Finding ways of optimizing coagulant dosage, for a more sustainable wastewater treatment process

WATER AND ENVIRONMENTAL ENGINEERING | DEPARTMENT OF CHEMICAL ENGINEERING | LUND UNIVERSITY
HANNA SJÖMAN | MASTER THESIS 2023

Finding ways of optimizing coagulant dosage, for a more sustainable wastewater treatment process

av

Hanna Sjöman

Examensarbete nr: 2023-01

Vattenförsörjnings- och Avloppsteknik
Institutionen för kemiteknik
Lunds universitet

Maj 2023

Supervisor: **Åsa Davidsson**
Co-supervisor: **Patrik Winér**
Examiner **Michael Cimbritz**

Bild på framsidan: En AI tolkning av ”Random forest Machine Learning Used To Optimize A Wastewater treatment”. Genererad av starryai

Postadress	**Besöksadress**	**Telefon**
Box 124	Naturvetarvägen 14	046-222 82 85
221 00 Lund		046-222 00 00
Hemsida		
www.vateknik.lth.se		

Förord

Det finns ett antal människor som jag vill tacka och utan vars hjälp jag aldrig hade kunnat slutföra detta. Först och främst så vill jag rikta ett jättestort tack till min handledare Åsa och examinator Michael som båda har bidragit med såväl stöd som värdefulla råd. Jag är tämligen övertygad om att de har gjort mer än vad som har krävts av dem båda. Från Oatly är det också ett antal människor som jag vill tacka och allra mest min handledare på företaget, Patrik. Tack för att du har gett mig friheten att få chansen att forma detta projekt till mitt eget samtidigt som du alltid har funnits som stöd längst vägen. Även ett stort tack till alla som jobbar på reningsverket, tack för att ni har besvarat alla mina frågor och har orkat lyssna på mig när jag har visat en massa oklara grafer.

Utöver alla personer som på olika sätt har hjälpt till att göra innehållet i denna rapport lite bättre så finns det oändligt många människor som jag vill tacka för att ha gjort min studietid till det den har varit. Personer som funnits där i ettan när jag satt varje torsdagskväll och var övertygad om att jag borde hoppa av utbildningen eftersom jag inte fick skriptet i Matlab att fungera. Personer som jag hängt med under bord på fester. Personer som jag har sprungit runt rondeller med. Personer som jag har kört snurra flaskan tillsammans med. Personer som jag har burit fanor tillsammans med. Personer som jag har cyklat runt delphi iförd mantel tillsammans med. Personer som jag har städat gasqueköket tillsammans med. Personer som jag har druckit vin och snackat skit tillsammans med. Personer som jag har diskuterat Brunnshög tillsammans med. Personer som jag har haft sittning tillsammans med i en bastu. Personer som jag har bott och kollat på legally blond tillsammans med. Personer som jag har ätit Max på KC tillsammans med. Personer som jag har suttit på AF-borgens tak tillsammans med. Personer som jag har kört sopkärl tillsammans med. Personer som jag har gjort en karneval tillsammans med. Tack.

Slutligen, att komma i tid har aldrig varit min starka sida så det känns bara passande att jag inte heller tog examen i tid men som en vis man sa första skolveckan när jag gick i ettan,

Detta kan antingen vara de fem värsta åren i era liv eller de sex roligaste.

Hanna Sjöman

Lund, 14 april 2023

Hur maskininlärning kan hjälpa dig att klara tentan och beräkna hur mycket kemikalie som behövs!

Maskininlärning kan liknas vid en student som läser en väldigt stor kurs med ett väldigt brett innehåll. När tentan börjar närma sig så inser studenten att det är omöjligt att tentaplugga genom att lära sig allt som står i kurslitteraturen. Så istället fokuserar hen på att göra extentor för att lära dig svara på de frågor som brukar komma på tentan. För varje extenta så svarar studenten på frågorna, jämför med facit och justerar sitt resonerande därefter. Målet för både studenten och med maskininlärning är att ha tillräckligt bra koll på hur man bör svara på en viss typ av fråga att på tentan kunna applicera det på frågor som man inte har övat på.

