50	**GROEPERING_TYPE**

Appendix F – Python code

It is important to note that most of the code that is referenced was adjusted for specific case. Additionally, ChatGPT was used to debug when required. This is explained more in Appendix X.

Data understanding (Patience, 2018)

#code by Patience (2018)

#Import Libraries Required

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

from scipy.stats import norm, probplot

```
#Data source: Rijkswaterstaat

#Source Query location:

#path =  'C:/Users/gorke/Desktop/University/thesis(non-python)/uncleaned data/uncleaned csv file(surface water height compared to NAP).csv'

#this first path was for the initial data which included different measurement intervals and many columns that were not useful for the project

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/precleaning/converting all to daily.csv'

#this second path is the one where all the data were converted to daily average values, and the unnecessary columns were removed

path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/precleaning/daily average without outliers(1969-2024)(unfilled).csv'

#this third path is where the outliers are removed

# reads the data from the file - denotes as CSV, it has no header, sets column headers

#df =  pd.read_csv(path, sep=',')

df =  pd.read_csv(path, sep=';', encoding= 'utf-sig-8', parse_dates=True, index_col=0)

df.columns

df.shape

df.dtypes

df.describe()

df.info()

df.head(5)

#Missing data

df.isnull().sum()

def missing_values_table(df):
        mis_val = df.isnull().sum()
        mis_val_percent = 100 * df.isnull().sum() / len(df)
        mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)
        mis_val_table_ren_columns = mis_val_table.rename(
        columns = {0 : 'Missing Values', 1 : '% of Total Values'})
        mis_val_table_ren_columns = mis_val_table_ren_columns[
            mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
        '% of Total Values', ascending=False).round(1)
```

```
        print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
            "There are " + str(mis_val_table_ren_columns.shape[0]) +
              " columns that have missing values.")
        return mis_val_table_ren_columns
missing_values_table(df)
# Get the columns with > 50% missing
missing_df = missing_values_table(df);
missing_columns = list(missing_df[missing_df['% of Total Values'] > 50].index)
print('We will remove %d columns.' % len(missing_columns))
# Drop the columns
df = df.drop(list(missing_columns))
#Distributions
def count_values_table(df):
        count_val = df.value_counts()
        count_val_percent = 100 * df.value_counts() / len(df)
        count_val_table = pd.concat([count_val, count_val_percent.round(1)], axis=1)
        count_val_table_ren_columns = count_val_table.rename(
        columns = {0 : 'Count Values', 1 : '% of Total Values'})
        return count_val_table_ren_columns
# Histogram
def hist_chart(df, col):
        plt.style.use('fivethirtyeight')
        plt.hist(df[col].dropna(), edgecolor = 'k', log=False, bins='fd');
        plt.xlabel(col); plt.ylabel('Number of Entries');
        plt.title('Distribution of '+ col);
col = 'Daily_Average'
#qq
def qq_plot(df, col):
    plt.style.use('fivethirtyeight')
    data = df[col].dropna()

    # Q-Q plot
```

```
    probplot(data, dist="norm", plot=plt)
    plt.title('Q-Q Plot of ' + col)
    plt.ylabel('Ordered Values')
    plt.xlabel('Theoretical Quantiles')
    plt.show()
# Histogram & Results
qq_plot(df, col)
hist_chart(df, col)
count_values_table(df.Daily_Average)
plt.show()
```

Differencing

```
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.graphics.tsaplots import plot_pacf
import pandas as pd
import matplotlib.pyplot as plt


#read csv files
#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1969-2024 daily average water level.csv'
path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1997-2024 daily average water level mmarith.csv'


df =  pd.read_csv(path, sep=';', encoding= 'utf-8-sig', parse_dates=True, index_col=0)


diffed1 = df.Daily_Average.diff(2)


diffed1.plot()
plot_acf(diffed1.dropna())
plot_pacf(diffed1.dropna())
#df.Daily_Average.diff(1).plot()
plt.show()
```

