# Forecasting effluent flow at Egå wastewater treatment plant

## Data science - exam project

Malte Meinert Olsen

201605945

Primary education:

MSc. in Engineering (Biotechnology and Chemical Engineering), Aarhus University

Elective course in Data Science at:

MSc. Cognitive Science, Aarhus University

Spring 2022

Aarhus University

Denmark

31-05-2022

## Preface

This article is a part of a larger master's thesis study, aiming to develop an alternative control strategy for Egå wastewater treatment plant (WWTP) during operation in weather situations with heavy rainfall. This article investigates data collected by the online sensors at Egå WWTP, in the attempt to develop a forecasting model for effluent conditions of the WWTP.

All data manipulation and analysis were conducted in R using RStudio.

All codes used for this project can be found at:

`https://github.com/Malte-Meinert-Olsen/data-science`

Charter count including space: 35040

Corresponding number standard pages: 14.6

## Abstract

Peak emissions of ammonium from a wastewater treatment plant (WWTP) at events with heavy rain precipitation is harmful to the aquatic ecosystems. Forecasting of such peak emissions are of utmost importance as the WWTP can prepare for the incoming rain event. In this article, data obtained from Egå WWTP has been analysed to develop a model to forecast the effluent flow of the WWTP. The analysis was based on hourly average values of rain precipitation in mm and effluent flow rate in $m^3/h$, obtained by sensors located at the WWTP, and the daily drought index, extracted from DMI's weather achieves. A total of 19 different linear models with and without utilising an autoregressive integrated moving average (ARIMA) model to model the errors were evaluated. The superior model was found to be a linear model without ARIMA errors, including multiple parameters of previous and current rain precipitation; accumulated rain precipitation; drought index; interaction between drought index and accumulated rain; and a dynamic harmonic regression with Fourier terms to model daily, weekly and yearly seasonality. The superior model yielded adjusted $R^2$ of 0.7563 when training the model and a mean absolute error of 163 $m^3/h$ and a mean absolute percentage error of 19.4% when evaluating the trained model on the test set. The results of this article provides the foundation to develop an control strategy, which ultimately can reduce the numbers of ammonium peak at Egå WWTP.

# Contents

# 1 Introduction

## 1.1 Introduction to wastewater treatment

Adequate wastewater treatment and handling are a part of the critical infrastructure of any society. If neglected, severe consequences can arise such as spread of potential lethal illnesses through pathological bacteria or, ultimately, collapse of entire aquatic ecosystems, due to increased algae formation caused by increased amount of nutrient emitted from the wastewater. To lower the ordinary citizen's exposure to wastewater, and thereby potential illnesses, all wastewater in Denmark are collected in an underground sewage system. From the sewage, the wastewater is lead to a wastewater treatment plant (WWTP) where various nutrients are removed from the wastewater before it is lead out to a lake or the sea, called the recipient. If the nutrient removal is insufficient, the nutrients are emitted to the recipient, potentially damaging the aquatic ecosystem. Many nutrients have a toxic effect on the recipient. Ammonium is one of the main chemical compounds of concern, and therefore, this project the will focus on ammonium emission. Other potential nutrients are omitted to reduce complexity. The concentration of ammonium (mg/l) in the emitted water, is a way of quantifying the degree of damage of an emission of ammonium - the higher the concentration, the greater the damage. The water leaving the WWTP is called the effluent, and each WWTP has their own guidelines for how high the ammonium concentration in the effluent is allowed to be. These guidelines are determined based on the nature of the recipient. Therefore, sufficient ammonium removal is a core task for every WWTP. Usually, ammonium is removed in a biological system, consisting of large tanks of concrete (process tanks), utilizing types of bacteria which can remove ammonium in the present of oxygen.

At times with heavy rain precipitation, the rainwater entering the sewage system pushes the present wastewater through the sewage system to the WWTP, leading to increased flow at the WWTP. The increased flow results in a shorter time span for the bacteria in the process tanks to remove ammonium from the wastewater, that in general has a high ammonium concentration. This can potentially cause an increase in the ammonium concentration in the effluent, as the bacteria are unable to remove the usual amount of ammonium in the shortened time span. Contrary, rainwater contains a low concentration of ammonium, and when the initial sewage water has flushed through the plant, the rainwater dilutes the wastewater, consequently decreasing the ammonium concentration in the effluent. Therefore, at events with heavy rain precipitation, a peak in the ammonium concentration in the effluent can be observed. An ammonium peak has consequences for recipient, but can also result in legal consequences for the WWTP, such as injunction.

A way of reducing the peak emission of ammonium is to increase the dissolved oxygen concentration in

the process tanks by increasing the amount of air that is supplied to the process tanks (Metcalf et al., 2014, Chapter 7). The supply of air is a major economic expenditure, thus to much aeration would be expensive. To develop a control strategy which can provide the sufficient air supply in the process tanks will be addressed in the ongoing master's thesis project. It is crucial for the design of this intelligent control system to be able to forecast when a ammonium peak will appear, which is the aim of this article. The time span needed to lower the ammonium concentration in the process tanks before an incoming rain event (hereafter called the forecasting horizon), is a crucial parameter. The forecasting horizon is equal to the time that the WWTP have to prepare for a rain event.
The data analysed in this project is obtained from Egå WWTP, which is located in the northern part of Aarhus and operated by Aarhus Vand.

### 1.2 Flow to the plant

An ammonium peak is physically caused by an increasing flow and an understanding of the flow and its sources is necessary. At Egå WWTP the flow into the plant can be split into four separate sources: Domestic wastewater, industrial wastewater, direct entry of rainwater to the sewage system, and groundwater infiltration. The domestic wastewater is the wastewater produced by the citizens of Aarhus. This flow is determined by human activity, and it is assumed that humans produce wastewater in a fixed pattern, therefore the domestic wastewater can be accurately predicted based on time. The industrial wastewater can be omitted as there is no major industrial contribution regrading flow or ammonium at Egå WWTP. The third source is rainwater entering the sewage system directly, through openings in the road etc. This source can be predicted using the total rain precipitation in mm. Groundwater infiltration into the sewage systems happens due to cracks in the sewers. This is dependent on how much water is absorbed in the first few meters of the soil.

### 1.3 Data science in wastewater treatment

Traditionally, modelling of WWTP is based on solving numerous differential equations, given multiple experimentally determined constants (Metcalf et al., 2014, Chapter 7). However, usage of data driven techniques, such as time series analysis and neural networks, has showed promising results in various studies. Kang et al. successfully utilized Long Short-Term Memory neural networks to predict the flow rate of at a WWTP, and further managed to interpolate missing values in their data set (Kang et al., 2020). Zhang et al. tested the performance of both an Autoregressive Integrated Moving Average (ARIMA) model and a neural network with regards to forecasting flow rates into the WWTP. The study found sufficient performance for both models, however the ARIMA model showed the best performance (Zhang

et al., 2019). Several other papers show the usefulness of various data driven models to predict effluent concentration, models such as regression tree and regression forest (Wang et al., 2021 and Granata et al., 2017), different kinds of neural networks (Guo et al., 2015, Cheon et al., 2008, and Hansen et al., 2022) and Autoregressive Moving Average (ARMA) models (Van Dongen et al., 1998 and Ellis et al., 1990).

### 1.4 Autoregressive Intergraded Moving Average

ARIMA is a model that can be used to forecast time series and consists of three parts: autoregression, differencing, and moving average. The autoregressive part of an ARIMA model follows equation 1:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{1}$$

It calculates $y_t$ based on a linear combination of the past observed values $y_{t-p}$.
The moving average part of an ARIMA model follows equation 2:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_p \varepsilon_{t-p} \tag{2}$$

It calculates $y_t$ based on a linear combination of the past observed errors $\varepsilon_{t-q}$.
The 'I' stand for integrated or integration in the sense of reversed differentiation. The ARIMA model assumes that the time series are stationary, meaning no trend or seasonality. If this criterion is not fulfilled, differencing the time series can make it stationary.
Combining the three parts yields equation 3:

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_p \varepsilon_{t-p} + \varepsilon_t \tag{3}$$

Equation 3 is differenced once as an example. (Hyndman et al., 2021, Chapter 9)

### 1.5 Linear regression

When modelling time series, linear regression or multiple linear regression are easy ways of including various parameters which influence the time series. Multiple linear regression follows equation 4:

$$y_t = \beta_0 + \beta_1 * x_{1,t} + \beta_2 * x_{2,t} + ... + \beta_k * x_{k,t} \tag{4}$$

Multiple linear regression is a way of improving a time series model, by providing information of various parameters, which are correlated with $y_t$, but not included in the trend or seasonality. (Hyndman et al., 2021, Chapter 7)

### 1.6 Dynamic harmonic regression

Models such as ARIMA fail to handle long seasonal periods such as weekly and daily seasonality. Therefore, a dynamic harmonic regression (DHR) with Fourier terms can be used to model the seasonality. A DHR uses a linear combination of pairs of sine and cosine functions to estimate the periodic function of the seasonality. Fourier, a French mathematician, showed that a linear combination of sine and cosine terms with the right frequency can estimate any periodic function. Complex seasonality in a time series, e.g. seasonality containing yearly, weekly, and daily variation, can be viewed as a fixed periodic function, and thereby easily estimated with a DHR. A disadvantage is the assumption of fixed seasonality, as many systems in nature change over time. However, in this project this assumption is acceptable, as the seasonality is caused by humans' daily activity, which does not change rapidly with time. (Hyndman et al., 2021, Chapter 10)

### 1.7 Evaluation of the models

All the above-described models will be used throughout this project to forecast the effluent conditions of Egå WWTP.

To compare different models against one another, three different values will be used to evaluate the models. Adjusted $R^2$ describes how well the model fitted the data, but unlike normal $R^2$ it penalizes the value for every parameter added to the model. The adjusted $R^2$ should be maximized (1 corresponding to a fully fitted model). Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two other metrics for model evaluation, which should be minimized. AIC and BIC often select the same model, however BIC favors models with fewer parameters, as each parameter is more heavily penalized than in the AIC. (Hyndman et al., 2021, Chapter 7)

# 2 Methods

## 2.1 Data description

Egå WWTP was subjected to a major revamp, which was completed in the last half of 2016. The plant has afterwards been the object to many optimizations, which are still ongoing. During the first year of operation many major changes were performed. Data from before 2018 is not representative of how the plant is operated today and thus the data for this project is obtained in 2018 and onwards. However, as the plant has been optimized and changed through the entire period, the most recent data has the highest quality.

The data available in this project is measured from January 1st, 2018 to April 19th, 2022 and is obtained from online sensors giving an average value each minute. Data from 3 different sensors was used in this project, measuring flow of the effluent, ammonium concentration in the effluent and rain precipitation respectively. According to internal experiences from Aarhus Vand, the sensors measuring flow and ammonium concentration yield reliable and stable measurements. A note however is, that the flow of the effluent is a calculated value based on the flow into the plant. A possible consequence of this is that the flow reported at time $t$ might not correspond to the actual flow at time $t$, as the time is based on the influent flow time. To address this uncertainty is beyond the scope of this project. However, the assumption of instantaneous increase in the effluent flow is acceptable, as it can be assumed that water cannot accumulate in the plant and the incoming flow would displace the existing water. The third sensor of interest, the rain sensor, is placed at Egå WWTP and will yield highly reliable data of the rain precipitation in the area around the plant.

## 2.2 Data pre-processing and cleaning

Duplicate values based on the time variable was removed, and only the first observation was kept. The data from the rain sensor contained outliers, defined as negative values or values over 5.4 mm rain precipitation pr. minute, which is the Danish national record (Cappelen, 2022), and 52602 missing minute values. Both the outliers and the missing values were replaced with values found in DMI's weather archives for rain precipitation in Aarhus. These data were on an hourly basis and was divided by 60 to get a value for each minute, assuming uniformly distributed rain precipitation within the hour.

To create less noisy data, the data was aggregated into bins of 5, 15, 30 and 60 minutes. Within each time interval, the data points for the flow and ammonium concentration were averaged, and the rain was accumulated. For the 1-hour data, missing values in the effluent flow and the ammonium concentration were estimated with a linear regression, between the two data points in either end of the period.

### 2.3 Predicting ammonium peak emissions with effluent flow

An ammonium peak emission is a physical event, which can be caused by an increase in flow. A graphical visualization of two time series, effluent flow and effluent ammonium concentration, can be seen in Appendix A. By comparing the two time series it becomes apparent that flow is a more stable time series than ammonium concentration. The ammonium time series is more unstable as other parameters than increasing flow can cause an increase in the ammonium concentration in the effluent, e.g. a change in biological conditions leading to a more insufficient ammonium removal or change in control strategy, which has been done numerous times over the four-year period (Metcalf et al., 2014, Chapter 7). As the plant size has not changed doing the measured period, the data quality of the flow time series does not change with time. Due to the causal relationship between flow and ammonium peak emissions, a forecast of the flow is assumed to yield better prediction of ammonium peaks, as the ammonium concentration time series the more unstable and the data quality varies with time.

### 2.4 Modelling of water saturation of soil

Determining the water saturation of the soil is a complex task. In this project eight parameters will be used to estimate the water saturation of soil. The first seven parameters is the accumulated rain precipitation over the last 1 to 7 days. The last parameter is the drought index, extracted from DMI's weather archives, and is a scale of the dryness in nature between 0 and 10. A drought index of 0 means low risk of drought or high water saturation of the soil, and a drought index of 10 means high risk of drought or low water satuation of the soil. The drought index builds on DMI's hydrogeological models, evaporation estimation, and rain precipitation (*Tørkeindeks* 2022).

To summarize, the forecasting task of this project is to predict the effluent flow out of Egå WWTP based on the three flow components: domestic wastewater (assumed to follow a fix pattern), rainwater entering the sewage system directly (assumed to be determined by the current rain precipitation), and groundwater infiltration (assumed to be correlated with accumulated rain and drought index)

### 2.5 Determination of flow response time and hydraulic retention time

Six days (2018-08-28, 2018-09-07, 2019-08-27, 2020-06-19, 2020-08-18, and 2021-06-20) were chosen for analysis as these represent days with a significant rain event after a long period without rain. The six days were subjected to a graphical investigation (Appendix B) to determine the amount of time lag from when rain was observed to an increase in flow could be observed, henceforth called the flow response time. Based on the graphs in Appendix B the flow response time was manually determined.

The hydraulic retention time (HRT) specifies the time it takes for the water to pass through part of the plant. Throughout this article, HRT refers to the time it takes for the water to pass through the process tanks. Aarhus Vand defines days with less than 2 mm rain precipitation as dry weather operation. The mean and the 25% quantile of the flow into the process tanks for dry weather operation was used to determine the HRT by dividing the process tank volume with the flow.

### 2.6 Determination of data aggregation and forecasting horizon

Before the forecasting models could be trained, the appropriate data type needed to be determined. The raw data values were based on a data point for each minute, which yielded a noisy time series. To reduce this noise, the data was aggregated into different time bins. Graphically, it can be seen that the numbers of outliers reduces as the time interval becomes broader (Appendix C).

Based on the HRT, the forecasting horizon of this project was determined to be 24 hours. The data aggregated into 1-hour intervals was chosen and will be used through the rest of this project.

### 2.7 Modeling

Before training the models, the data was divided into a training set (data from 2018 to 2021 included) and a test set (the 3.5 months of 2022). Afterwards various linear models, with and without ARIMA errors, were fitted to the training data. All the models tested can be seen in Table 4. The optimal number of parameters in the ARIMA models was found automatically by the fable 0.3.1 package in R (O'Hara-Wild et al., 2021a), using a variation of the Hyndman-Khandakar algorithm (Hyndman et al., 2021, Chapther 9.7). To reduce run time, the degree of differencing was defined beforehand to be 0 or 1, and seasonal differencing was excluded.

### 2.8 Validating the flow models

Two models were created and validated on the test set: A linear model including all parameters, henceforth called full linear model, and a linear model including all parameters and ARIMA modelling of the errors, henceforth called full ARIMA model. The mean absolute error (MAE) and the mean absolute pertentage error (MAPE) were calculated for both models by using the fabletools 0.3.2 packages (O'Hara-Wild et al., 2021b). Furthermore, the residual of both models was visualised with the gg_tsresiduals function from the feasts 0.2.2 package (O'Hara-Wild et al., 2021c).

# 3 Results

## 3.1 Flow response time

The flow response time is the time from when the rain begins to the flow begins to increase. The results from the graphical determination of the flow response times can be seen in Table 1:

*Table 1: shows the assigned time point for when the rain starts and when the flow at the wastewater treatment plant starts to increase. The difference between this two values are the flow response time, and is a measure of the time it take before the rainwater disturb the flow.*

| Date | Time rain begins | Time flow begins to increase | Flow response time [min] |
|---|---|---|---|
| 2018-08-28 | 20:20 | 20:50 | 30 |
| 2018-09-07 | 10:30 | 11:10 | 40 |
| 2019-08-27 | 17:10 | 17:45 | 35 |
| 2020-06-19 | 11:15 | 12:00 | 45 |
| 2020-08-18 | 12:10 | 13:10 | 60 |
| 2021-06-20 | 10:30 | 11:10 | 40 |

Based on the time in the table the flow response time was determined to be between 30 and 60 minutes.

## 3.2 Hydraulic retention time

The mean and 25% quantile flow in dry weather operation (less than 2 mm of rain in one day), and the total process tank volume of 22750 m3 were used to calculate the HRT. The results can be seen in Table 2.