Maskininlärning är inte bara användbart under studietiden utan går även att applicera i verkliga livet, exempelvis i en reningsprocess av avloppsvatten. Utanför Landskrona har Oatly en fabrik där man tillverkar havreprodukter och avloppsvattnet från den produktionen renar man i ett tillhörande vattenreningsverk. Reningsprocessen består av ett antal olika steg varav ett är ett kemiskt reningssteg där man använder en fällningskemikalie för att rena vattnet från bland annat organiskt material, COD. För att uppskatta mängden fällningskemikalie som man behöver tillsätta så genomför man ett antal testar där man tar prover, tillsätter olika mängder kemikalie och ser vad som verkar vara en tillräcklig mängd. Problemet är att varje test tar ganska lång tid och när mängden organiskt material i avloppsvattnet eller andra faktorer ändras så måste man göra nya tester. Oatlys avloppsvatten varierar väldigt mycket när det kommer till organiskt material, temperatur, pH och annat. Det innebär att väldigt många justeringar i mängden fällningskemikalie man tillsätter krävs och därför också väldigt många tester.

Man renar också vattnet mer än vad som behövs enligt sitt miljötillstånd och man vill därför se över om det går att minska mängden kemikalie.

Hur effektiv en fällningskemikalie är beror på väldigt många olika faktorer och det är därför svårt att få en överblick, kurslitteraturen är enormt bred och komplicerad. Däremot så kan man använda maskininlärning, mer specifikt, Random Forest, för att försöka hitta ett mönster i sammansättningen av avloppsvattnet som kommer in, mängden fällningskemikalie man tillsätter och hur rent vattnet ut blir. På så sätt så får man både ett verktyg som de som jobbar kan använda för att förhoppningsvis slippa göra så många tester och man kan också undersöka hur rent vattnet beräknas bli vid olika doseringar.

Resultatet är någonting som kan användas för att göra miljön en tjänst samtidigt som man sparar pengar, och det enda som krävs är att låta en dator göra jobbet.

Summary

The wastewater treatment process at Oatly consists of several treatment steps of which, two are chemical treatment steps involving coagulation followed by flotation to remove phosphorous and COD.
The coagulant dosage is mainly determined based on the operator's experience and supported using results from jar tests. Jar tests are, however, highly dependent on the wastewater quality parameters making it difficult to apply the result if there are large variations in the composition of the wastewater, which is the case at Oatly. As wastewater parameters such as concentration of COD and Tot-P in the primary influent, pH, temperature, and flow rate all varies multiple jar tests have to be conducted daily and the coagulant dosage is often adjusted. Despite this, the BOD_7 concentration of the final effluent is well within the concentration limits of the environmental permit. This in combination with the large amounts of coagulant used in the process has raised the question, is it possible to decrease the dosage and still fulfill the requirements of the permit?
The large variations of the primary influent make this difficult to achieve as there is no simple linear relationship between the amount of coagulant used and the coagulation efficiency, instead, it is influenced by several parameters.

Instead, machine learning is used where a random forest regressor algorithm is used to predict the concentration of COD in the effluent based on the flow rate, pH, temperature, PAC dosage, the influent concentration of COD, and Tot-P. To train the model data from the first chemical treatment step is used as more data is available compared to the second chemical treatment step. After training the model it is used to predict the COD concentrations of the effluent from the first flotation.

The results show an r^2 value of 0.85 and an RMSE value of approximately 49, indicating a strong relationship between the independent and dependent variables used in the model. The RMSE value on the other hand should preferably have been lowered, it means that the difference between the concentration predicted by the model and the real concentrations is 49 mg/L on average.

To illustrate the possible benefits of the model it is used to estimate the COD concentration based on test data where the dosage of PAC is varied. The predicted COD concentrations are then compared to the COD concentration of the effluent with a 57 % removal efficiency. A 57% removal efficiency was used to determine what would have been a sufficient amount of COD in the effluent as this is the removal efficiency that was used when the WWTP was designed.
The model predictions showed that it, according to the model, would have been possible to decrease the coagulant dosage and save around 280 kg of coagulant/day on average. The carbon footprint of PAC is approximately 0.54 kg CO_2-eq/kg PAC so this would mean a decrease in CO_2 emission by around 150 kg CO_2-eq per day. Decreasing the dosage would also make it possible to save approximately 1000 SEK.