Run sequence plots

```
import pandas as pd
```

```
from matplotlib import pyplot

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/precleaning/daily average without outliers(1969-2024)(unfilled).csv'

path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/cleaned/monthly avg.csv'


df =  pd.read_csv(path, sep=';', encoding= 'windows-1254', parse_dates=True, index_col=0)

df.plot()

pyplot.show()
```

Stationarity tests (Sheppard et al., 2024)

```
#code from arch library website itself (Sheppard et al., 2024)


import pandas as pd

from arch.unitroot import PhillipsPerron

from arch.unitroot import ADF

from arch.unitroot import KPSS


#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/cleaned/cleaned 1964-2024 unfilled.csv'

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1969-2024 daily average water level.csv'

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1997-2024 daily average water level mmarith.csv'

path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/2009-2010.csv'


df =  pd.read_csv(path, sep=';', encoding= 'utf-8-sig', parse_dates=True, index_col=0)


df = df.dropna()

pp = PhillipsPerron(df.Daily_Average)

adf = ADF(df.Daily_Average)

kpss = KPSS(df.Daily_Average)


#pp = PhillipsPerron(df.Daily_Average.diff().dropna())

#adf = ADF(df.Daily_Average.diff().dropna())
```

```
#kpss = KPSS(df.Daily_Average.diff().dropna())


print(pp.summary().as_text())

print(adf.summary().as_text())

print(kpss.summary().as_text())
```

Static forecast (Sony, 2020)

```
#code is from (Sony, 2020)


from statsmodels.tsa.statespace.sarimax import SARIMAX

import pandas as pd

import matplotlib.pyplot  as plt

import timeit

from math import sqrt

from math import isnan

start = timeit.default_timer()


path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1969-2024 daily average water level.csv' #all observations

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1997-2024 daily average water level mmarith.csv' #excluding visual readings


df =  pd.read_csv(path, sep=';', encoding= 'utf-8-sig', parse_dates= True, index_col=0, header =0, dayfirst=True)


size = int(len(df) - 100)  #adjust this value as desired, which will give the last x observations to be predicted


train, test = df[1:size], df[size:len(df)]


# Fit the model once on the training data

model = SARIMAX(train, order=(6,1,7), missing='drop')

model_fit = model.fit(disp=False)


# Forecast the entire test set at once
```

```
predictions = model_fit.forecast(steps=len(test))

#error metrics
count = 0
errorsum = 0
for i in range(len(test)):
    if isnan(test.iloc[i]) == False:
        errorsum = (test.iloc[i]-predictions.iloc[i]) ** 2 + errorsum
        count= count + 1

#reset and evaluate error metrics
mserror = errorsum/count
rmserror = sqrt(mserror)

count = 0
errorsum = 0

#calculate MAPE
for j in range(len(test)):
    if isnan(test.iloc[i]) == False:
        errorsum = abs((test.iloc[i]-predictions.iloc[i])/test.iloc[i]) + errorsum
        count = count + 1

maperror = 100*errorsum/count

print('Test Mean Squared Error: ' + str(mserror))
print('Test Root Mean Squared Error: ' + str(rmserror))
print('Mean Absolute Percentage Error: ' + str(maperror))

# plot
plt.plot(test, label = 'true water level', marker = '*')
plt.plot(predictions, color='orange', label = 'predicted water level', marker = '*')
```

```
plt.title('ARIMA(6,1,7) all observations')

plt.xlabel('Future Steps')

plt.ylabel('Water Level')

plt.legend()


stop = timeit.default_timer()

plt.show()


print('Runtime: ' + str(stop - start))
```