*Table 2: shows the flow and hydraulic retention time for days of dry weather operation based on the mean and the 25% quantile*

| | FLow [$m^3$/h] | Hydraulic retention time [hours] |
|---|---|---|
| 25% quantile | 904 | 24 |
| Mean | 1211 | 19 |

## 3.3 Seasonality

To investigate the seasonality, the flow in $m^3$/h was plotted on an daily (see Figure 1), weekly (see Figure 2), and yearly basis (see Figure 3):

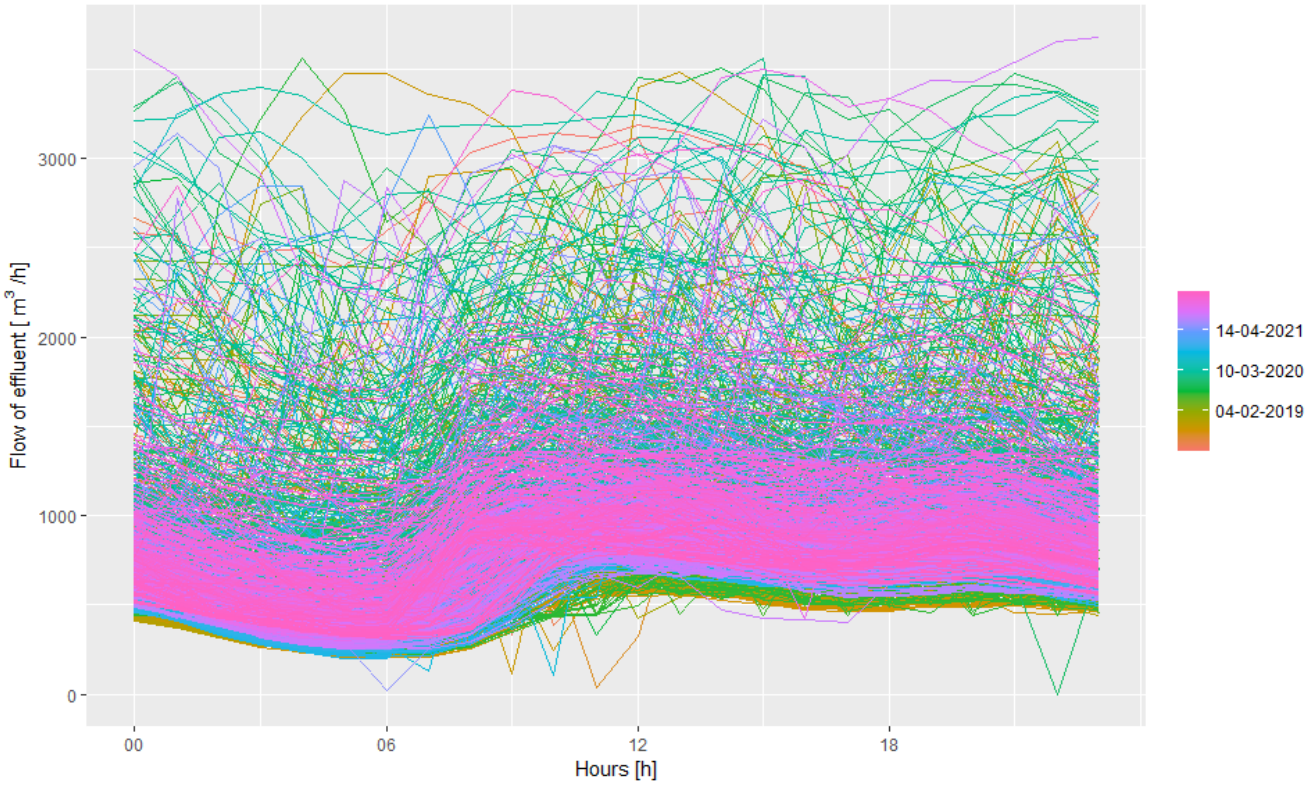


*Figure 1: The seasonal plot of the daily variation.*

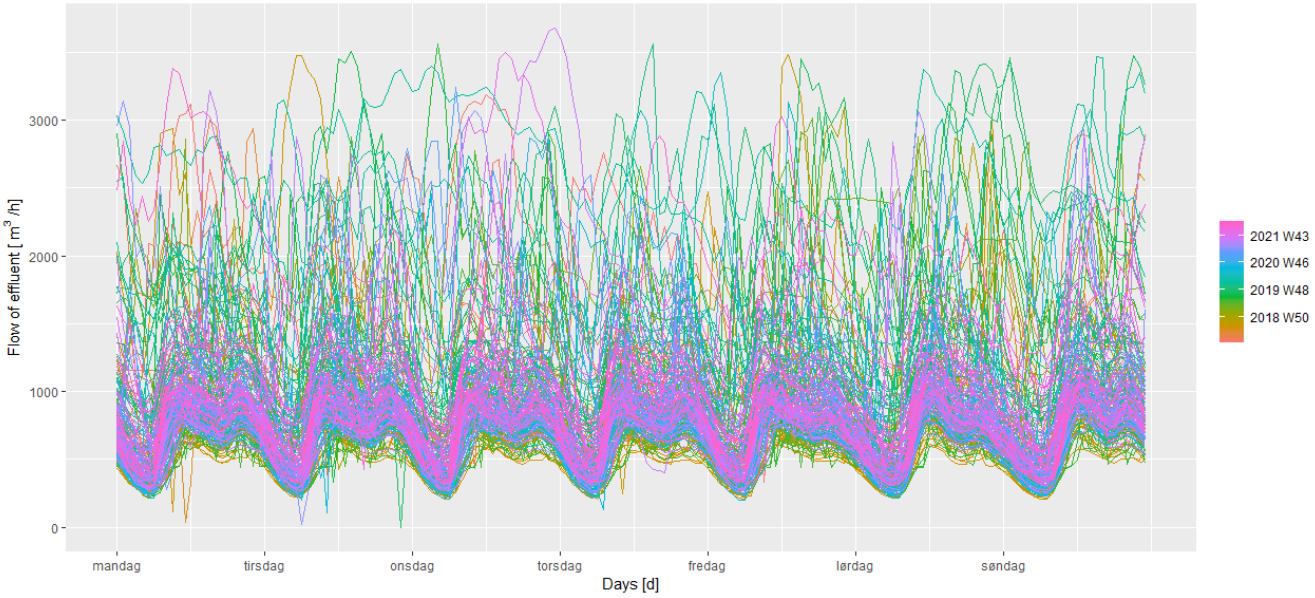


*Figure 2: The seasonal plot of the weekly variation.*

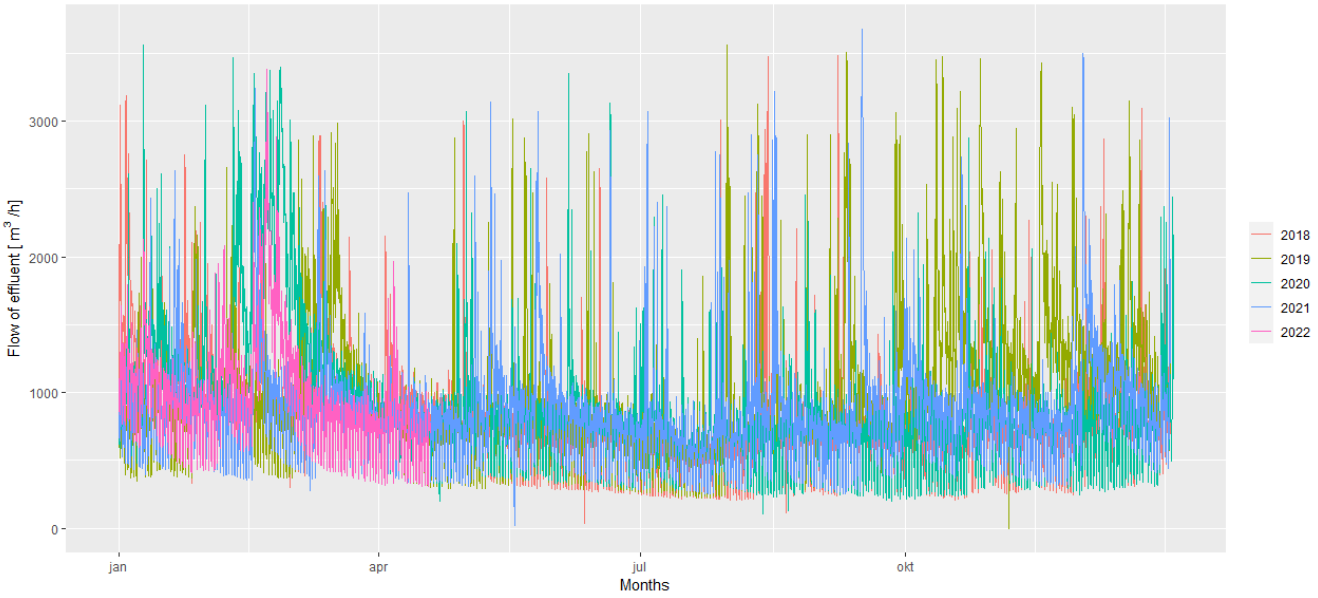


*Figure 3: The seasonal plot of the yearly variation.*

### 3.4 Linear modelling of effluent flow

Table 3 shows the used abbreviations when describing the different models.

*Table 3: shows the abbreviations used when describing the tested models*

| **Abbreviation** | **Explanation** |
|---|---|
| R | Rainfall in mm at the given $t$ |
| R1-R7 | Accumulated rain over the last 1 to 7 days |
| DI | The Drought index for the current day |
| DIxR1 | The interaction of one day accumulated rain and drought index |
| F | Fourier terms |
| D,W,Y | Day, week or year |
| K | Pairs of sine and cosine term |
| lag# | Takes the last, second or third last number of the given parameter, e.g. lag1R is the last observation of rain, corresponding to rain at $t-1$ |
| (p,d,q) | p number of parameters in the autoregression part, d degree of differencing, q numbers of parameters in the moving average part of the ARIMA model |
| TSLM | Linear model |
| ARIMA | Linear model with ARIMA errors |

Table 4 shows the 19 different linear models tested in this project.

*Table 4: shows the different linear models and their adjusted $R^2$, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and the numbers of parameters.*

| **Model number** | **Model name** | **Adjusted $R^2$** | **AIC** | **BIC** | **Numbers of parameters** |
|---|---|---|---|---|---|
| 1 | TSLM(R) | 0.0834 | 425898 | 425923 | 1 |
| 2 | TSLM(R-R1-R2-R3-R4-R5-R6-R7-DI) | 0.445 | 408248 | 408341 | 9 |
| 3 | TSLM(R-R1-R4-R7-DI) | 0.4444 | 408278 | 408337 | 5 |
| 4 | TSLM(R-R1-R7-DI) | 0.4435 | 408338 | 408388 | 4 |
| 5 | TSLM(R-R1 -R7-DIxR1) | 0.4253 | 409461 | 409511 | 4 |
| 6 | TSLM(R-R1-R7-DIxR1-DI) | 0.4672 | 406808 | 406868 | 5 |
| 7 | TSLM(R-lag1R-R1-R7-DIxR1-DI) | 0.527 | 402626 | 402694 | 6 |
| 8 | TSLM(R-lag1R-lag2R-R1-R7-DIxR1-DI) | 0.5634 | 399814 | 399890 | 7 |
| 9 | TSLM(R-lag1R-lag2R-lag3R-R1-R7-DIxR1-DI) | 0.579 | 398531 | 398615 | 8 |
| 10 | TSLM(lag1R-lag2R-lag3R-R1-R7-DIxR1-DI) | 0.5782 | 398593 | 398669 | 7 |
| 11 | TSLM(lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK10-FWK5-FYK3) | 0.7531 | 379863 | 380244 | 43 |
| 12 | TSLM(lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK7-FWK5-FYK3) | 0.7531 | 379856 | 380186 | 37 |
| 13 | TSLM(lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK5-FWK3-FYK1) | 0.7488 | 380449 | 380678 | 25 |
| 14 | TSLM(lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK7) | 0.7061 | 385948 | 386143 | 21 |
| 15 | TSLM(R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) | 0.7563 | 379407 | 379788 | 43 |

| Model number | Model name | Adjusted $R^2$ | AIC | BIC | Numbers of parameters |
|---|---|---|---|---|---|
| 16 | ARIMA(pdq(5,0,0)-lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK7-FWK5-FYK3) | - | 425667 | 426040 | 42 |
| 17 | ARIMA(pdq(0,1,0)-lag1R-lag2R-lag3R-R1-R7-DIxR1-DI-FDK7-FWK5-FYK3) | - | 427933 | 428255 | 37 |
| 18 | ARIMA(pdq(5,0,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) | - | 425447 | 425870 | 48 |
| 19 | ARIMA(pdq(0,1,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) | - | 427775 | 428147 | 43 |

### 3.5 Model accuracy

The full linear model (model 15) and the full ARIMA model (model 18) were tested on the test set and the results can be seen in table 5.

*Table 5: Reports the mean absolute error (MAE) and the mean absolute percentage error (MAPE) for the two linear models including all parameters with and without ARIMA errors.*

| Models | MAE [$m^3/m$] | MAPE [%] |
|---|---|---|
| Full linear model | 163 | 19.4 |
| Full ARIMA model | 168 | 20.1 |

To graphically investigate the two models, residual plots and a visualization of the two models were made. Figure 4 and Figure 5 shows the residual plots of the full linear model and the full ARIMA model respectively.

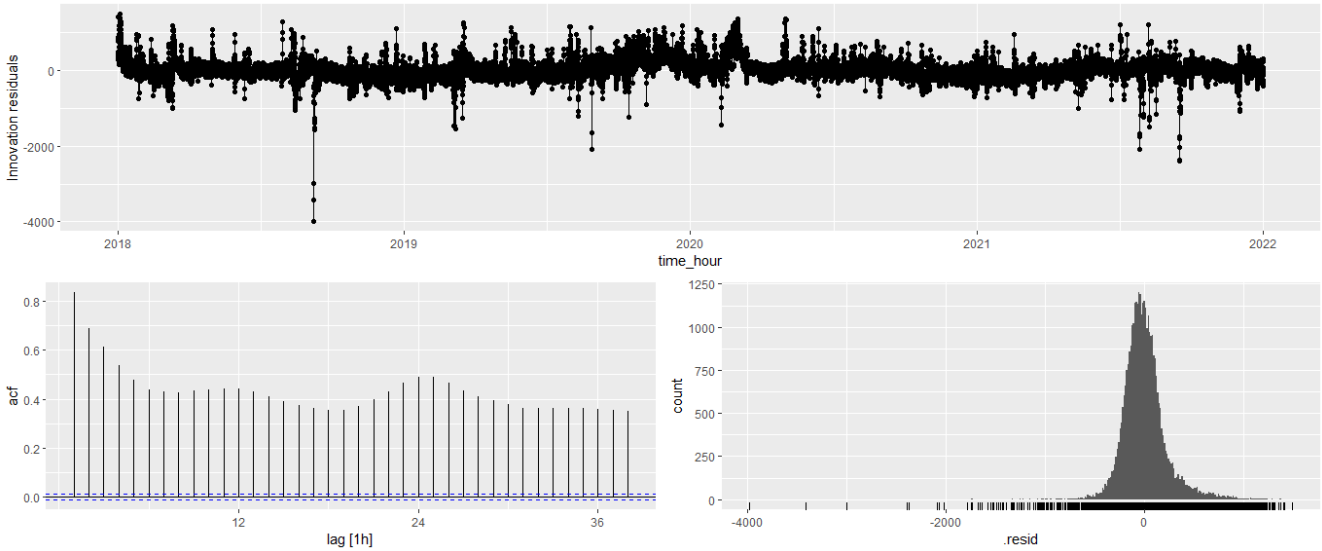


*Figure 4: shows the residual plots of the full linear model.*

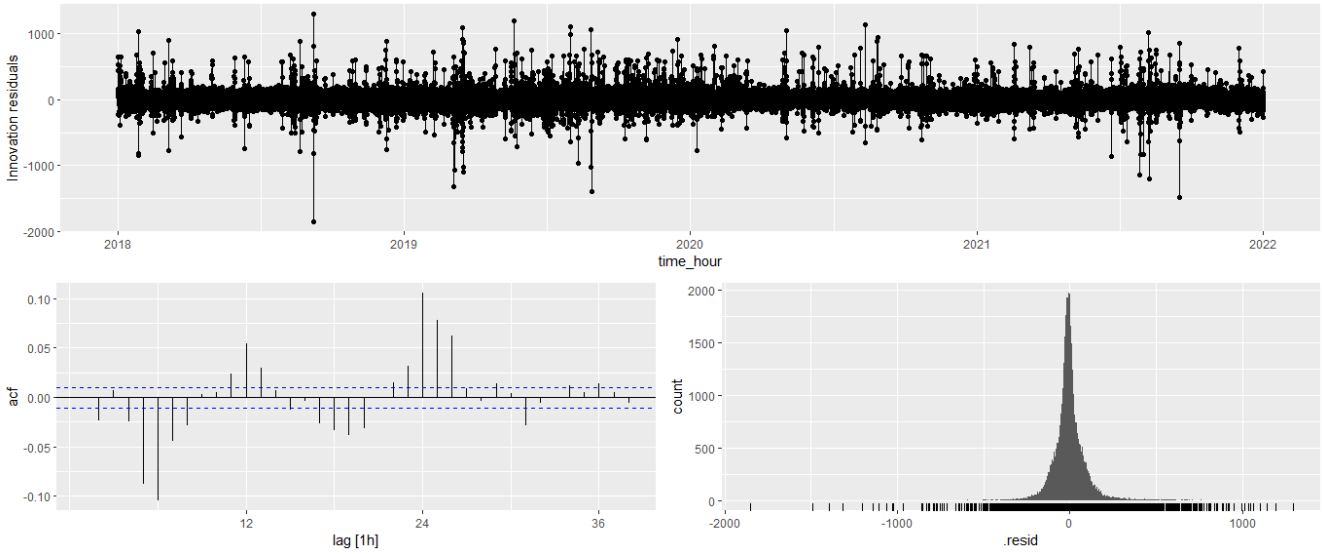


*Figure 5: shows the residual plots of the ARIMA linear model.*

Figure 6 and Figure 7 visualize how two models, the full linear model and the full ARIMA model, fitted the data from the test set respectively.

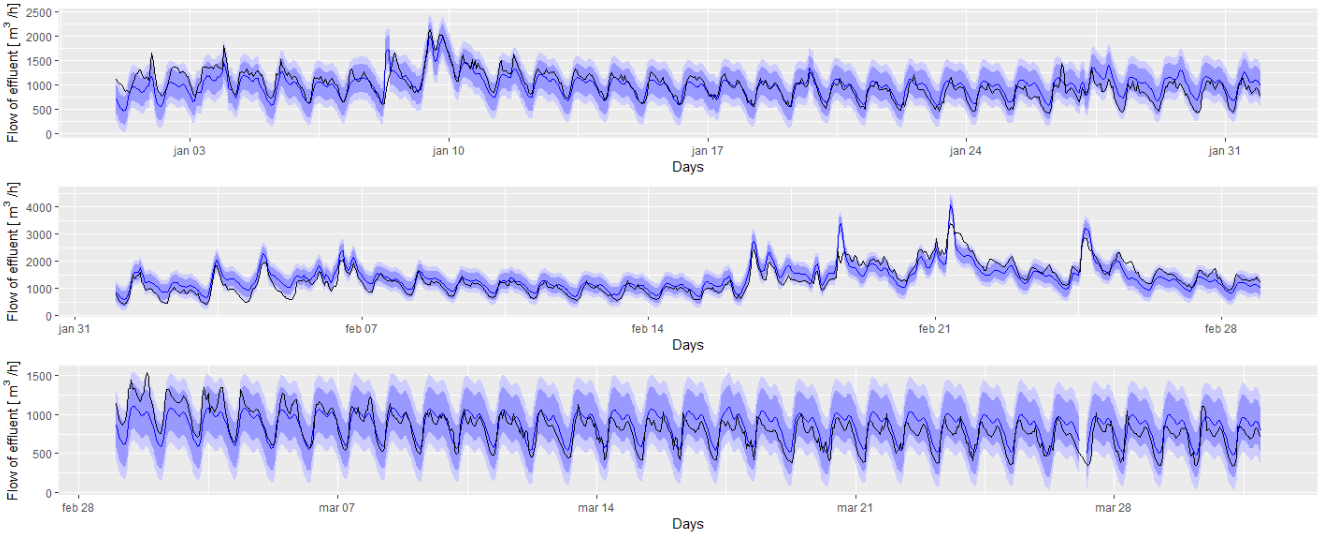


*Figure 6: shows the full linear model predictions of the flow and the data from the test set.*

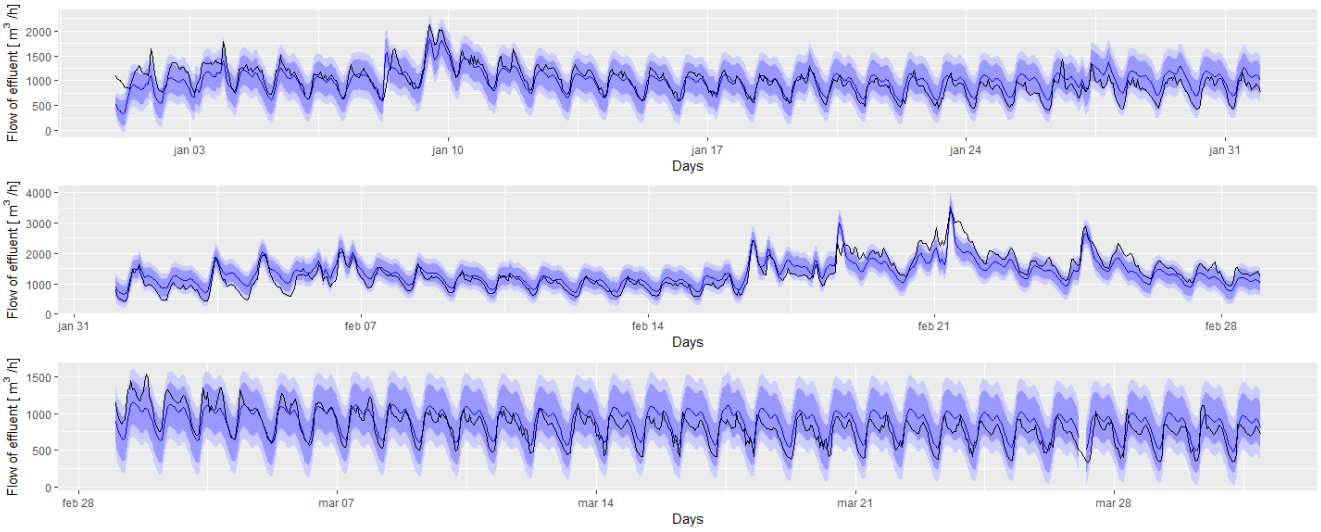


*Figure 7: shows the full ARIMA model predictions of the flow and the data from the test set..*

# 4 Discussion

## 4.1 Forecasting horizon

The graphical investigation (Appendix C) showed that the aggregated data with time intervals of 30 minutes or 1 hour were robust towards outliers and did not show noisy behavior.
In Table 2, the mean and 25% quantile of the HRT are reported to be 19 and 25 hours respectively. In other words, in 25 hours all the water in the process tanks would have been replaced at least once for 75% of the days with the rain precipitation less than 2 mm. In section 3.1, the flow response time was determined to be between 30 to 60 minutes. This is additional time that the process tanks have to prepare for the rain event, as the rainwater has to travel through the sewage system. Therefore, the flow response time can be subtracted from the HRT to calculate the forecasting horizon. Thus, the forecasting was determined to be 24 hours. Furthermore, the rain precipitation forecast 24 hours ahead are assumed to be reliable, hence a forecasting horizon of 24 hours can be accepted. The flow data aggregated to 1-hour intervals was chosen as this graphically showed to be robust towards outliers and have no noisy behavior. Statistical uncertainty increased with numbers of observations to forecast (Hyndman et al., 2021, Chapter 5.5). For this reason, the hourly data was selected over the data aggregated over 30 minutes, as this reduced the number of forecasting points from 48 to 24.