The main suggestion for areas of future work is to increase the size of the dataset to improve the performance of the model. However, the model shows potential as a tool to predict the COD concentrations of the effluent and could be used to estimate the effect of different coagulant dosages, thereby opening up the possibility of optimizing the coagulant dosage.

Table of content

1 Introduction

One of Sweden's environmental objectives is to achieve a reduced climate impact by reducing emissions of carbon dioxide and other greenhouse gases to achieve zero net emissions of greenhouse gases by 2045 (Naturvårdsverket, n.d.). It is approximated that around one-third of the emissions in Sweden are from industries which is 24 % less than in 1990 (Naturvårdsverket, n.d.). Poly-aluminum chloride (PAC) is a coagulant commonly used in wastewater treatment processes and is used in Oatly's wastewater treatment (WWT) mainly for the removal of phosphorous and colloidal particles. The carbon footprint of PAC is 537 kg of CO_2-eq per ton of product. For the WWT at Oatly, approximately 52 tons of PAC are used per month which, means that just the usage of coagulant in the process is responsible for around 28 000 kg CO_2-eq emissions each month. In comparison, a pickup truck is estimated to emit approximately 3510 kg CO_2/year or 290 kg CO_2/month (Armstrong, 2022). This means that the amount of coagulant used each month is equal to the emissions from almost 96 pickup trucks. Although this number might seem high it is important to note that during 2021 the total emissions in Sweden from industries were 15.7 million tons of CO_2-eq.

One possible way of decreasing the amount of emission is of course to ensure that the coagulant dosage used is sufficient for the treatment purpose without being excessive. Oatly's environmental permit regulates the requirements that must be met by the process, for example, the maximum concentration of BOD_7 in the treated water. According to Oatly's environmental permit, the maximum concentration of BOD_7 in the final effluent is 8 mg/L calculated over 6 months. Currently, the concentration is 4 mg/L indicating that it might be possible to decrease the coagulant dosage, causing higher concentrations in the effluent, while still meeting the demands of the permit.

The most common approach for determining coagulant dosages is to perform jar tests to determine how much coagulant is required; however, these results may be difficult to apply because it is highly dependent on the wastewater quality parameters of the sample. At Oatly there are large variations in the composition of the primary influent, which means that a large number of jar tests and adjustments in coagulant dosage are required. It is therefore difficult to systematically determine what could be a sufficient dosage to achieve the required results while minimizing the amount of coagulant used in addition to being very time-consuming (Heddam, et al., 2011).

A more novel approach is to use machine learning to train a system to predict concentrations in the effluent based on historical data. By learning from previous data, the system has the potential to make accurate predictions even with varying wastewater parameters. Research also indicates that it could provide a possible substitute for jar tests thereby also decreasing the workload for the operators (Zhang et al., 2013).

Although the concentration of BOD_7 in the final effluent is the parameter regulated in the permit, it can be substituted by measuring the chemical oxygen demand (COD) as there is a strong. correlation between BOD_7 and COD (Hu and Grasso, 2005). Due to this, and the fact that measuring COD is much faster, Oatly has measured the concentrations of COD at multiple points in the process during a period of approximately 5 months. This data, in combination with data from several online sensors, makes it possible to investigate the possibility of using machine learning to predict the concentration of COD in the effluent of

the first chemical treatment step. Using the coagulant dosage as a parameter in this model also opens up the possibility of estimating what could be a sufficient amount that results in a low enough concentration of COD in the effluent. The ambition is that by continuing this work it could eventually be possible to make real-time adjustments to the dosage to ensure that just enough coagulant is used and not too much.

1.1 Aim and research questions

The overall aim of the project is to investigate ways of possibly decreasing the amount of coagulant used in the WWT at Oatly. To achieve this machine learning will be used to try to create a model that could be used for this purpose.

Based on this overall aim, the research questions that the project aims to answer are:

- Is it possible to use the data available to create a model that can be used to predict COD concentrations in the effluent?
- How could such a model be used and what could be the possible consequences from an environmental and economical perspective?
- What is needed to improve the performance of the model?