Rolling forecast ARIMA (Sony, 2020)

```
#code is from (Sony, 2020)


from statsmodels.tsa.statespace.sarimax import SARIMAX

import pandas as pd

import matplotlib.pyplot  as plt

import timeit

from math import sqrt

from math import isnan

start = timeit.default_timer()


#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1969-2024 daily average water level.csv' #all observations

path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1997-2024 daily average water level mmarith.csv' #excluding visual readings


df =  pd.read_csv(path, sep=';', encoding= 'utf-8-sig', parse_dates= True, index_col=0, dayfirst=True)

X = df.Daily_Average


size = int(len(df) - 100)  #last 100 observations excluding visuals


train, test = X[1:size], X[size:len(X)]

history = [x for x in train]
```

```
predictions = []

point=0
for t in range(len(test)):
    model = SARIMAX(history, order=(6,1,2), missing = 'drop')
    model_fit = model.fit(disp=False)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test.iloc[t]
    history.append(obs)
    point = point+1
    print(point)

#error metrics
count = 0
errorsum = 0
for i in range(len(test)):
    if isnan(test.iloc[i]) == False:
        errorsum = (test.iloc[i]-predictions[i]) ** 2 + errorsum
        count= count + 1

#reset and evaluate error metrics
mserror = errorsum/count
rmserror = sqrt(mserror)

count = 0
errorsum = 0

#calculate MAPE
for j in range(len(test)):
    if isnan(test.iloc[i]) == False:
```

```
        errorsum = abs((test.iloc[i]-predictions[i])/test.iloc[i]) + errorsum
        count = count + 1

maperror = errorsum/count

print('Test Mean Squared Error: ' + str(mserror))
print('Test Root Mean Squared Error: ' + str(rmserror))
print('Mean Absolute Percentage Error: ' + str(maperror))

# Correct indices for predictions
test_index = test.index  # Get the index of the test data (last 100 observations)

# plot
plt.plot(test_index, test, label='True Water Level', marker='*')
plt.plot(test_index, predictions, color='red', label='Predicted Water Level', marker='*')
plt.title('ARIMA(6,1,7) all observations')
plt.xlabel('Future Steps')
plt.ylabel('Water Level')
plt.legend()

stop = timeit.default_timer()
plt.show()

print('Runtime: ' + str(stop - start))
```

Forecast rolling ARIMAX (Sony, 2020)

```
#code from Sony (2020) adjusted using ChatGPT
from statsmodels.tsa.statespace.sarimax import SARIMAX
import pandas as pd
import matplotlib.pyplot as plt
import timeit
from math import sqrt
from math import isnan
```

```
# Start the timer
start = timeit.default_timer()


# Load the dataset
path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1990-2024 temperature included.csv'
df = pd.read_csv(path, sep=';', encoding='utf-8-sig', parse_dates=True, index_col=0, dayfirst=True)


# Endogenous variable
X = df['Daily_Average']


# Exogenous variable
exog = df['Temperature']


# Define the size of the training set
size = int(len(df) - 100)  # Last 100 observations excluding visuals


# Split the data into training and testing sets
train_X, test_X = X[1:size], X[size:]
train_exog, test_exog = exog[1:size], exog[size:]


# Initialize history with the training data
history_X = [x for x in train_X]
history_exog = [x for x in train_exog]
predictions = []


# Iterate over the test set to make rolling forecasts
point = 0
for t in range(len(test_X)):
    model = SARIMAX(history_X, order=(6,1,7), exog=history_exog, missing='drop')
    model_fit = model.fit(disp=False)
```

```
    output = model_fit.forecast(exog=test_exog.iloc[t:t+1])  # Forecast with the corresponding exogenous value
    yhat = output[0]
    predictions.append(yhat)
    obs = test_X.iloc[t]
    history_X.append(obs)
    history_exog.append(test_exog.iloc[t])
    point += 1
    print(point)

# Error metrics
mse = ((test_X - predictions) ** 2).mean()
rmse = sqrt(mse)
mape = (abs((test_X - predictions) / test_X).mean()) * 100

print('Test Mean Squared Error: ' + str(mse))
print('Test Root Mean Squared Error: ' + str(rmse))
print('Mean Absolute Percentage Error: ' + str(mape))

# Correct indices for predictions
test_index = test_X.index  # Get the index of the test data (last 100 observations)

# Plot
plt.plot(test_index, test_X, label='True Water Level', marker='*')
plt.plot(test_index, predictions, color='red', label='Predicted Water Level', marker='*')
plt.title('ARIMA(6,1,7) with Exogenous Variable (Temperature)')
plt.xlabel('Future Steps')
plt.ylabel('Water Level')
plt.legend()

# Stop the timer and display the plot
stop = timeit.default_timer()
plt.show()
```

```
print('Runtime: ' + str(stop - start))
```