## 4.2 Seasonality

The daily seasonality can be seen in Figure 1. The daily seasonality is consistent with expectations, as the flow is lower at night, when people are asleep and higher doing the day, when people are awake. The daily pattern is also apparent on in the weekly seasonality plot (Figure 2), a slight smoothing of the daily pattern can be observed doing the weekend. This aligns with expectations as peoples' behavior in the weekend is less stringent. The yearly seasonality plot (Figure 3) shows no apparent information about the flow variations. Based on the plots it can be concluded that the daily seasonality expresses the strongest pattern. In appendix D, the flow time series was decomposed with the STL function from the fable package (O'Hara-Wild et al., 2021a). Similarly the decomposition found that the daily pattern expresses the strongest seasonality.

## 4.3 Modelling

To model the effluent flow, 19 different linear models were created and tested, see Table 4. A graphical visualisation of how the model fits the data in the year of 2020 can be seen in Appendix E. Model 15 showed the best performance of the linear models without ARIMA modelling of the errors, and model 18

showed the best performance of the linear models with ARIMA modelling of the errors. Common for both models is that they include all the tested parameters.

Model 1 was a regression of the flow only based on the current rain precipitation. This model yielded a terrible result. The inclusion of the accumulated rain and the drought index significantly increased the adjusted $R^2$ values. Comparing model 2 and model 4 shows that the exclusion of the accumulated rain from the last 2 to 6 days only slightly deteriorates the model, indicating a relationship between the accumulated rain parameters. As the accumulated rain from yesterday is included in the accumulated rain from 2 days ago and so on, logically some autocorrelation exists.

The interaction between the one day accumulated rain and the drought index was tested in model 4 to 6. The drought index is zero when the soil is saturated with water, hence a high drought index will result in considerable amount of rainwater absorbed into the soil instead of being transported to the WWTP. The interaction parameter has a negative slope, resulting in a negative value and reduce the flow to the WWTP. When the drought index is high despite rain precipitation in the last 24 hours, the interaction parameter would be largely negative. This aligns with the expectation of the rainwater being absorbed in the soil, and the inclusion of the interaction parameters improves the models' adjusted $R^2$. The current rain precipitation at time $t$ might not be the best predictor of the flow at time $t$, as the flow response time was between 30 and 60 minutes (shown in section 3.1). Models 6 to 9 test the inclusion of predicting the flow based on lagged rain precipitation values. Model 9 includes the lagged rain precipitation for the last 1 to 3 hours, which yielded a better model compared to model 6. Model 9 and 10 show that the exclusion of the current rain precipitation only slightly deteriorates the model. Models 10 to 14 test the effect of adding a DHR model to include the daily, weekly and yearly seasonality. The number of Fourier terms in model 11 was based on a similar problem described in Hyndman et al. (2021, Chapter 12.1). Models 12 to 14 tested the effect of reducing the numbers of parameters and exclusion of the weekly and yearly seasonality, as these appeared to be weaker seasonalities as earlier described. The best model was model 12 which included a daily variation with 7 pairs of sine and cosine terms, 5 pairs for the weekly seasonality, and 3 pairs for the yearly seasonality. Model 15, a linear model that includes all the parameters tested in the previous models, was the model with the highest adjusted $R^2$, and lowest AIC and BIC. Hence, this model is superior to the rest of the models. Model 12 (including parameters that have been chosen subjectively) and 15 (including all described parameters) were tested with the assumption that the errors could be modelled by an ARIMA model. The ARIMA parameters were chosen based on an inbuilt algorithm in the fable package (O'Hara-Wild et al., 2021a), and models with 0 or 1 degree of differencing was tested. Model 18 included all parameters plus an autoregressive model with 5 parameters was found to be the superior model with ARIMA modelling of the errors.

### 4.4 Validating models

To validate the performance of the linear model including all parameters with and without ARIMA modelled errors (model 15, full linear model, and 18, full ARIMA model, respectively), the accuracy was calculated on the test set, which has been withhold in the training of the models.
When comparing the MAE and MAPE, the linear model showed slightly better results. The MAE is the mean of the errors, i.e. the average error to expect in the full linear model, which is 163 $m^3/h$. The MAPE finds the mean of the percentage error, corresponding to a average error to expect. In the full linear model this is 19.4%. An error of 19.4% percent is considerable error, which preferably should be reduced. When investigating the residual plot (Figure 4), the autocorrelation function (ACF) plot shows that $y_t$ correlates with at least the 38 last observations. This indicates that there is information the model fails to extract. Another possible way of improving the model, and thereby also removing some of the autocorrelation in the residuals, is to find new meaningful parameters to include in the model. Figure 6 shows the measured data and the fitted model, which has been provided with the real measurement for rain precipitation and drought index. Graphically it can be seen that the model does not fit the data perfectly, and tends to slightly over estimate the peaks in the flow. Even though the model could be improved, Figure 6 shows that the model captures the seasonality and predicts the peaks in effluent flow well. As a WWTP is a complicated systems, some uncertainty is expected and a forecast with 19.4% error would still in many cases produce useful insight of the coming operation. The full ARIMA model had slightly higher MAE and MAPE values and an oscillating behaviour can be seen in the ACF plot (Figure 5). Figure 7 shows the full ARIMA model fitted to the test set and compared to Figure 6 no apparent distinction can be seen. It can be concluded that the full linear model is preforms slightly better.

### 4.5 Future work

This project lays the foundation for the further development of an alternative control strategy at Egå WWTP. Future work could include the design of models that can predict ammonium concentrations at multiple locations at the WWTP. Furthermore, testing other types of models to predict the flow e.g. by using long short-term memory neural network, could be useful to improve forecasting accuracy. Lastly, if the forecasting models should be able to predict when an ammonium peak will appear, it is of utmost importance to find a way to mathematically describe the causal relationship between flow and ammonium peaks.

# 5 Conclusion

Throughout this article, different models to forecast the flow out of Egå WWTP have been evaluated. It was found that the superior model was a linear model containing the rain precipitation for the current time $t$ and the last 1 to 3 hours; the accumulated rain from the last 1 to 7 days; the drought index; an interaction parameter of the drought index and the accumulated rain for the previous day; 7, 5, and 3 pairs of sine and cosine terms on a daily, weekly, and yearly period respectively. The model was evaluated and a mean absolute error of 163 $m^3/h$ and a mean absolute percentage error of 19.4% was found. This error is considerable and future work should aim to lower this error. However, it is expected that the investigated model still will provide useful insight of the operation of Egå WWTP.

As the current control strategy at Egå WWTP does not include forecasting, this article lay the foundation for the future development of an alternative control strategy, which takes heavy rain precipitation into account. This article proves that data driven models trained on large amount of data, are a feasible way of creating forecasting models of effluent flow at Egå WWTP. Therefore, these types of models are expected to yield good results in forecasting other parameters of Egå WWTP, and would be a recommendable element in the development of an alternative control strategy.

## References


Cappelen, John (2022). *Så vildt kan det regne i Danmark*. DMI. URL: http://www.dmi.dk/nyheder/2016/sa-vildt-kan-det-regne-i-danmark/ (visited on 05/26/2022).

Cheon, Seong-Pyo et al. (2008). "Learning Bayesian networks based diagnosis system for wastewater treatment process with sensor data." In: *Water Science and Technology: A Journal of the International Association on Water Pollution Research* 58.12, pp. 2381–2393. ISSN: 0273-1223. DOI: 10.2166/wst.2008.839.

Ellis, G. W., X. Ge, and D. Grasso (Jan. 1, 1990). "TIME SERIES ANALYSIS OF WASTEWATER QUALITY." In: *Instrumentation, Control and Automation of Water and Wastewater Treatment and Transport Systems*. Ed. by R. Briggs. Pergamon, pp. 441–448. ISBN: 978-0-08-040776-0. DOI: 10.1016/B978-0-08-040776-0.50059-7. URL: https://www.sciencedirect.com/science/article/pii/B9780080407760500597 (visited on 05/25/2022).

Granata, Francesco et al. (Feb. 2017). "Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators." In: *Water* 9.2. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 105. ISSN: 2073-4441. DOI: 10.3390/w9020105. URL: https://www.mdpi.com/2073-4441/9/2/105 (visited on 05/25/2022).

Guo, Hong et al. (June 1, 2015). "Prediction of effluent concentration in a wastewater treatment plant using machine learning models." In: *Journal of Environmental Sciences* 32, pp. 90–101. ISSN: 1001-0742. DOI: 10.1016/j.jes.2015.01.007. URL: https://www.sciencedirect.com/science/article/pii/S1001074215001278 (visited on 05/25/2022).

Hansen, Laura Debel, Mikkel Stokholm-Bjerregaard, and Petar Durdevic (Apr. 1, 2022). "Modeling phosphorous dynamics in a wastewater treatment process using Bayesian optimized LSTM." In: *Computers & Chemical Engineering* 160, p. 107738. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2022.107738. URL: https://www.sciencedirect.com/science/article/pii/S0098135422000795 (visited on 05/25/2022).

Hyndman, R.J. and G. Athanasopoulos (2021). *Forecasting: Principles and Practice (3rd ed)*. OTexts: Melbourne, Australia. OTexts.com/fpp3. (Visited on 05/25/2022).

Kang, Hoon et al. (Dec. 1, 2020). "Time Series Prediction of Wastewater Flow Rate by Bidirectional LSTM Deep Learning." In: *International Journal of Control, Automation and Systems* 18.12, pp. 3023–3030. ISSN: 2005-4092. DOI: 10.1007/s12555-019-0984-6. URL: https://doi.org/10.1007/s12555-019-0984-6 (visited on 05/25/2022).

Metcalf and Eddy, eds. (2014). *Wastewater engineering: treatment and resource recovery*. 5. ed., internat. student ed. New York, NY: McGraw-Hill. ISBN: 978-1-259-01079-8.

O'Hara-Wild, Mitchell, Rob Hyndman, and Earo Wang (2021a). *fable: Forecasting Models for Tidy Time Series*. R package version 0.3.1. URL: https://CRAN.R-project.org/package=fable.

— (2021b). *fabletools: Core Tools for Packages in the 'fable' Framework*. R package version 0.3.2. URL: https://CRAN.R-project.org/package=fabletools.

— (2021c). *feasts: Feature Extraction and Statistics for Time Series*. R package version 0.2.2. URL: https://CRAN.R-project.org/package=feasts.

*Tørkeindeks* (2022). DMI. URL: http://www.dmi.dk/dmis-vejrprodukter/torkeindeks/ (visited on 05/26/2022).

Van Dongen, G. and L. Geuens (Mar. 1, 1998). "Multivariate time series analysis for design and operation of a biological wastewater treatment plant." In: *Water Research* 32.3, pp. 691–700. ISSN: 0043-1354. DOI: 10.1016/S0043-1354(97)00249-2. URL: https://www.sciencedirect.com/science/article/pii/S0043135497002492 (visited on 05/25/2022).

Wang, Dong et al. (Aug. 25, 2021). "A machine learning framework to improve effluent quality control in wastewater treatment plants." In: *Science of The Total Environment* 784, p. 147138. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2021.147138. URL: https://www.sciencedirect.com/science/article/pii/S0048969721022087 (visited on 05/25/2022).

Zhang, Qianqian et al. (Aug. 1, 2019). "Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network." In: *Water Science and Technology* 80.2, pp. 243–253. ISSN: 0273-1223. DOI: 10.2166/wst.2019.263. URL: https://doi.org/10.2166/wst.2019.263 (visited on 05/25/2022).

# A Effluent flow vs ammonium concentration

This appendix show the the effluent flow of Egå WWTP for each year and the ammonium concentration in the effluent to the corresponding time. The ammonium concentration is in red and the flow is in black.

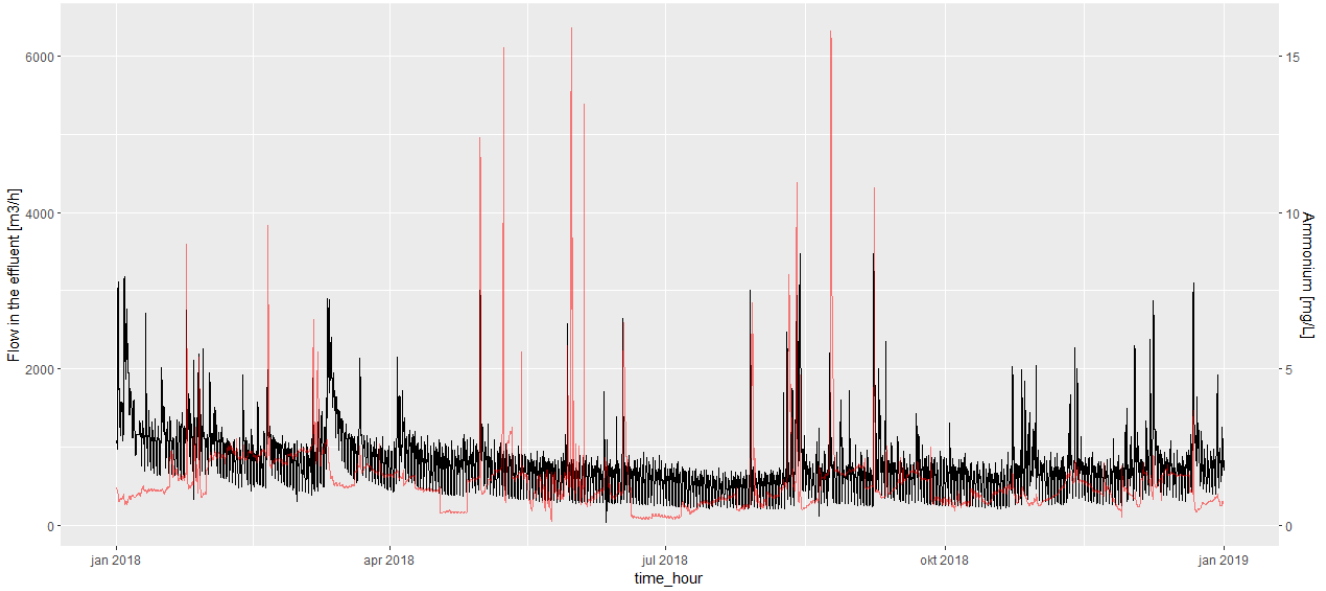


*Figure 8: The plot show the effluent flow (black) and the ammonium concentration in the effluent (red) in 2018.*

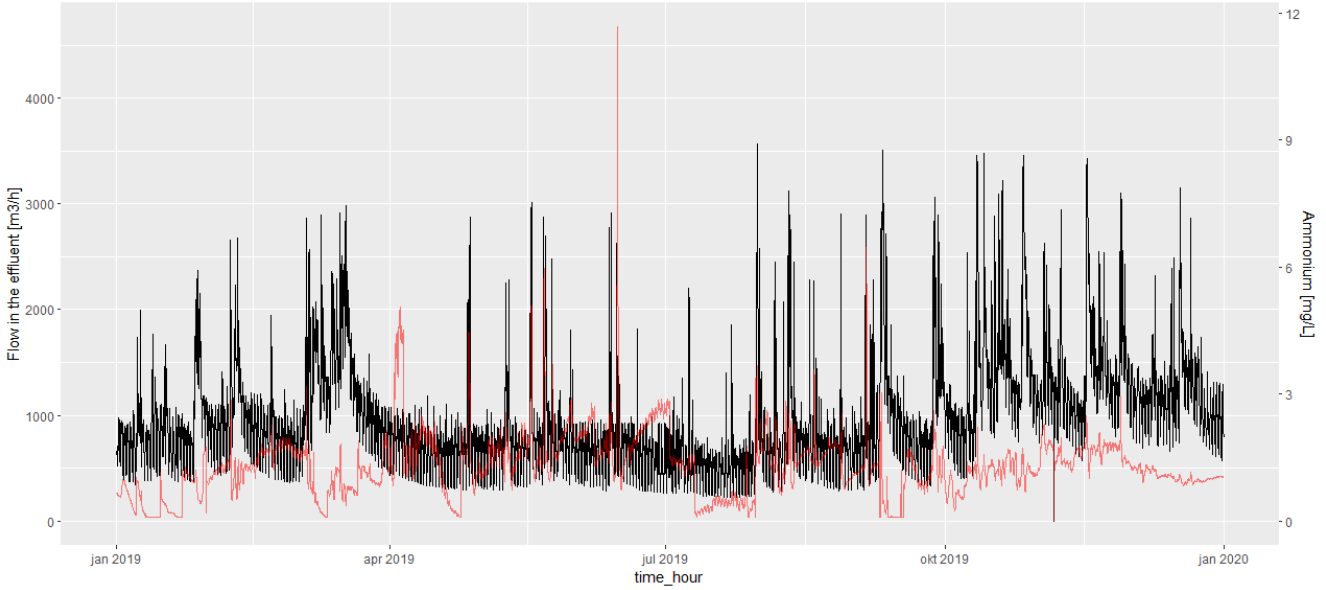


*Figure 9: The plot show the effluent flow (black) and the ammonium concentration in the effluent (red) in 2019.*