1.2 Project boundaries

The model is only based on data from the first chemical treatment step. This is motivated by the data available, and the choice is described in greater detail in section *4. The dataset*. The focus of the project is to investigate the possibility of creating a model and highlight its possibilities, but validating the results by applying them to the actual process is considered out of scope.

The task was only focused on the concentration of BOD_7 in the effluent and so all evaluations regarding coagulation efficiencies and so on, are only done in regards to the concentrations of BOD_7 and by extension, COD.

Finally, as the aim was to investigate the possibility of creating a model, there has been no systematic comparison of different algorithms and their strengths. The choice has been made based on conclusions from previous research, but no attempt has been made to evaluate which algorithm that is best suited nor has any optimization of the algorithm been attempted. The reason for this is that the main limitation of the model is believed to be the small size of the dataset which could make the effect of any optimization negligible.

2 Background

2.1 The production at Oatly

Just outside of Landskrona Oatly has a large factory used for the production of the company's different oat-based products. Oatly's products are produced from oat kernels which are peeled and heat-treated before arriving at the factory. At the factory, the oat kernels are milled and mixed with water before natural enzymes are added for the hydrolysis of the starch to create the oat base. After the separation of insoluble oat fibers, the main components of the oat base are beta-glucans, maltose, protein, fat, and carbohydrates. This oat base is the base of all of Oatly's products and different ingredients are added to form the different products before the solution is pasteurized or treated using ultra-heat treatment. Finally, the product is homogenized to break down large fat droplets before the product is packed and transported to wholesale companies (Oatly, n.d.).

The production generates wastewater with a high content of organic material, nitrogen, phosphorous, and suspended particles mainly. There have also been occasional discharges of for example large quantities of oil due to disturbances in production.

2.2 The wastewater treatment process at Oatly

To take care of the wastewater generated from the production Oatly built their wastewater treatment plant which began operating at the beginning of 2021 and is connected to the recipient Lundåkrabassängen. The major steps of the process are shown in Figure 1 and as can be seen, it consists of mechanical, chemical, and biological treatment steps.

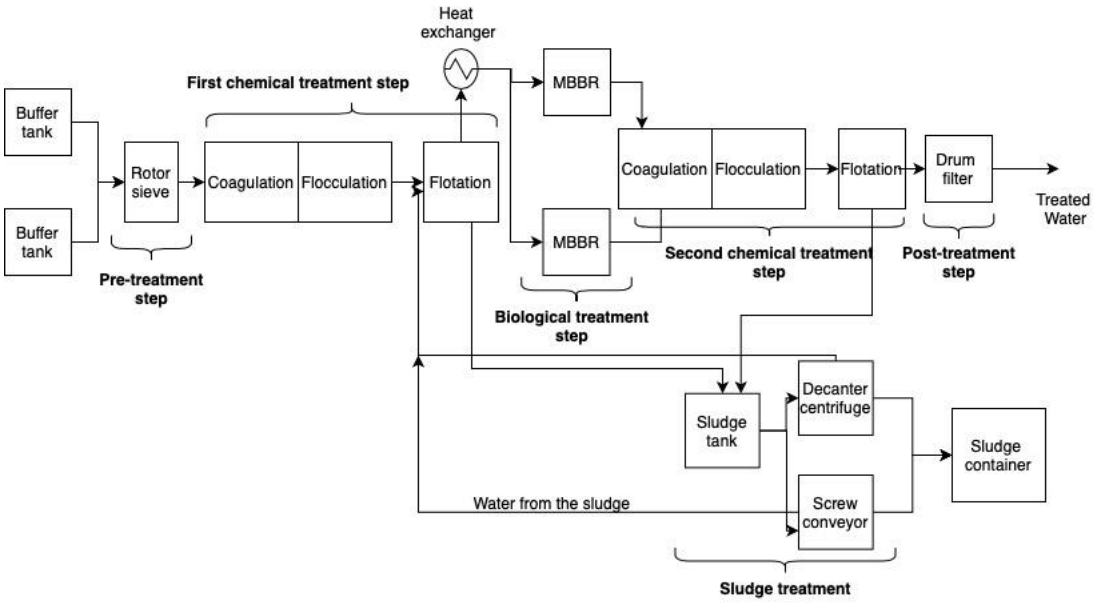

Figure 1 The major steps of the WWT at Oatly.