Information criterion (Brownlee, 2020)

```
#code by (Brownlee, 2020)


import pandas as pd

from statsmodels.tsa.statespace.sarimax import SARIMAX


# load dataset

path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1969-2024 daily average water level.csv' #all observations

#path = 'C:/Users/gorke/Desktop/University/thesis(non-python)/final used documents/1997-2024 daily average water level mmarith.csv' #excluding visual readings


df =  pd.read_csv(path, sep=';', encoding= 'utf-8-sig', parse_dates=True, index_col=0)

# fit model

model = SARIMAX(df.Daily_Average, order=(6,1,7), missing='drop')

model_fit = model.fit()

# summary of fit model

print(model_fit.summary())
```

Appendix G – Long term predictions last 100 observations

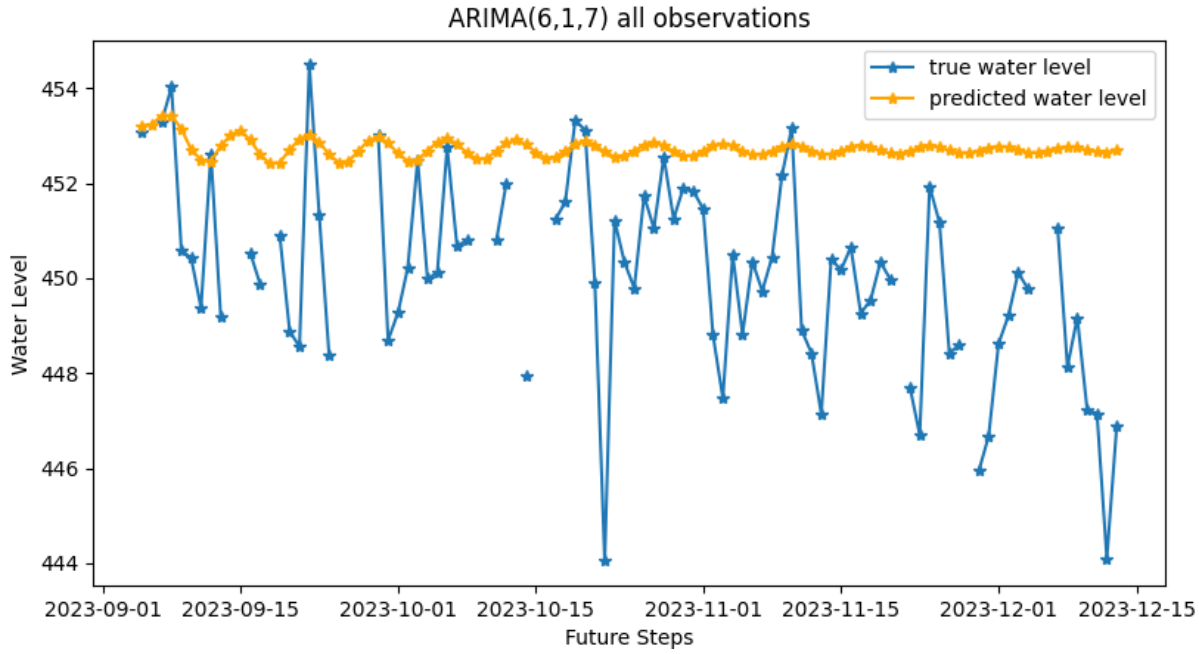

Figure G.1: ARIMA(6,1,7) all observations, last 100 observations

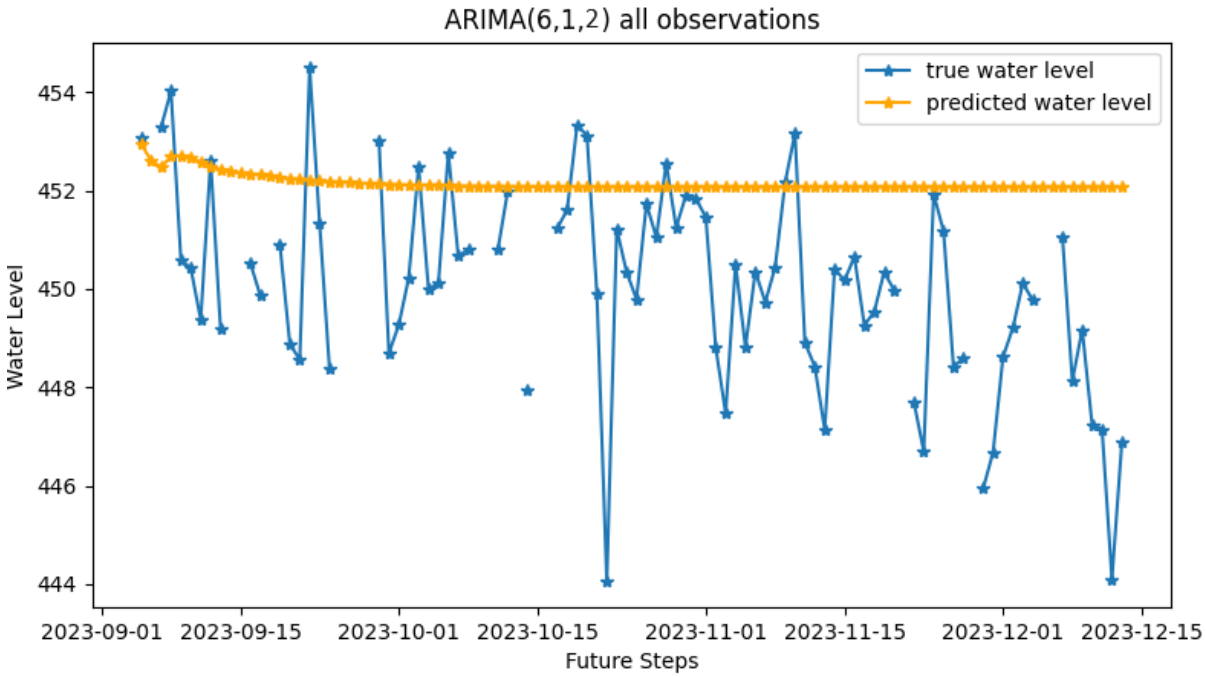

Figure G.2: ARIMA(6,1,2) all observations, last 100 observations

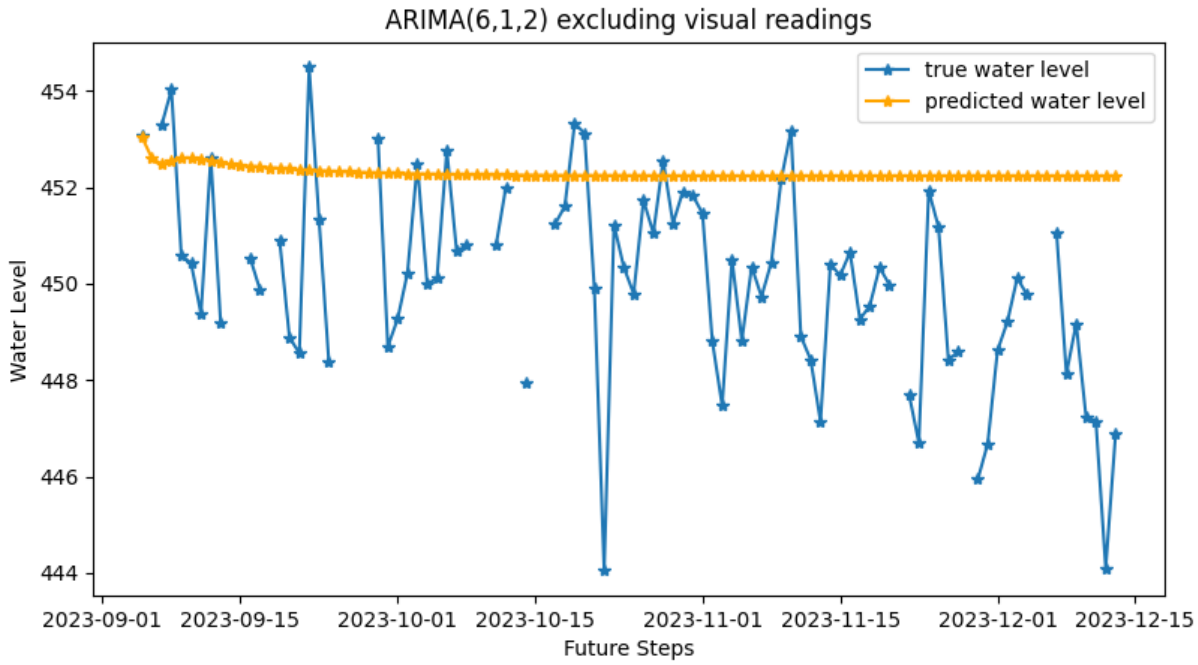