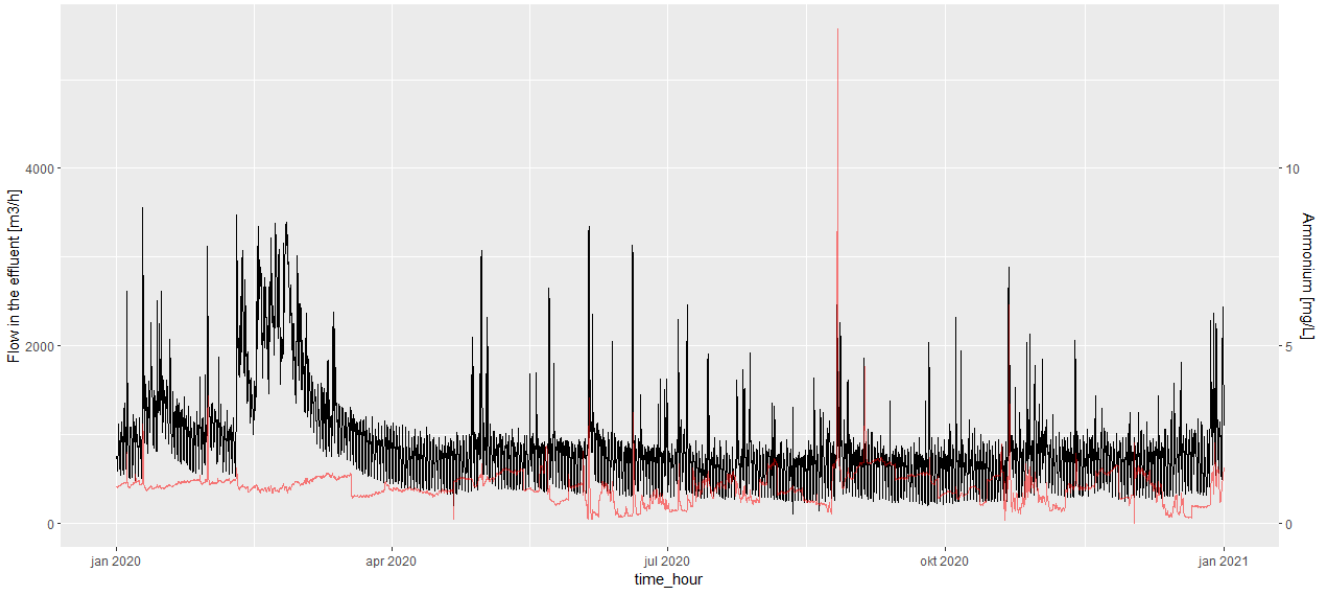


*Figure 10: The plot show the effluent flow (black) and the ammonium concentration in the effluent (red) in 2020.*

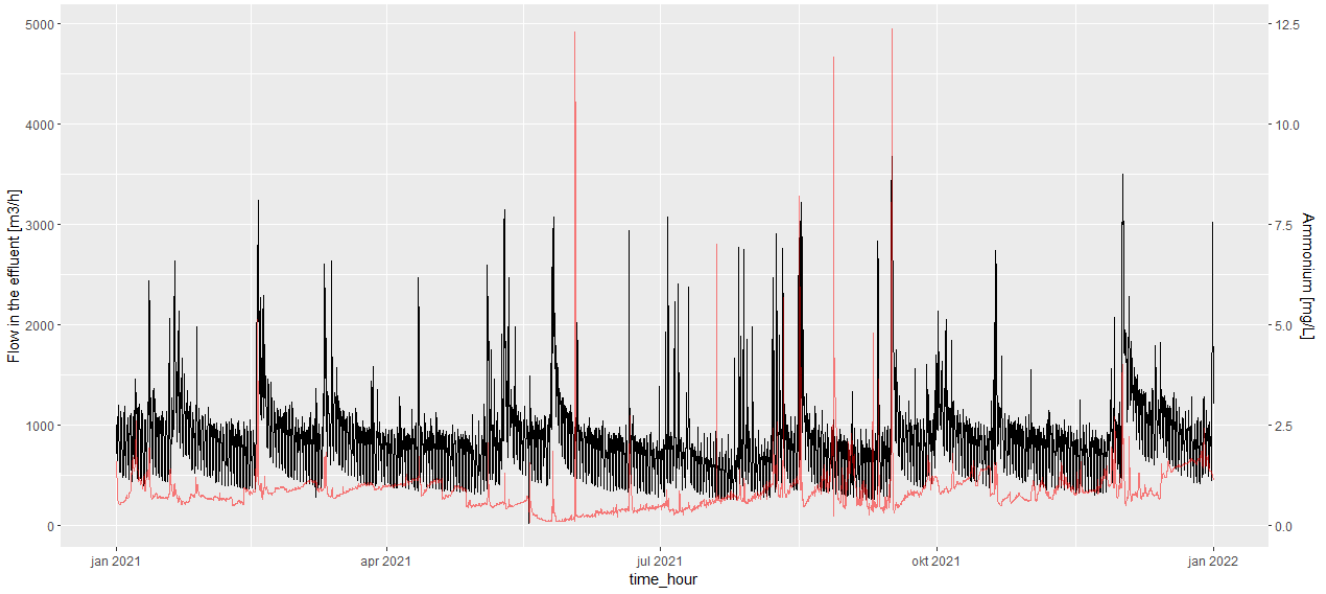


*Figure 11: The plot show the effluent flow (black) and the ammonium concentration in the effluent (red) in 2021.*

# B Flow response time

The following six graphs have the same layout. The rain precipitation in mm over time is the upper graph, the effluent flow in $m^3/h$ is the middle graph, and the bottom graph is the ammonium concentration in mg/L i.

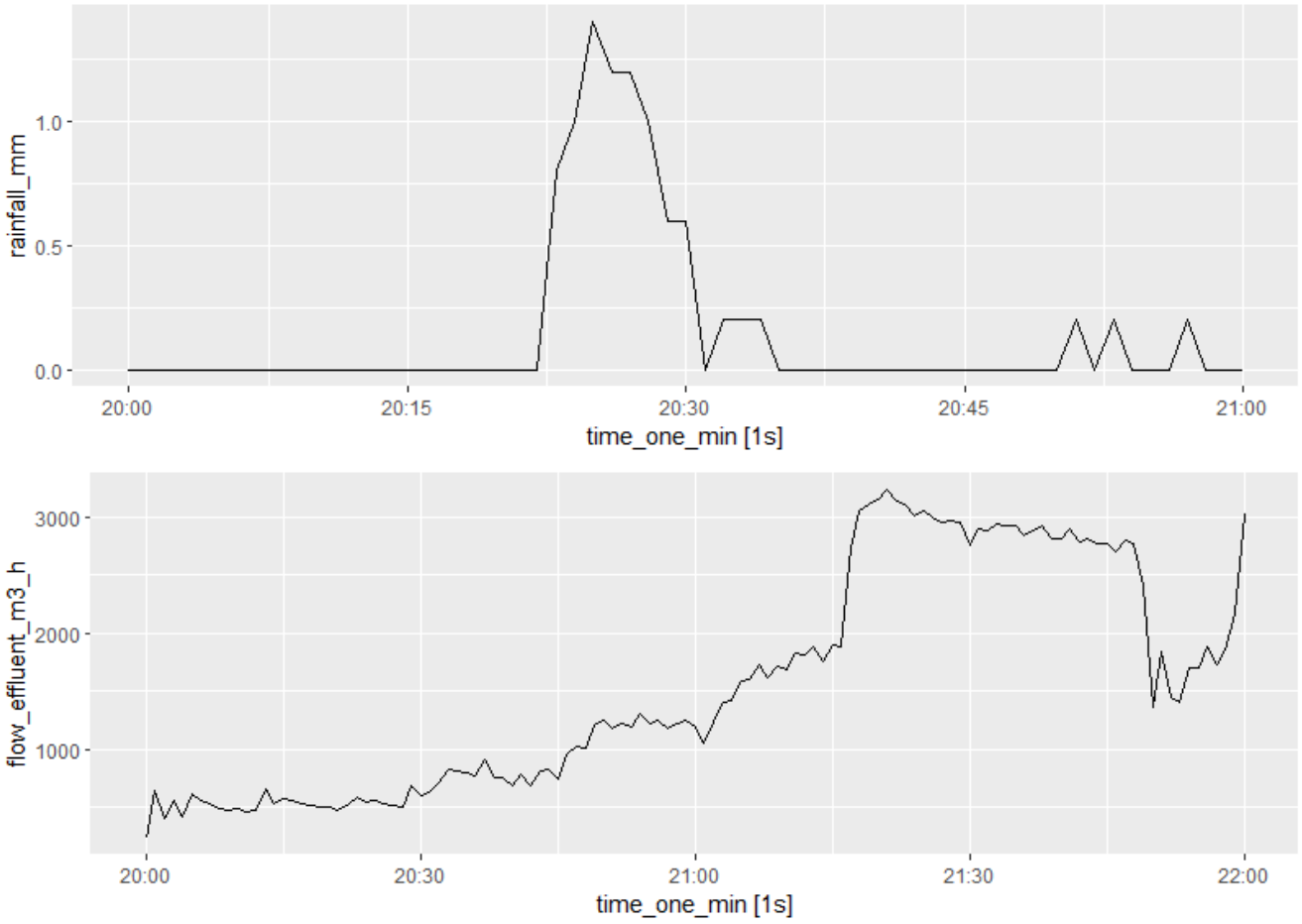


*Figure 12: The plot shows the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2018-08-28.*

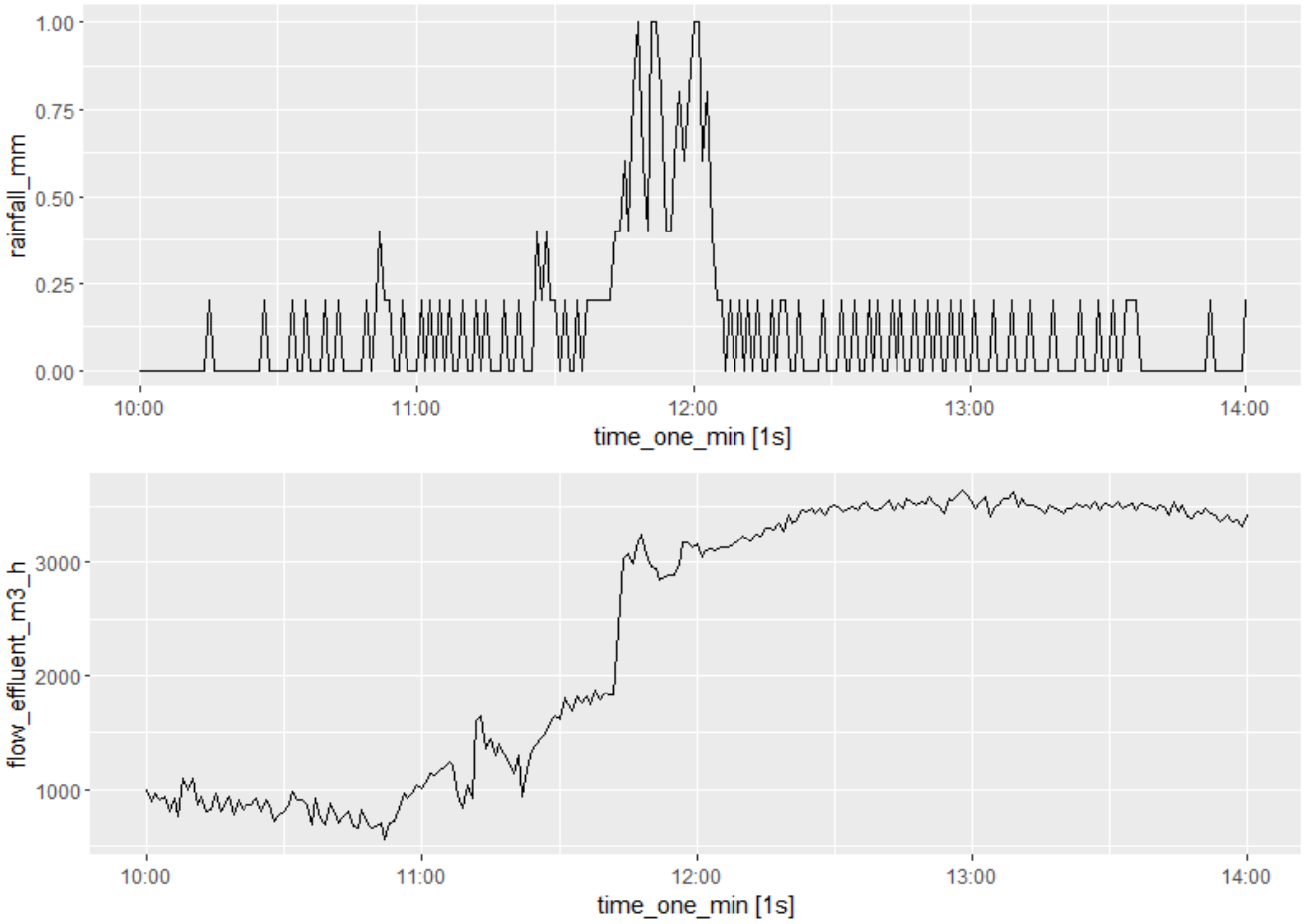


*Figure 13: The plot show the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2018-09-07.*

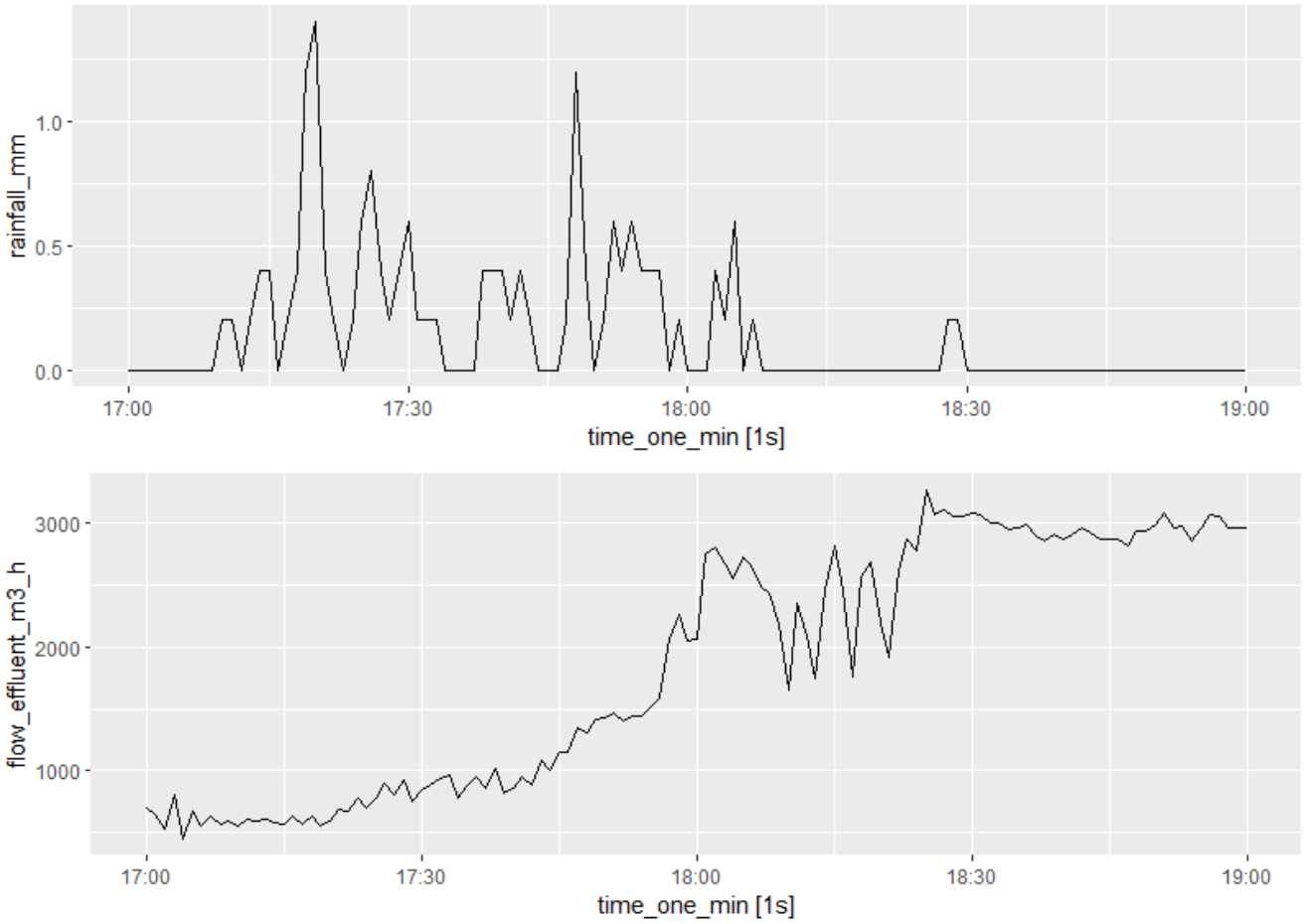


*Figure 14: The plot show the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2019-08-27.*

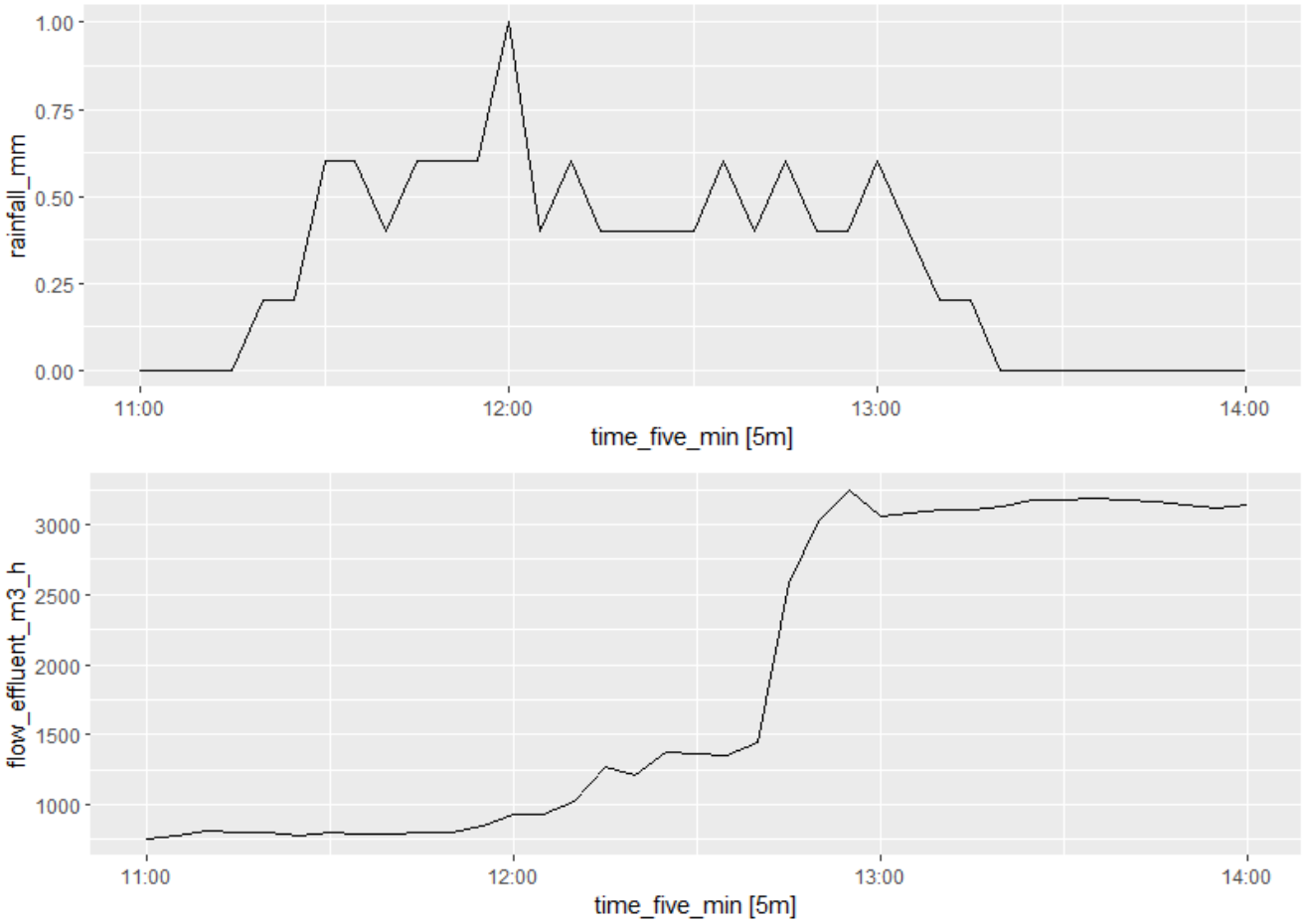


*Figure 15: The plot show the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2020-08-19.*