2.2.1 Before entering the WWT

The wastewater from the factory is short-time stored in two buffer tanks with a total capacity of 800 m^3 before entering the treatment process. The buffer tanks are usually filled to about half capacity to ensure that there is a possibility of negating possible peaks in temperature, pH, and concentration of containments using dilution. Both tanks are equipped with stirrers as well as on-line monitors measuring the temperature and pH in the tanks. There are also monitors to measure the flow rate, pH, and turbidity as the wastewater enters the plant. The influent samples are taken automatically and sent for analysis at an accredited laboratory. The samples are flow proportional to ensure that the concentrations are representative of the average concentration for 24 hours.

2.2.2 The pretreatment step

During the process's pretreatment step, large solids, for example, residual oats, are separated from the influent using a rotary drum screen. The screen has 3 mm perforations and an integrated screw press that dewaters the separated solids before being sent away for incineration.

2.2.3 Pre-precipitation

The chemical treatment step consists of coagulation followed by flotation for the removal of phosphorous and colloidal particles, thereby removing COD. The reaction tank consists of two areas separated by a partial divider, both of which are equipped with stirrers operating at different conditions. In the first area of the tank sulphuric acid (H_2SO_4) is used to adjust the pH to approximately 5 to enable the removal of oil from the wastewater. The acid reacts with the oil through emulsion breakage, meaning that it reacts with the oil's carboxylic groups forming carboxylic acid causing neutralization of the oil's stabilizing factors and thereby facilitating coagulation-flocculation (Kemmer and Ill, 1988). In the same area, the coagulant, poly-aluminum chloride (PAC) is added during turbulent conditions caused by the mixer operating at 1465 rpm.

In the second area of the tank flocculation takes place and to avoid breakage of the flocs formed, the stirrer is set to operate at 1390 rpm to ensure less turbulent conditions. The addition of the coagulant causes the pH to decrease and to negate the effect sodium hydroxide (NaOH) is added to increase the pH to approximately 6.5 to 7.5. The amounts of H_2SO_4 and PAC that are added for the coagulation are proportional to the flow rate but the dosage of NaOH is automatically adjusted to achieve a set pH.

After the flocculation, an anionic polymer solution is added to the wastewater to strengthen the formed flocs just before it passes through a static mixer and enters the flotation tank. For separation of the flocs dissolved air flotation (DAF) is used using an IdraFlot IFS 80 which has a volume of 79.5 m^3 and a flotation area of 83 m^2. Along the entire length of the tank, there are openings where the influent can enter after being mixed with a mixture of pressurized air (around 5 bar) and dispersion water. Once inside the tank, the pressurized air causes bubbles to form which the flocs will adsorb to and then float to the surface of the tank. At the top of the IdraFlot, the sludge produced is scraped off the surface using a drag skimmer. The produced sludge is pumped into a sludge tank before being dewatered and compressed. During this process, the secondary effluent, water separated from the sludge, is produced and then mixed with the pressurized air before being mixed with the influent.

2.2.4 The biological treatment step

The biological treatment step consists of two parallel Moving Bed Biofilm Reactors (MBBR) where nitrogen and the majority of the BOD_7 are removed from the wastewater through pre-denitrification and two BOD-reducing steps taking place in different areas of the reactors. Each reactor has a volume of approximately 400 m^3 divided into three chambers with different operating conditions adjusted for the biological process taking place. The first chamber is 60 m^3 and where denitrification takes place, BOD reduction occurs in two steps in two different parts of the reactor. The volume for the first part of the BOD reduction is 230 m^3 and for the second 110 m^3. All chambers contain biofilm carriers, AnoxK™ Z-400 biofilm-carriers, where the microbes and bacteria grow and are kept suspended in the reactor. The approximate residence time of the reactors is 15 hours according to the design specifications (VA Ingenjörerna, 2022).