Figure G.3: ARIMA(6,1,2) without visual readings, last 100 observations

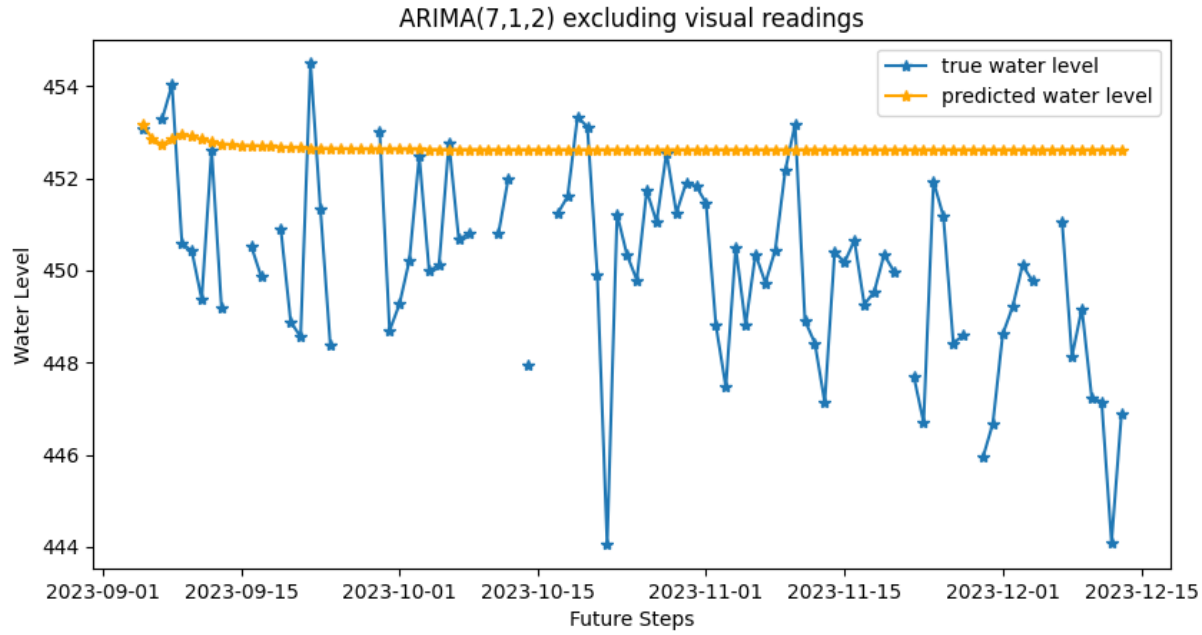

Figure G.4: ARIMA(7,1,2) without visual readings, last 100 observations

Appendix H – Use of AI

Throughout this research, ChatGPT by OpenAI was used extensively for the coding processes. It was used for debugging in almost all of the code files. The proposed solutions by ChatGPT were thoroughly checked before implementation. All responsibility belongs to the author and not on the tool.