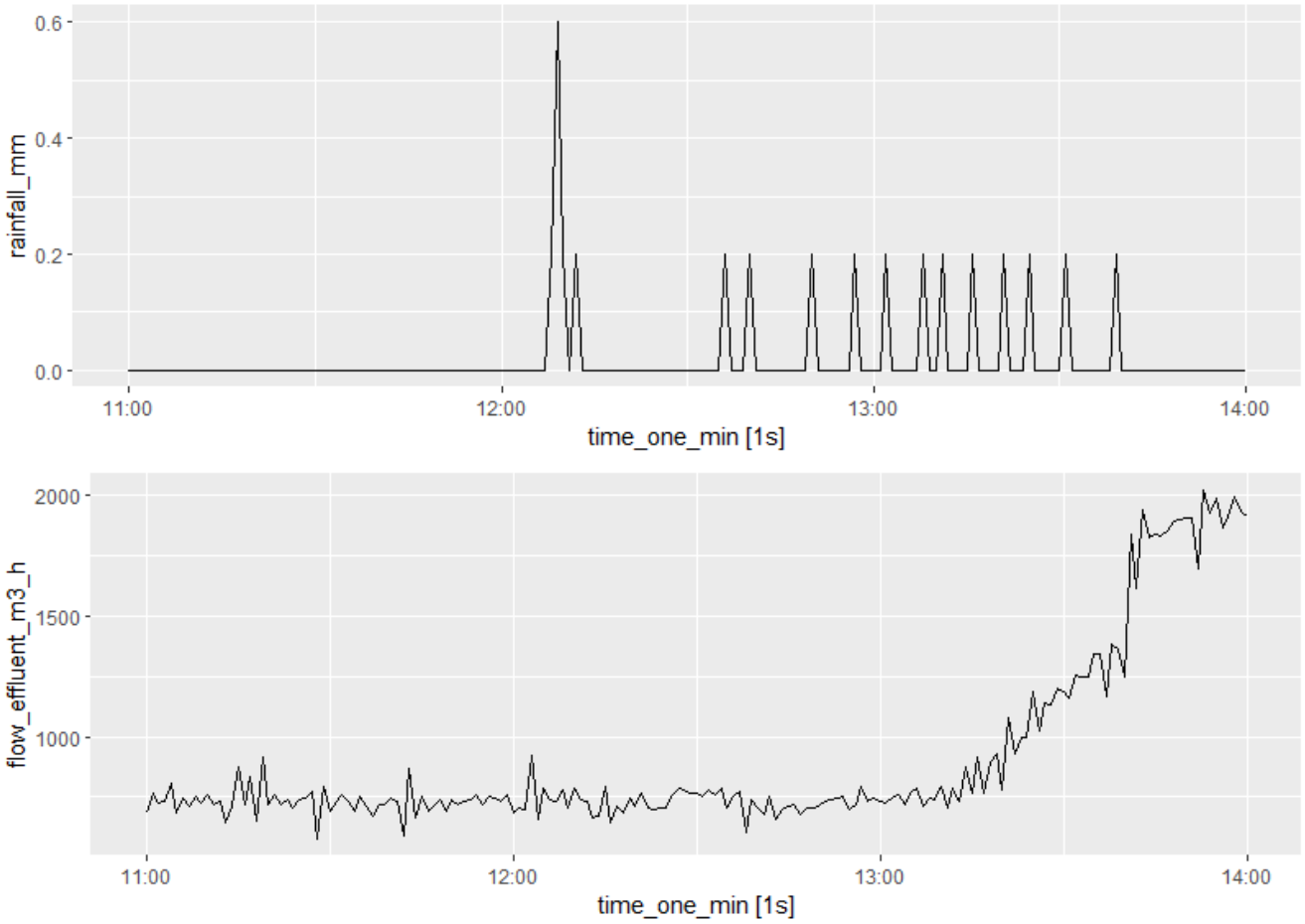


*Figure 16: The plot show the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2020-08-18.*

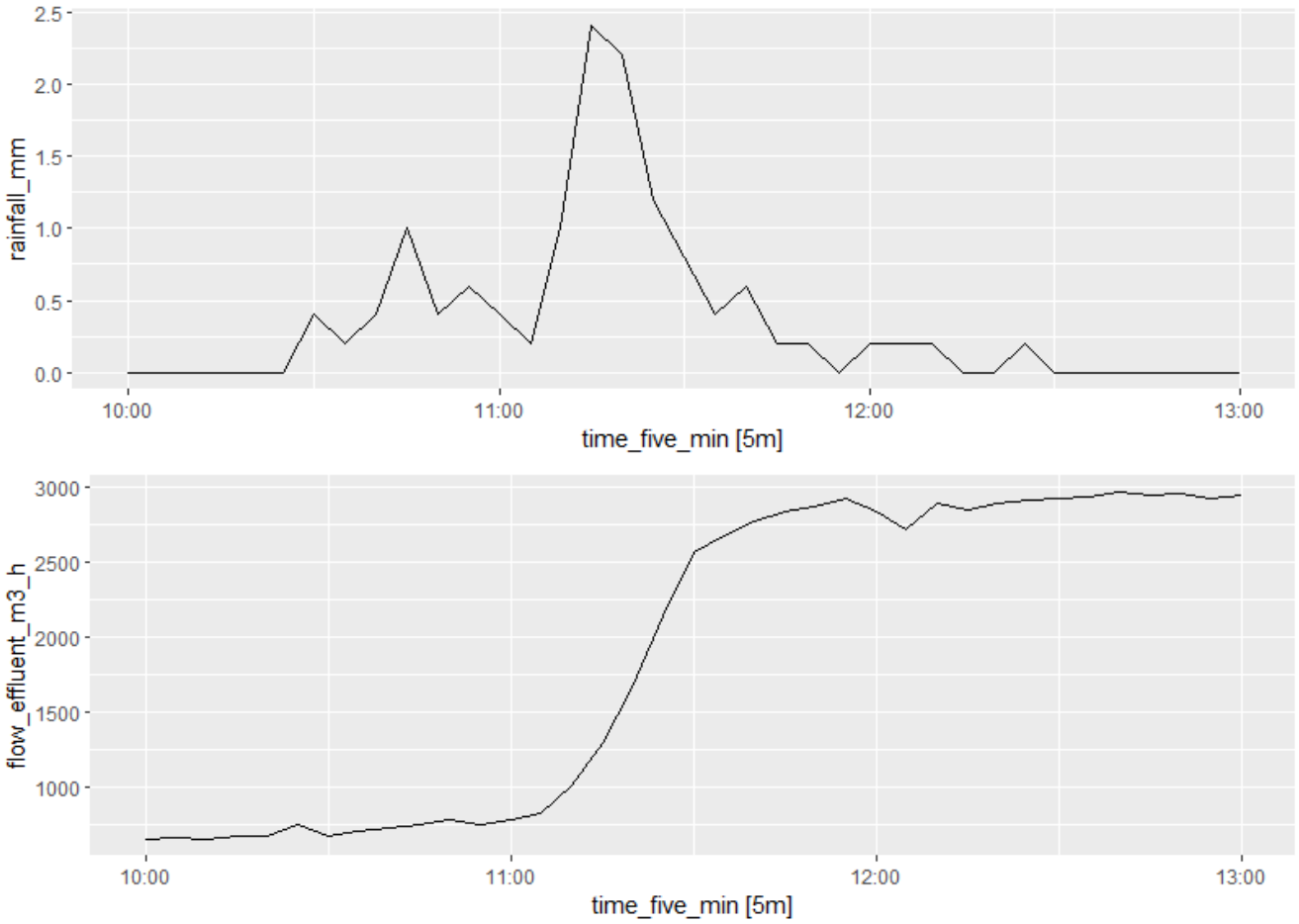


*Figure 17: The plot show the rain precipitation in mm over time in the upper graph, the effluent flow in $m^3/h$ in the middle graph, and the ammonium concentration in mg/L in the bottom graph for the date 2021-06-20.*

# C Data aggregation

This appendix show the the effluent flow of Egå WWTP in January of 2021 . The graphs in this appendix is a visualisation of how the averaging over different time intervals affects the data smoothness and information.

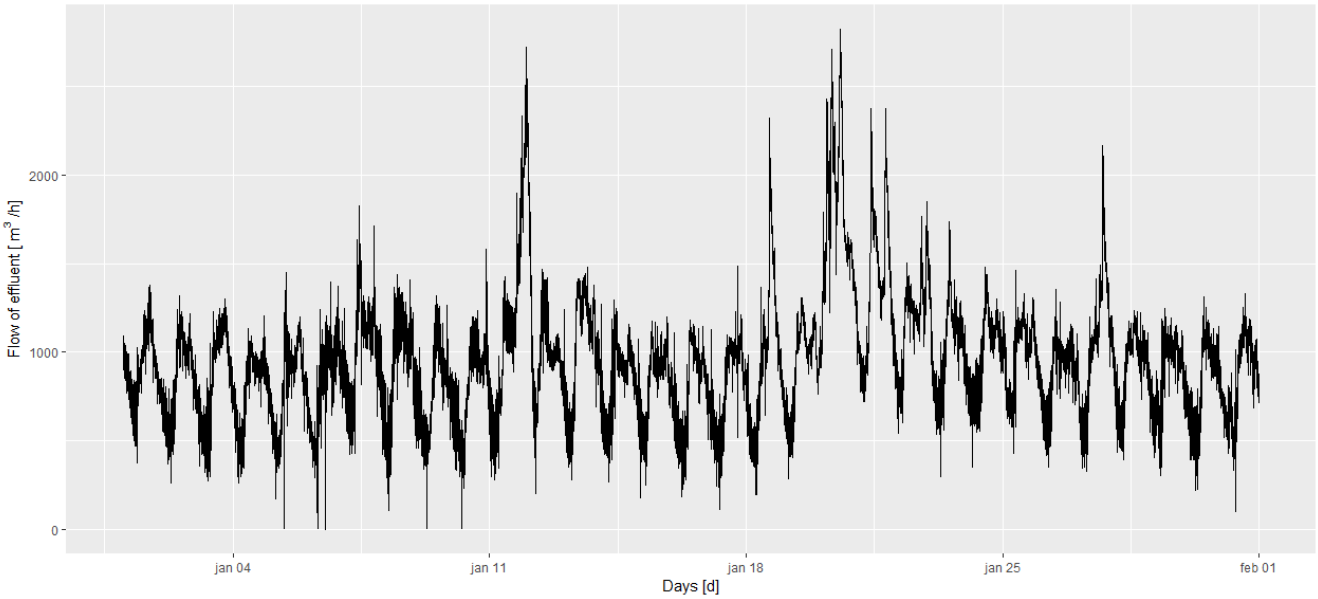


*Figure 18: The plot show the effluent flow in January of 2021 measured each 1 minute.*

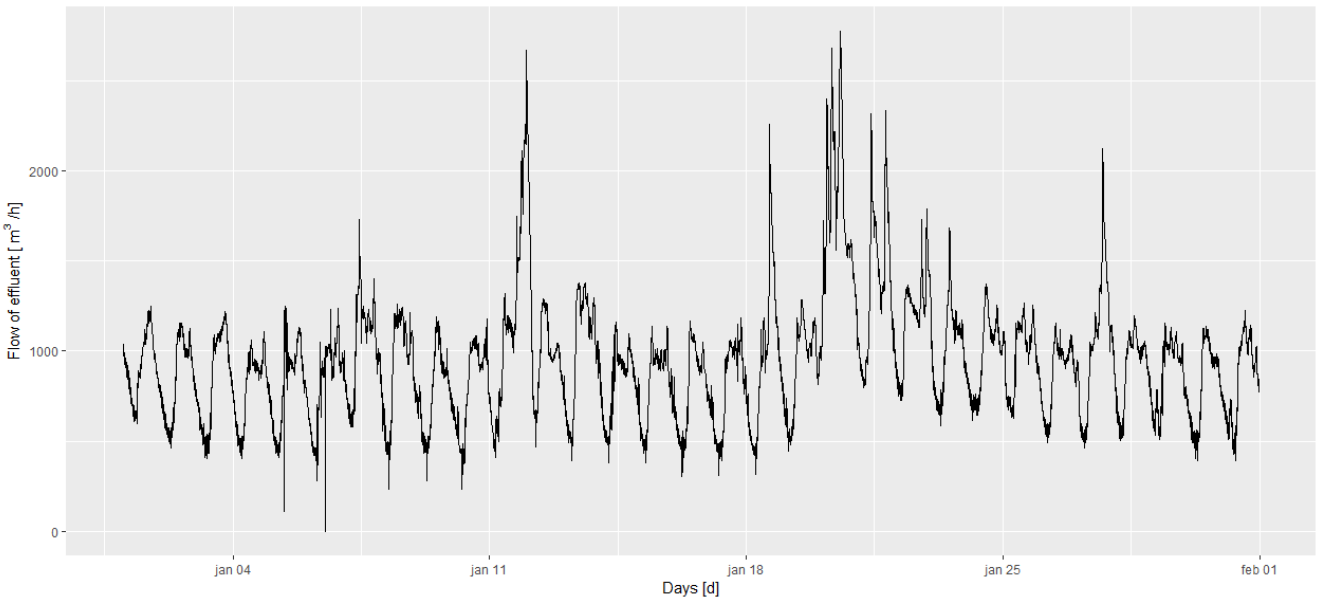


*Figure 19: The plot show the effluent flow in January of 2021 with average values for each 5 minutes.*

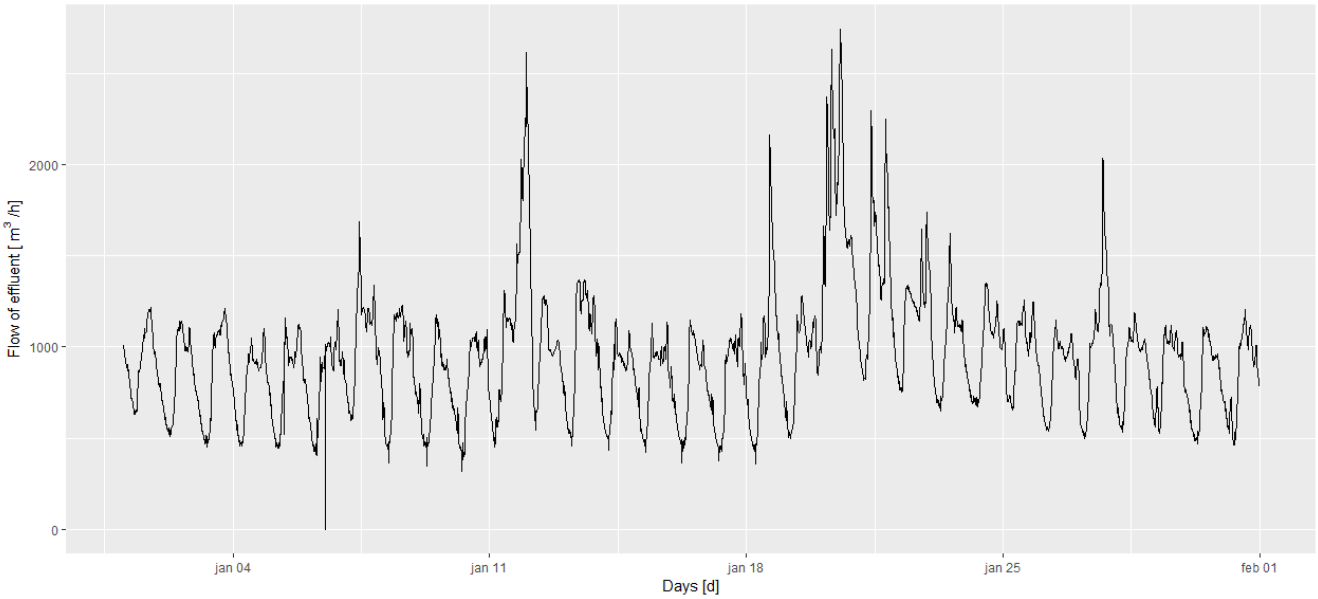


*Figure 20: The plot show the effluent flow in January of 2021 with average values for each 15 minutes.*

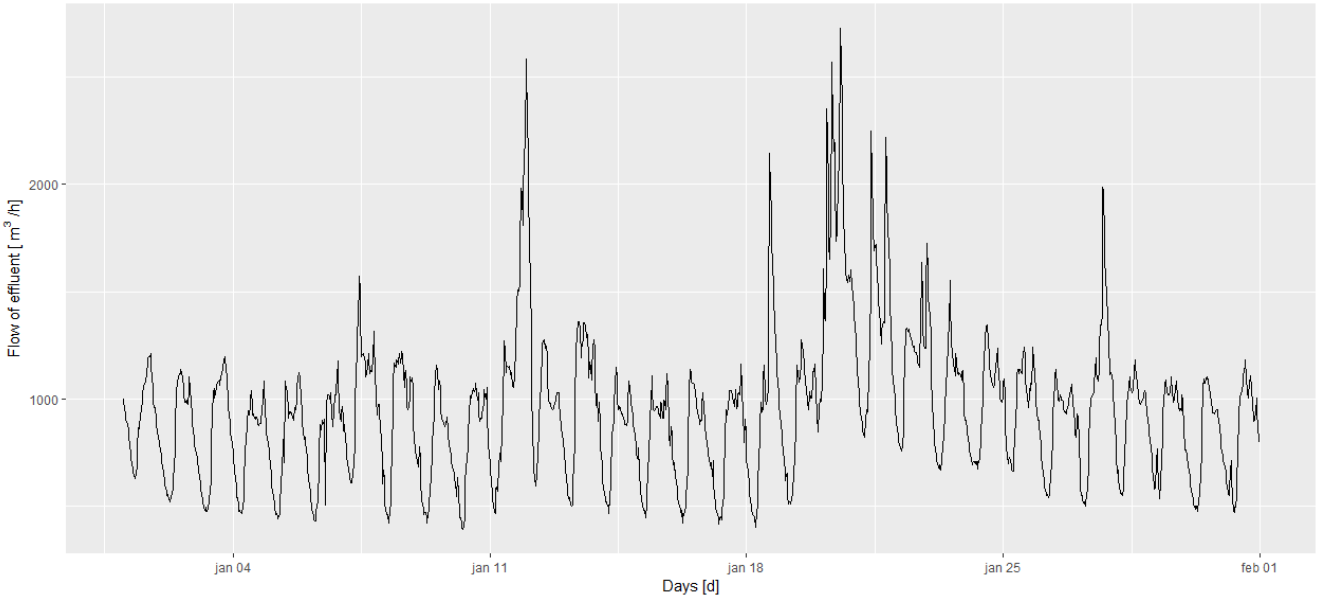


*Figure 21: The plot show the effluent flow in January of 2021 with average values for each 30 minutes.*

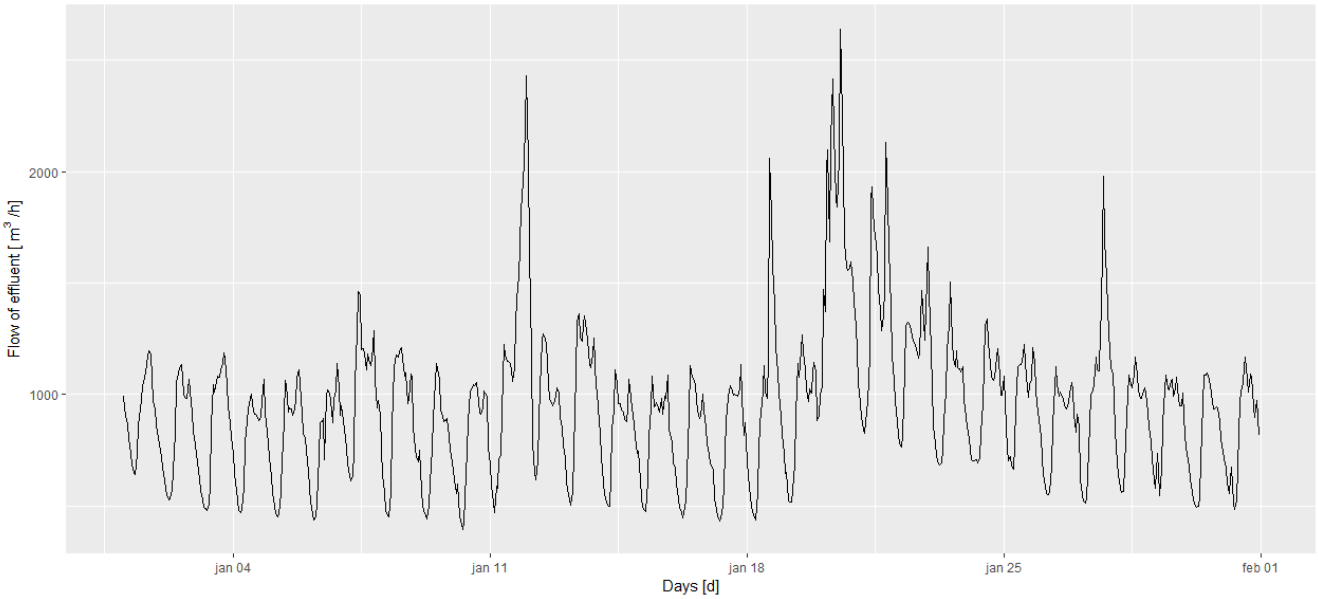


*Figure 22: The plot show the effluent flow in January of 2021 with average values for each hour.*

# D STL decomposition zoom on the first 1000 data points

The figure 23 show the STL decomposition of the flow time series, with an assumption of a fix seasonal pattern for both the daily seasonality (season_24), weekly seasonality (season_168), and the yearly seasonality (season_8766).

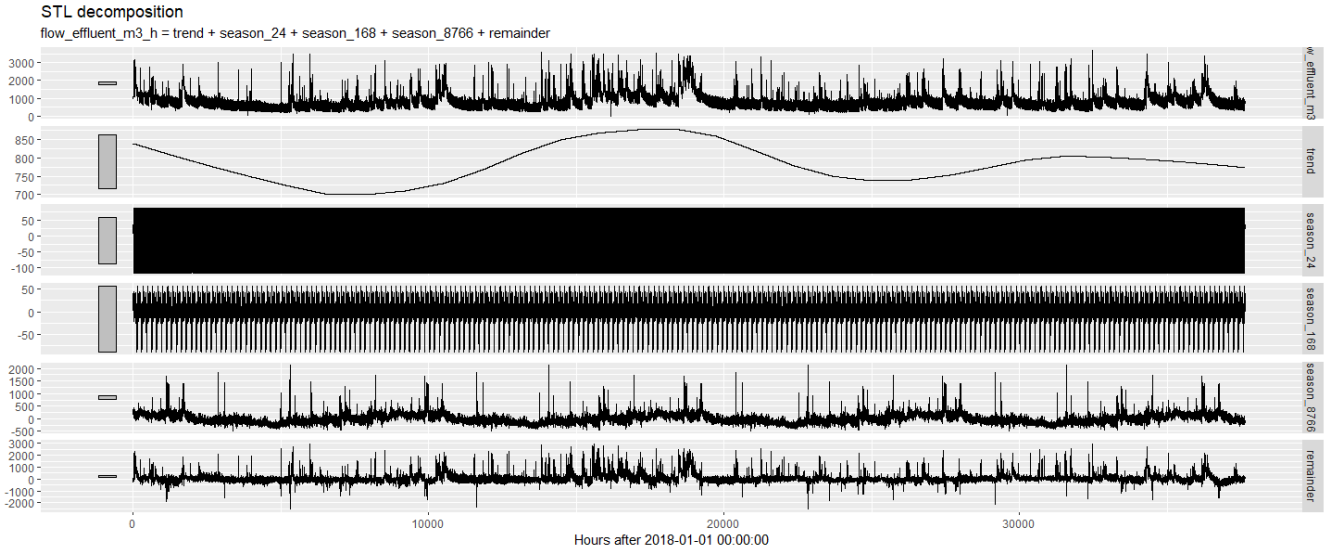


*Figure 23: shows the STL decomposition of the flow time series.*

The figure 24 show the first 1000 measurements of the STL decomposition.

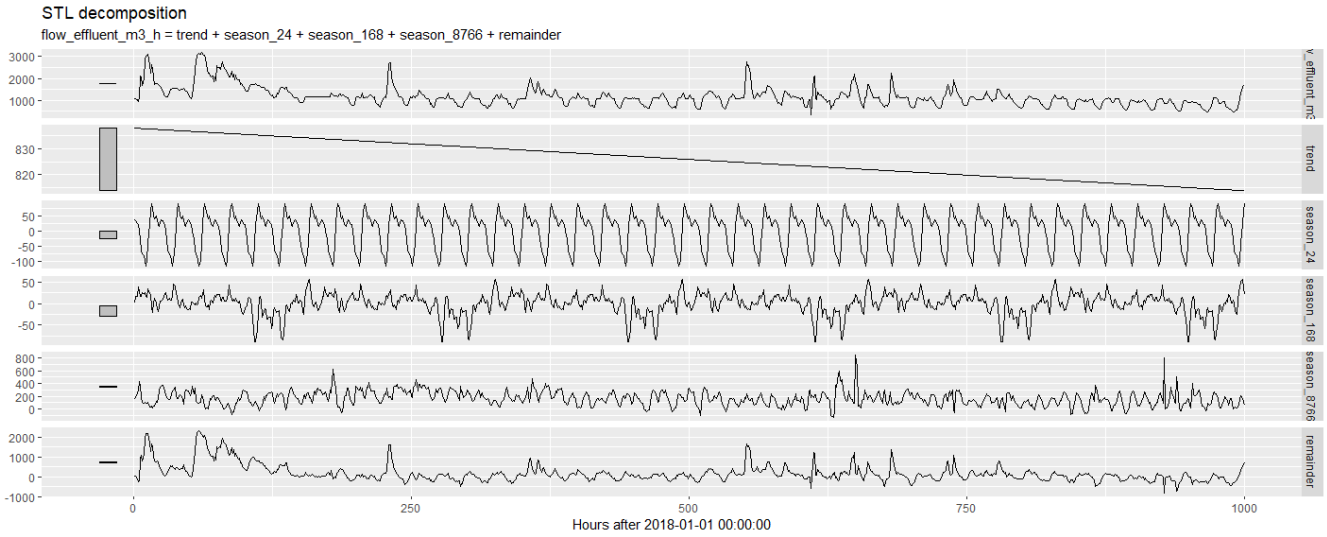


*Figure 24: show the first 1000 measurements of the STL decomposition.*

# E Model fitting - graphical visualisation

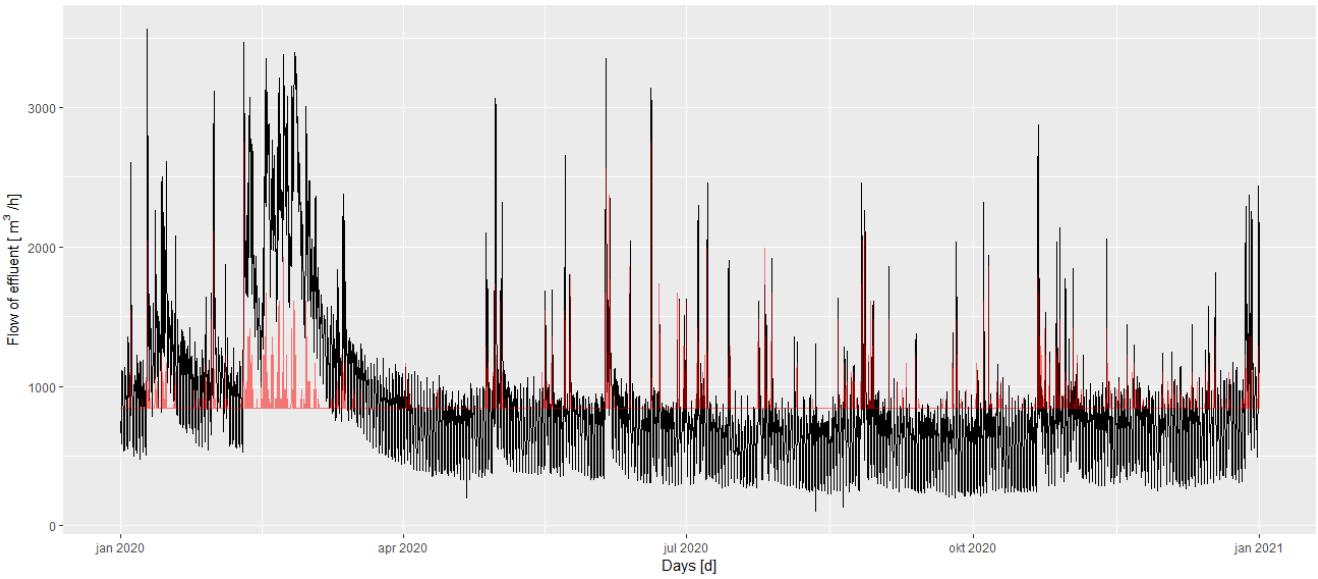


*Figure 25: The TSLM(R) plot.*

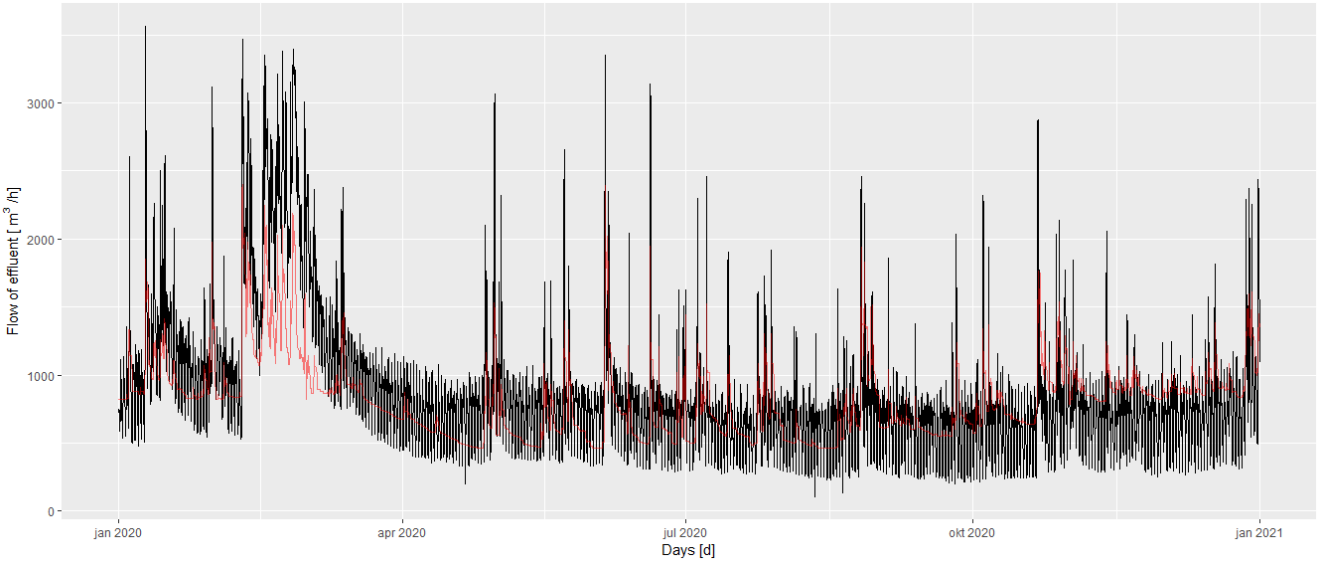


*Figure 26: The TSLM(R-R1-R2-R3-R4-R5-R6-R7-DI) plot.*

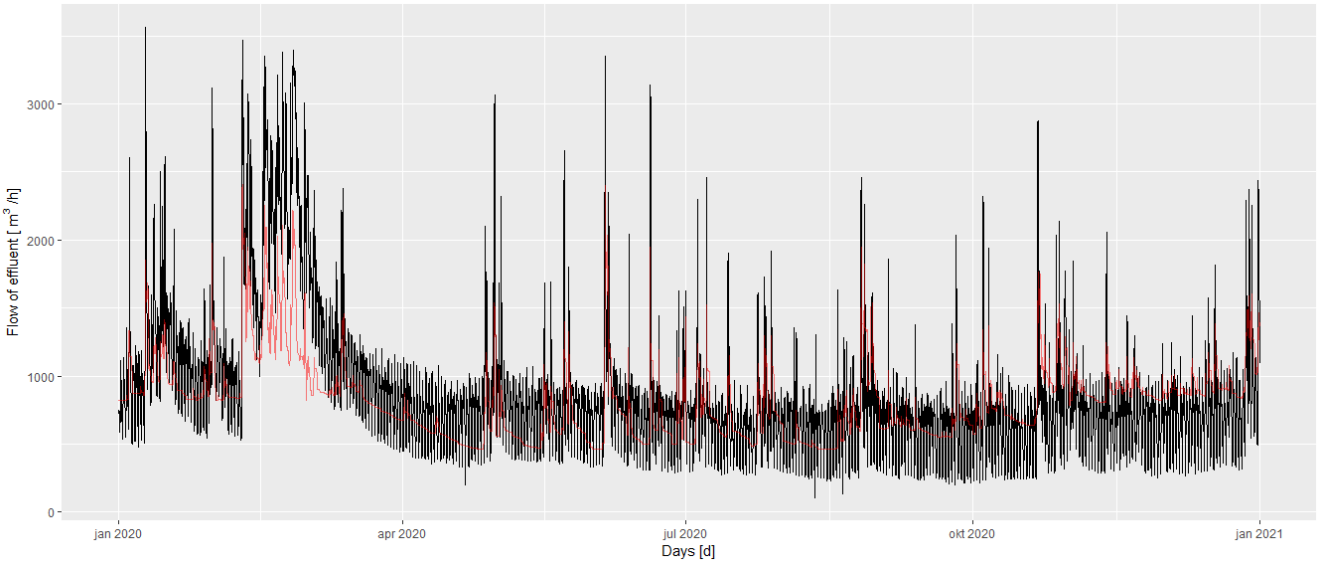


*Figure 27: The TSLM(R-R1-R4-R7-DI) plot.*

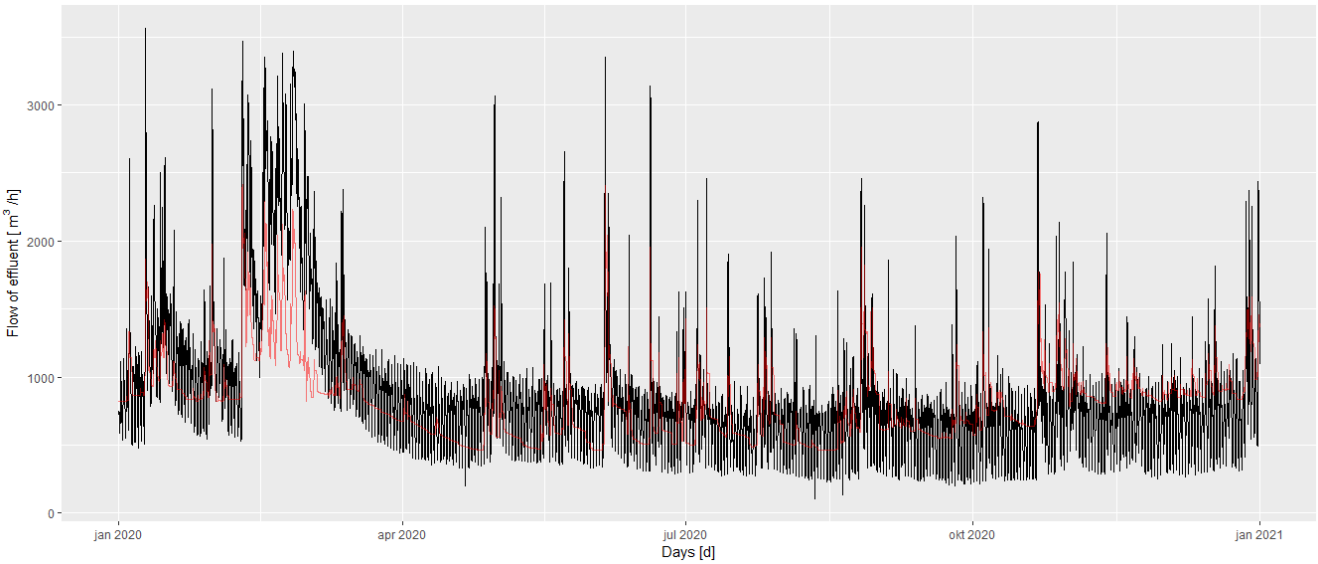


*Figure 28: The TSLM(R-R1-R7-DI) plot.*

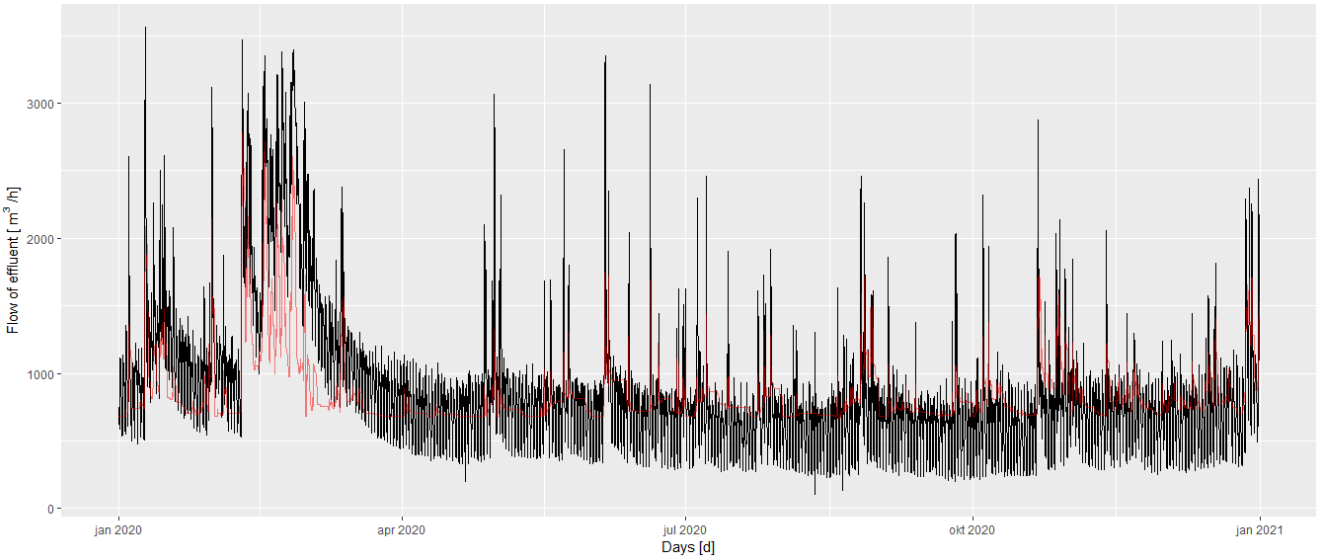


*Figure 29: The TSLM(R-R1-R7-DIxR1) plot.*

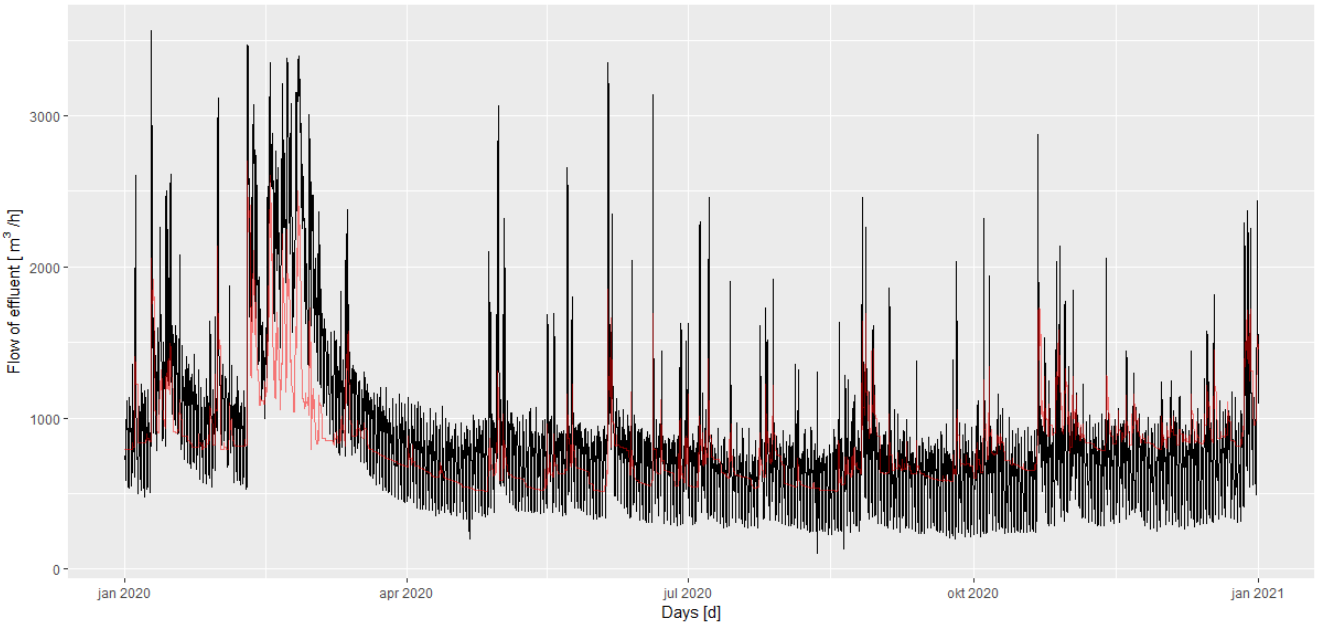


*Figure 30: The TSLM(R-R1-R7-DIxR1-DI) plot.*

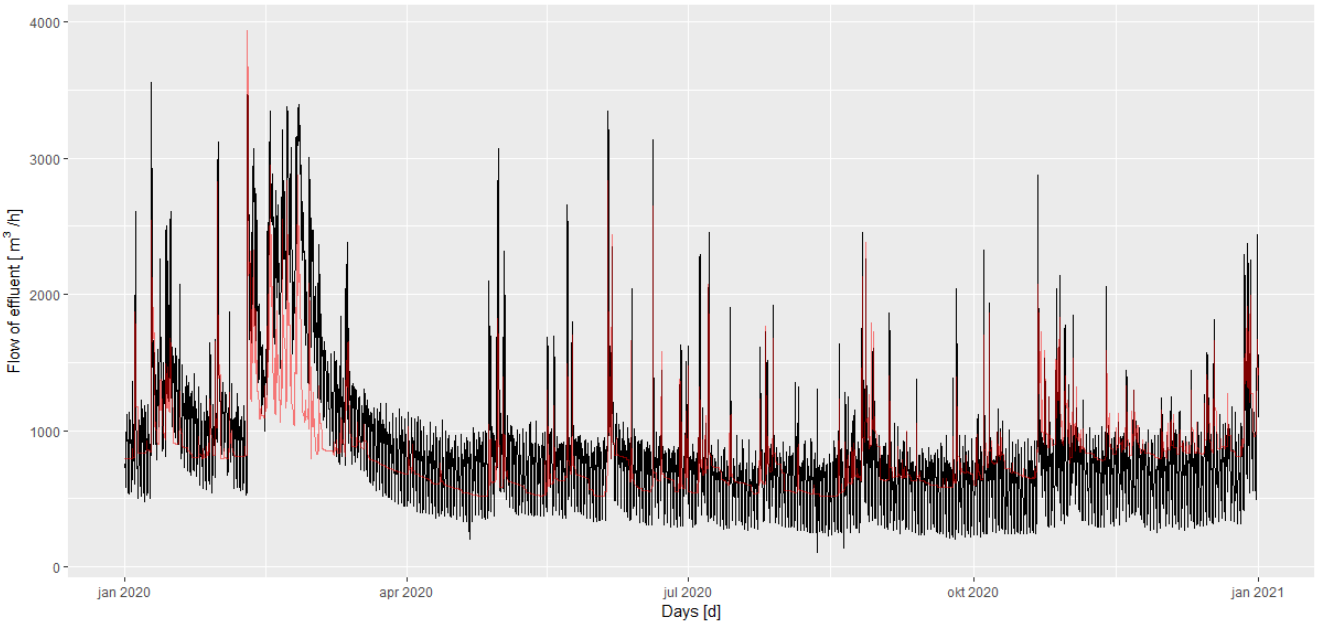


*Figure 31: The TSLM(R-lagR1-R1-R7-DIxR1-DI) plot.*

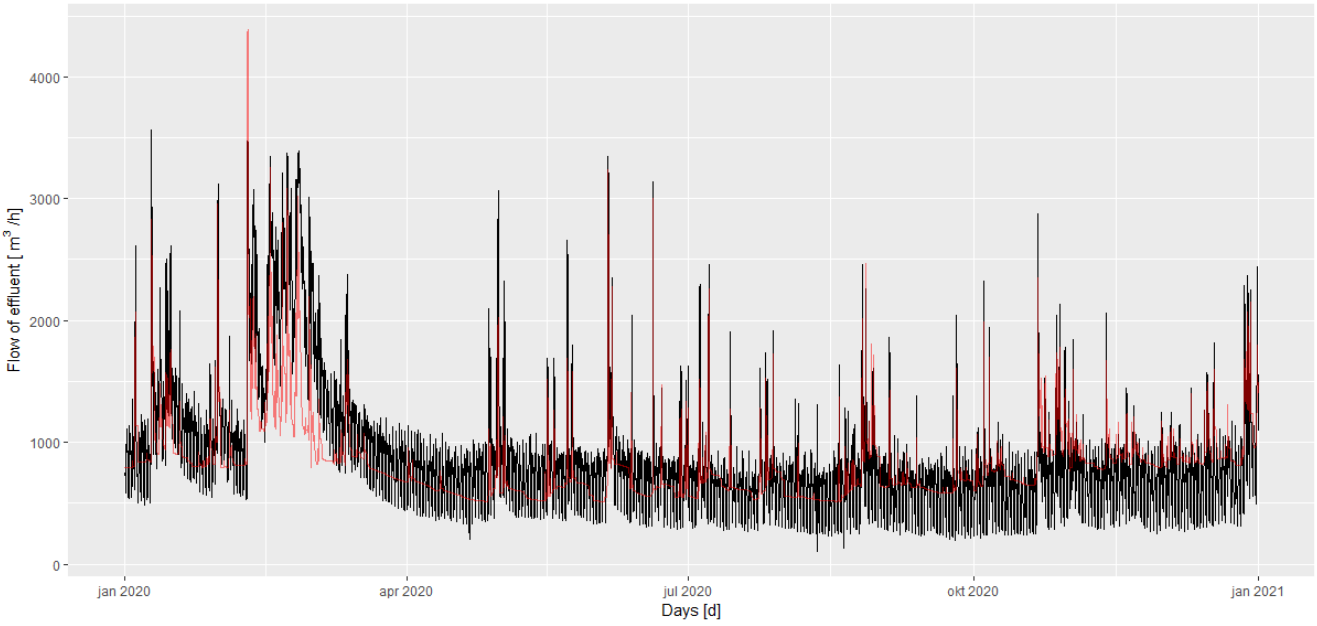


*Figure 32: The TSLM(R-lagR1-lagR2-R1-R7-DIxR1-DI) plot.*

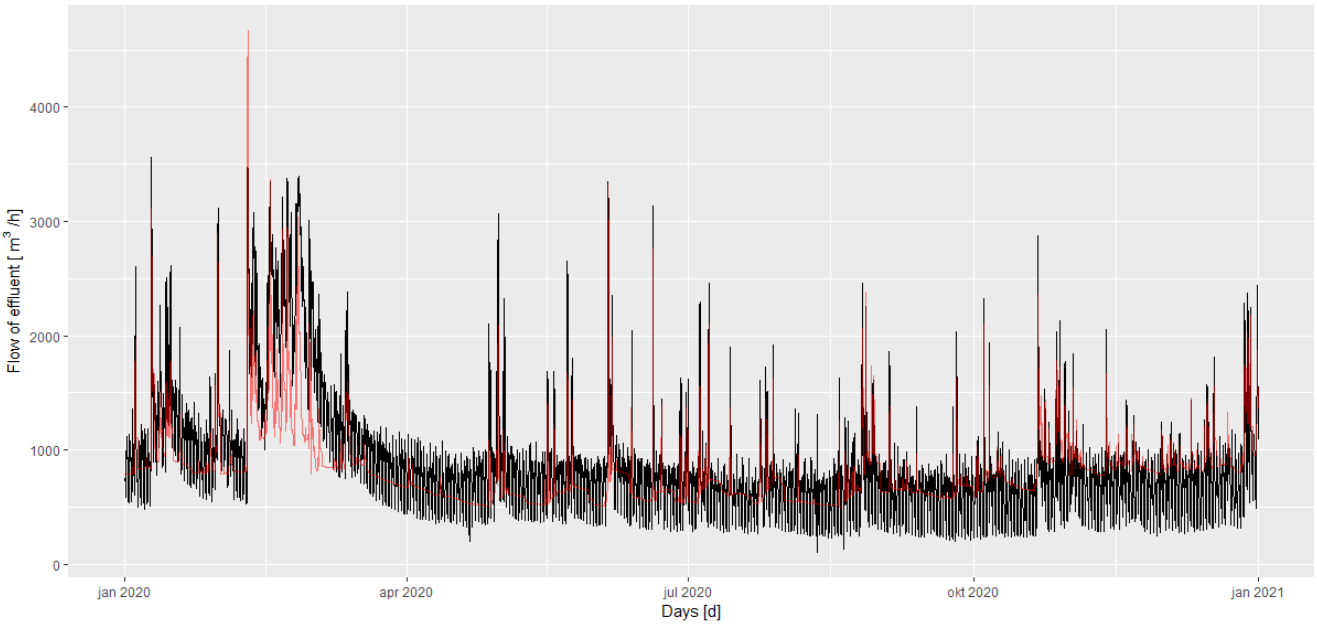


*Figure 33: The TSLM(R-lagR1-lagR2-lagR3-R1-R7-DIxR1-DI) plot.*

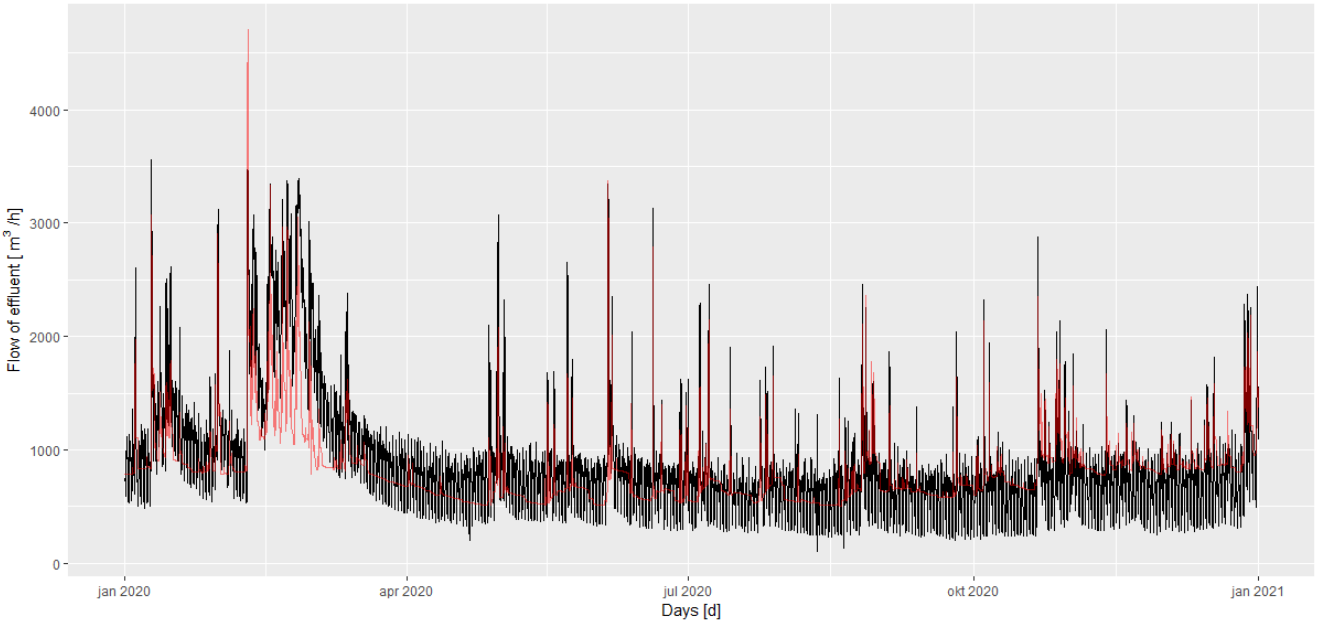


*Figure 34: The TSLM(lagR1-lagR2-lagR3-R1-R7-DIxR1-DI) plot.*

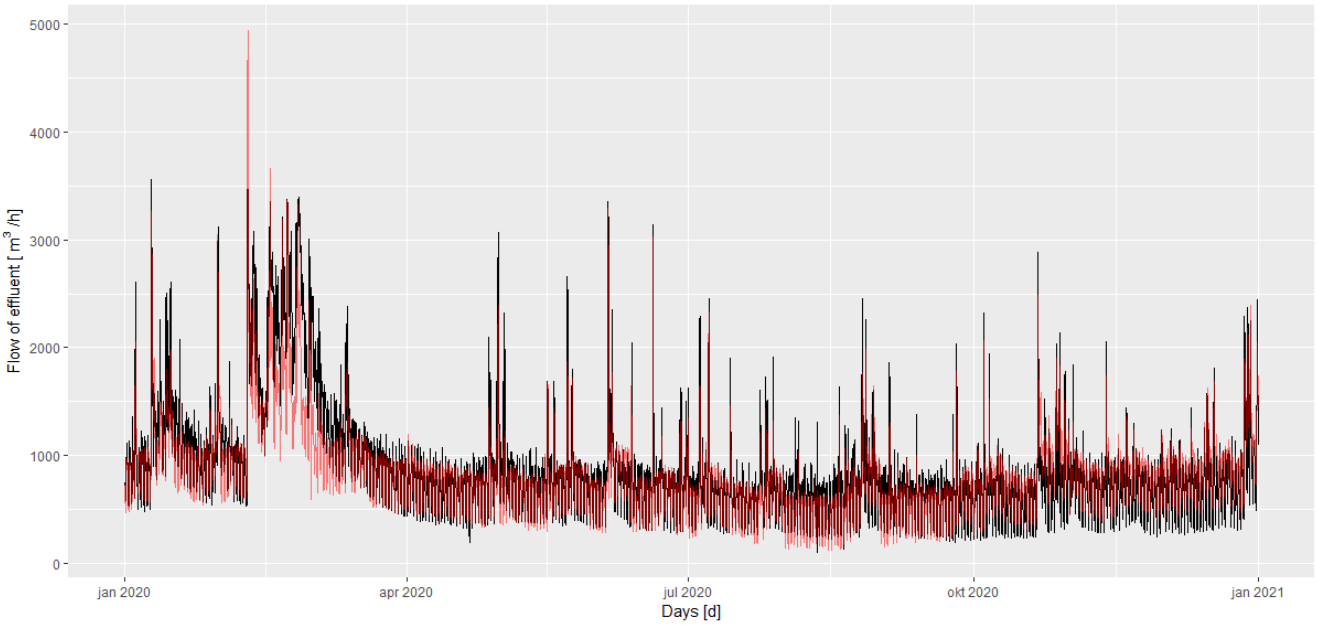


*Figure 35: The TSLM(lagR1-lagR2-lagR3-R1-R7-DIxR1-DI-FDK10-FWK5-FYK3) plot.*

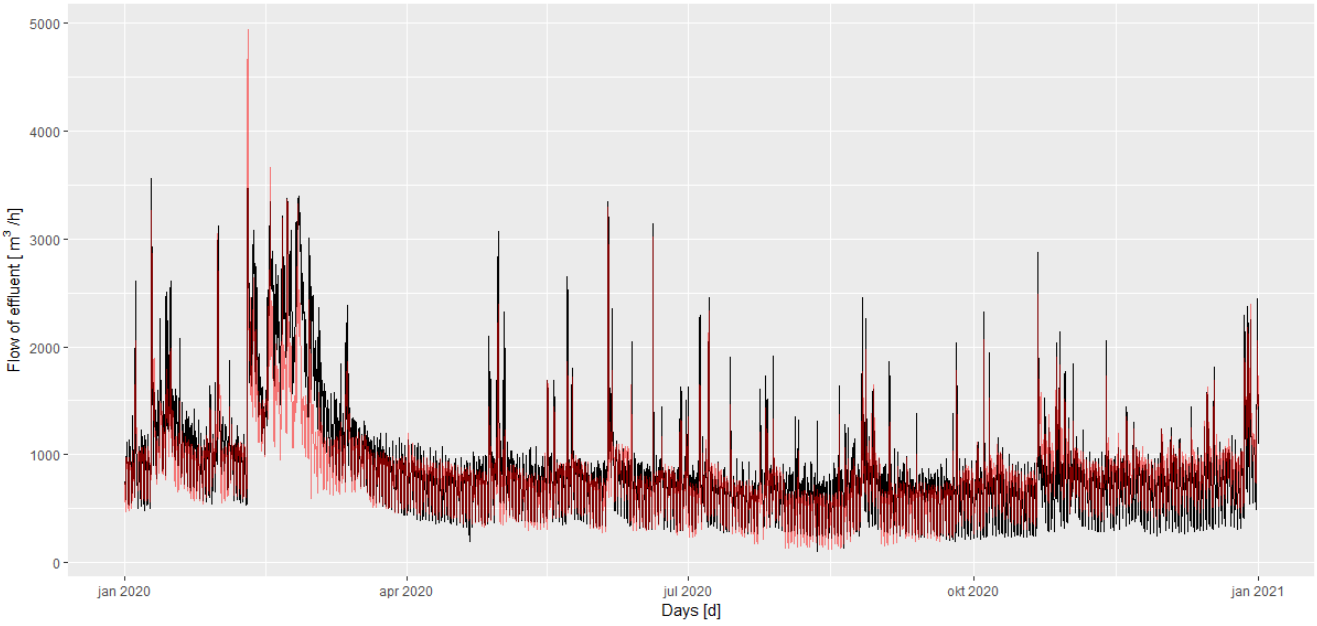


*Figure 36: The TSLM(lagR1-lagR2-lagR3-R1-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*

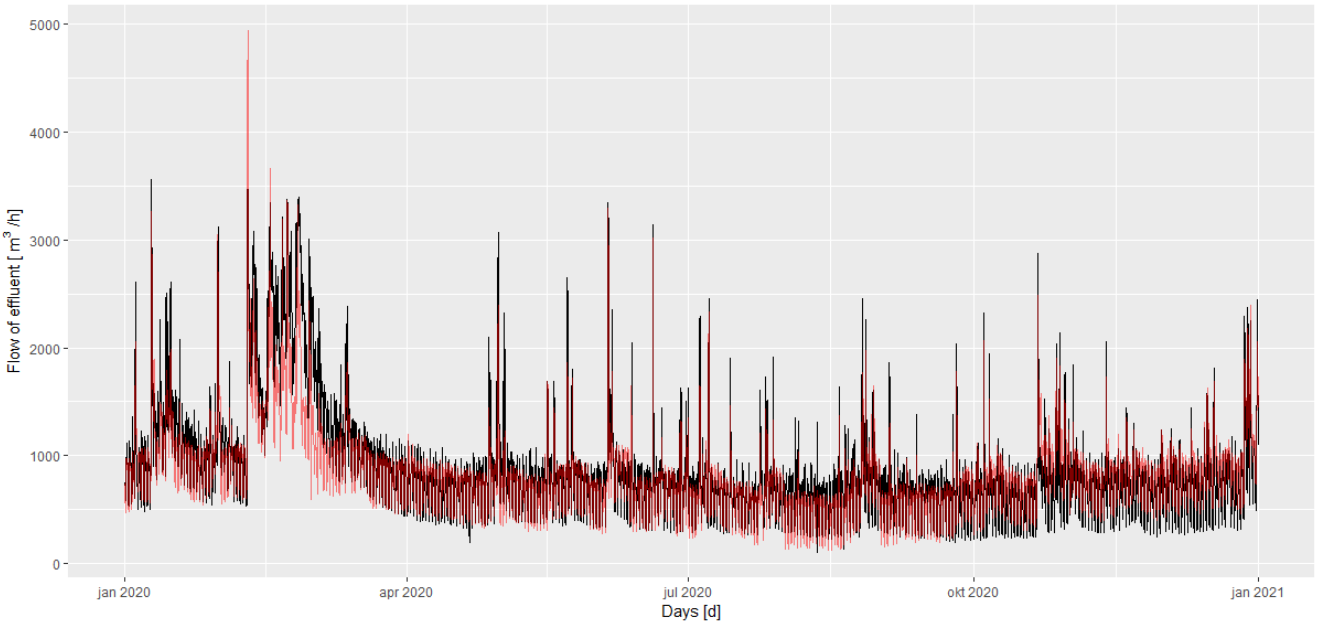


*Figure 37: The TSLM(lagR1-lagR2-lagR3-R1-R7-DIxR1-DI-FDK5-FWK3-FYK1) plot.*

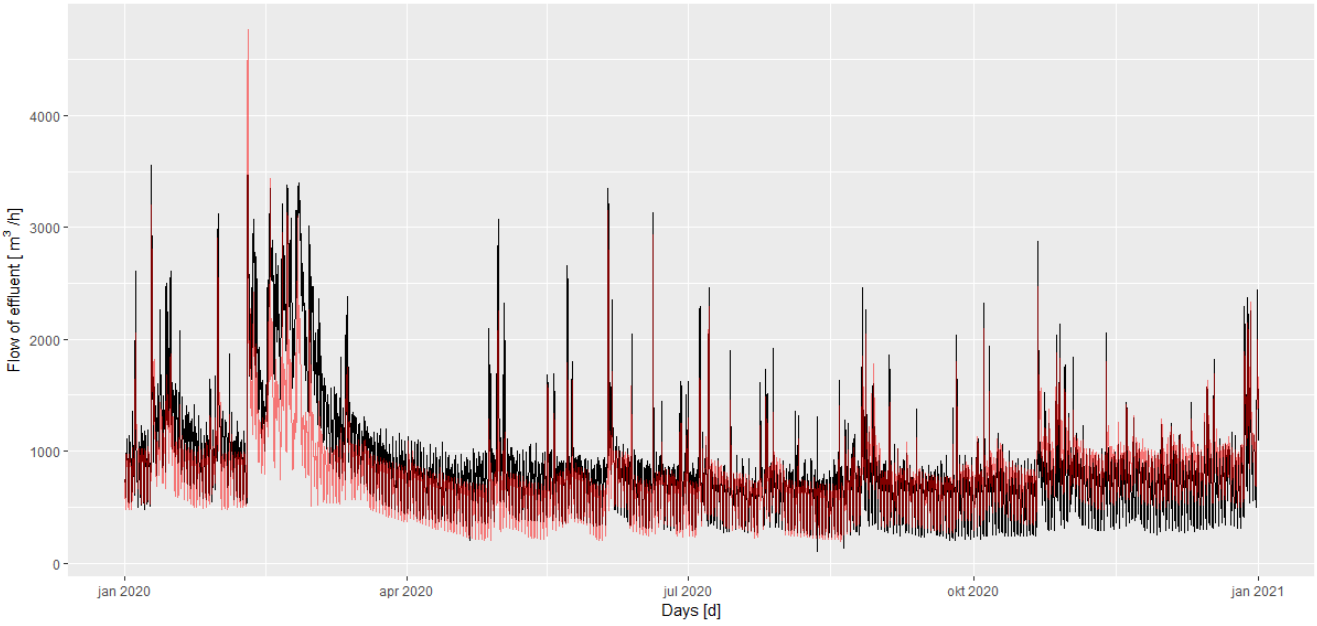


*Figure 38: The TSLM(lagR1-lagR2-lagR3-R1-R7-DIxR1-DI-FDK7) plot.*

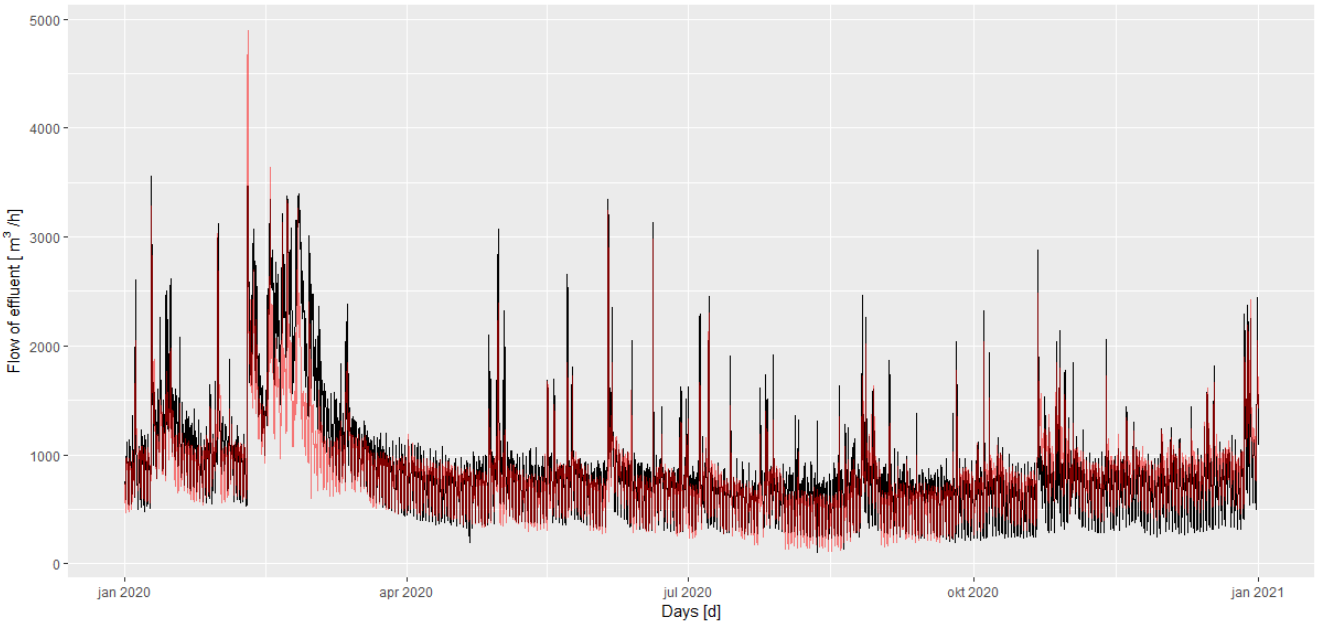


*Figure 39: The TSLM(R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*

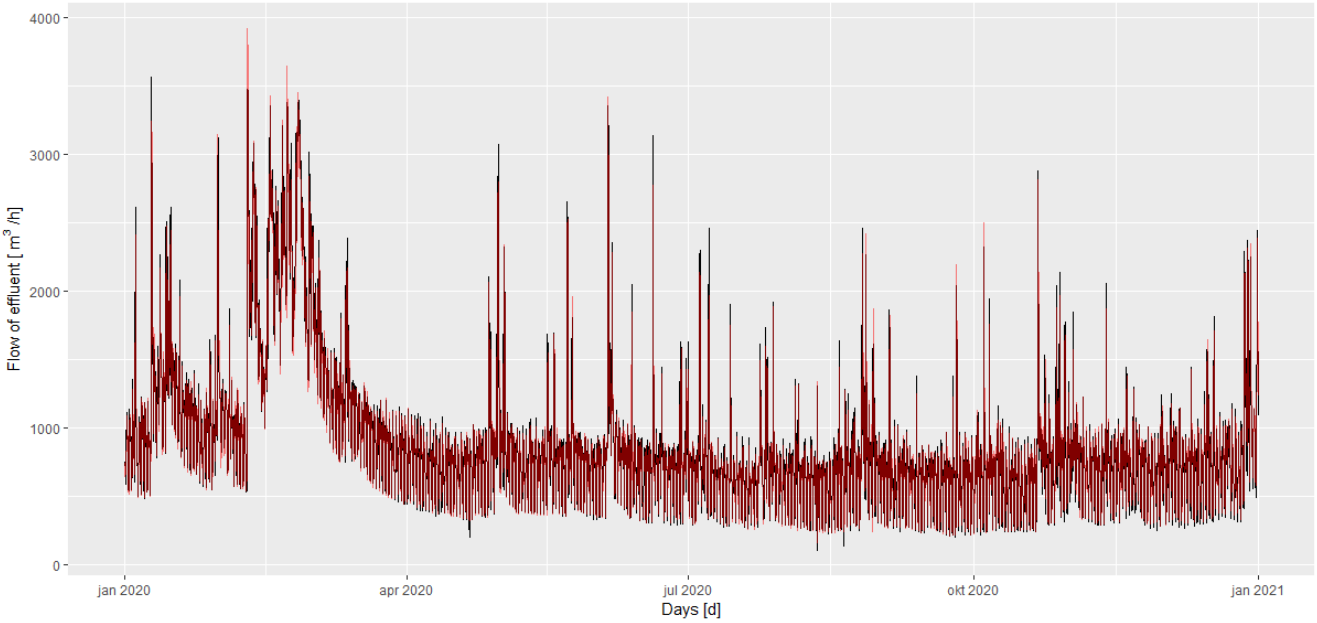


*Figure 40: The ARIMA(pdq(5,0,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*

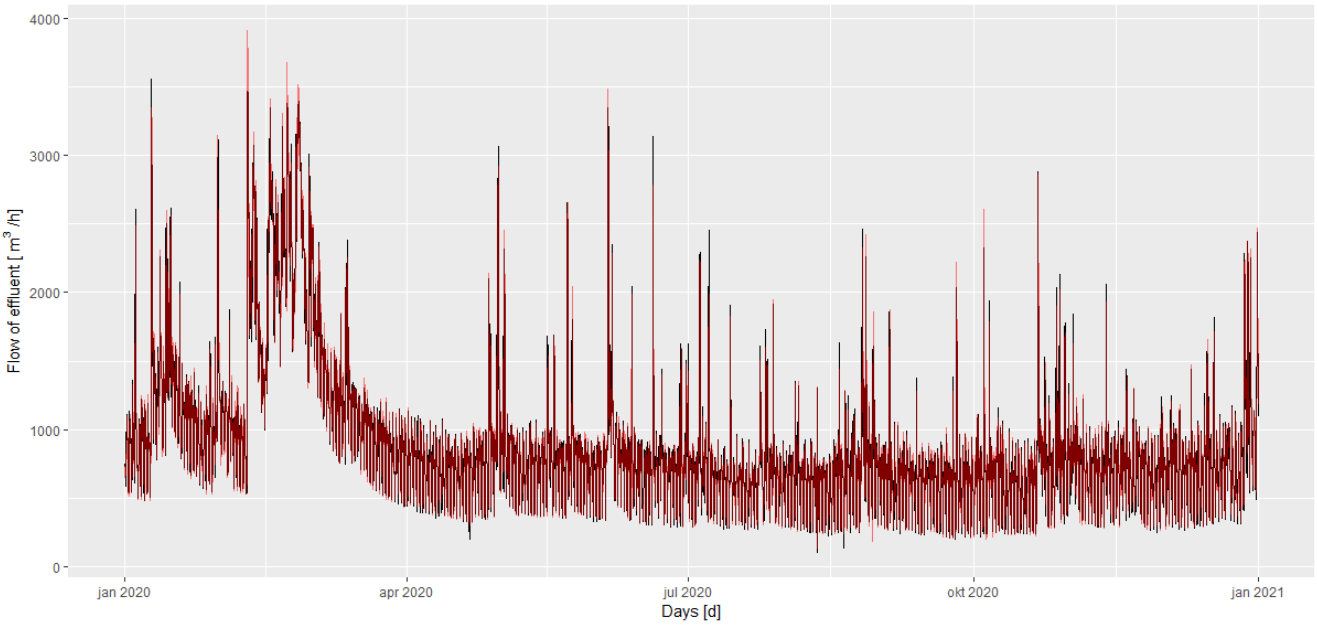


*Figure 41: The ARIMA(pdq(0,1,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*

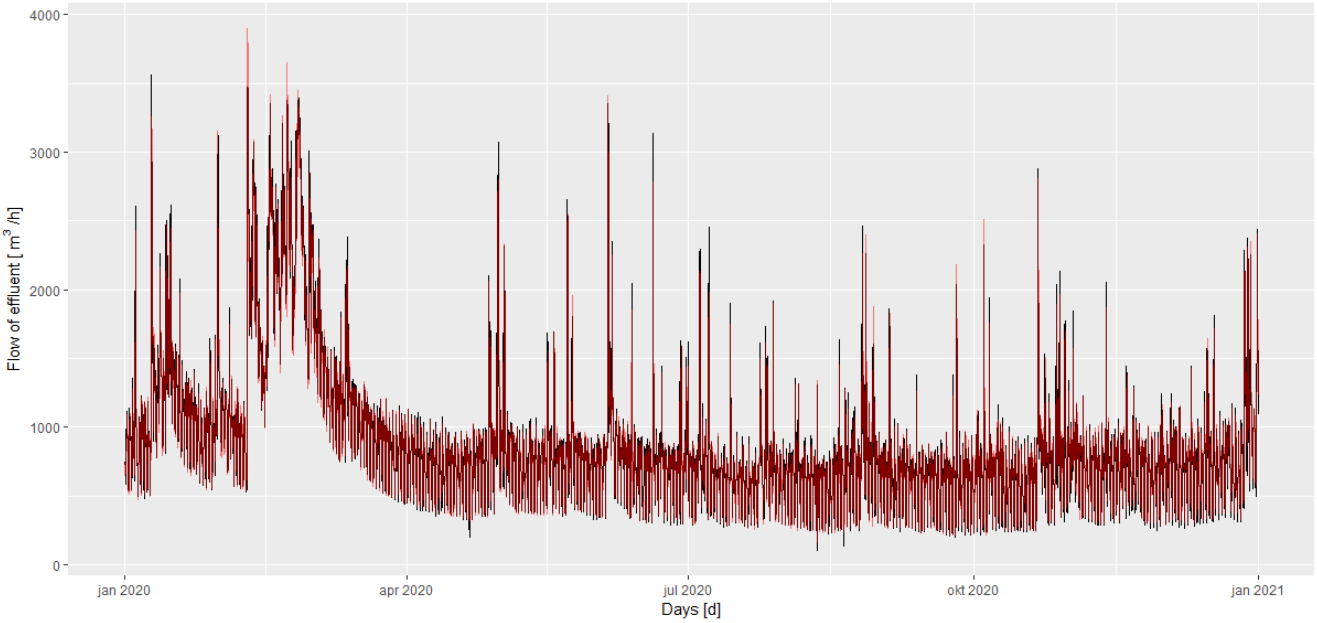


*Figure 42: The ARIMA(pdq(5,0,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*

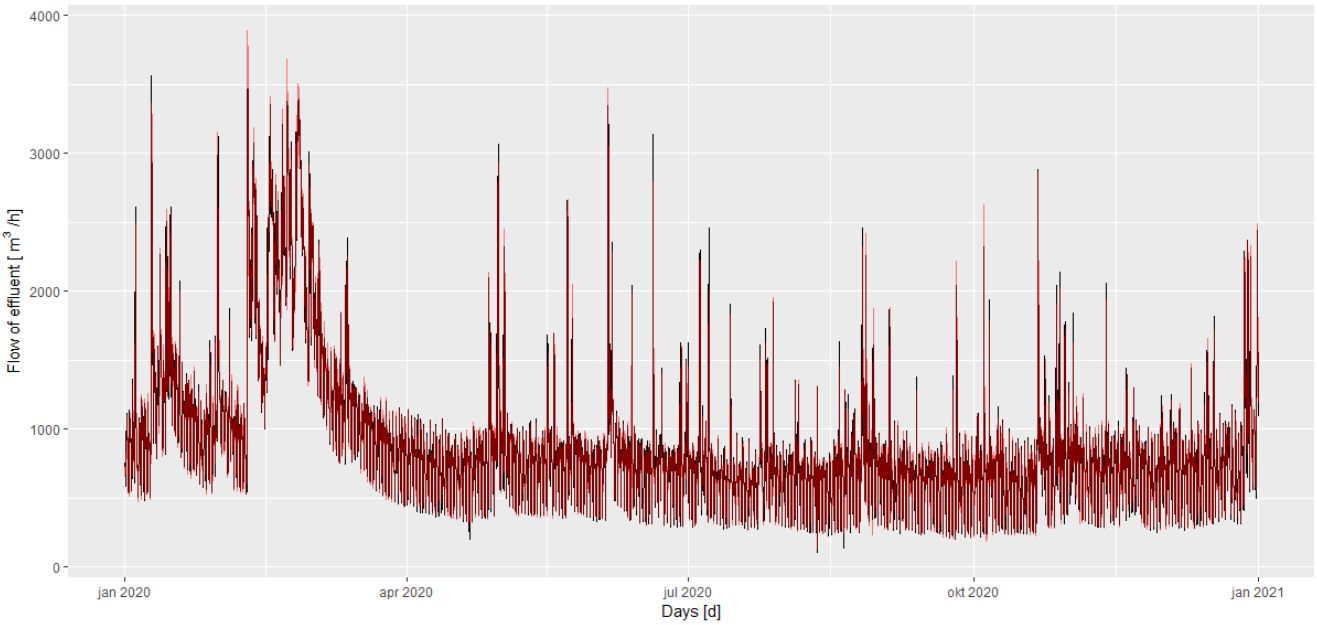


*Figure 43: The ARIMA(pdq(0,1,0)-R-lag1R-lag2R-lag3R-R1-R2-R3-R4-R5-R6-R7-DIxR1-DI-FDK7-FWK5-FYK3) plot.*