The wastewater enters the reactors at the pre-denitrification stage where nitrite is converted into nitrogen gas under anoxic conditions using BOD as an energy source. Due to the anoxic conditions, there is no air diffuser to keep the biofilm carriers suspended, instead, there is a mixer to ensure this and to facilitate good transportation of the nutrients to the biofilm (Veolia Water Technologies AB, 2021). The mixing also causes turbulent conditions which help keep the biofilm from growing too thick on the carriers which can hinder the transportation of substrates to the surface of the biofilm (Barwal and Chaudhary, 2014). Unlike the chamber for pre-denitrification, in the areas for BOD reduction, there is also an air diffuser connected to the bottom of the tank to enable aerobic conditions. An on-line monitor measures the pH, temperature, and concentration of dissolved oxygen which is automatically adjusted using an air diffuser to ensure that the concentration is approximately 2 to 4 mg/L. In the two BOD-reduction steps, BOD is reduced to form mainly CO_2 and water and during this reduction, oxygen is consumed (Veolia Water Technologies AB, 2021).

Important operating parameters for the bio-treatment are temperature and the influent concentration of phosphate. Before entering the bioreactors the temperature of the influent is adjusted to between 26 °C and 28 °C and it is also possible to increase the concentration of phosphate by adding phosphoric acid (H_3PO_4) in connection to the entrance of the reactor. The adjustment of temperature is required as it controls the growth and activity of the microbes with a temperature of around 27°C found to be optimal for COD removal according to one study (Madan et al., 2022; Majid, 2019). Phosphate is required for the growth and reproduction of the microbes and a rule of thumb is that an appropriate ratio of C:N:P in the wastewater is approximately 100:5:1 for aerobic treatment (Ammary, 2004). The dosage of H_3PO_4 is set to be flow-proportional and adjustments to the amount added have been made according to the BOD loading rate over time (Veolia Water Technologies AB, 2020).

2.2.5 Post-percipitation

Following the bio-step is a second chemical treatment step consisting of coagulation followed by flotation used for the removal of phosphorous and surplus biomass from the MBBR. The design of the reaction tank is similar to the first chemical treatment step, meaning that it consists of two parts divided by a partial divider. PAC is used as a coagulant and the stirrer in the coagulation area operates at 1465 rpm. To counteract the pH lowering caused by the coagulant, NaOH is added to increase it to approximately pH 7 to 8. The pH is chosen to ensure that the recipient fulfills the pH requirement of the environmental permit. After the pH adjustment, the

wastewater enters the flocculation part where flocs are formed under less turbulent conditions accomplished by a stirrer operating at 1390 rpm.

To strengthen the flocs an anionic polymer is added as the wastewater moves through a static mixer before entering the flotation tank. The model of the flotation tank is the same as that used for the first flotation, an IdraFlot 80 with a volume of 79.5 m^3 and a flotation area of 83 m^2. The flocs are scraped off the surface of the tank using a drag skimmer and the sludge is transported to the same sludge tank.

2.2.6 The post-treatment step

The effluent from the second chemical treatment step carries on to a drum filter for after-treatment before reaching the recipient Lundåkrabassängen. In connection to the drum filter is a coagulation tank with the possibility of adding coagulant and polymer based on the turbidity of the wastewater which, is measured using an on-line sensor located just after the second flotation tank. As a coagulant, ferric chloride, PIX-111, is used along with an anionic polymer and the set point of the automatic dosing is 3.5 NTU, meaning that the amount added to ensure a residual turbidity of 3.5 NTU or less is calculated. From the filter effluent samples are taken automatically using a Liquistation CSF48 which gathers a flow-proportional sample over a period of 24 hours.

2.2.7 The sludge treatment

The sludge produced from both the first and second chemical treatment steps is transported to a sludge tank for dewatering and thickening. From there the sludge is transported and a cationic polymer is added just before it reaches either the decanter centrifuge or screw conveyor. The water that is separated during the dewatering is then recirculated back to the first flotation and the dewatered sludge is stored for a short period before being sent away for incineration.

2.3 Requirements and parameters

The wastewater treatment plant (WWTP) began operating on 2021-01-19 and the second MBBR, drum filter, the second buffer tank, the second sludge tank, and a new sludge centrifuge was installed on 2021-12-22 except for the second buffer tank which began operating on 2022-05-19.

2.3.1 Environmental permit

In the environmental permit granted Oatly, the concentrations of BOD_7, total phosphorus, and total nitrogen are regulated for which the maximum concentrations are stated in Table 1. These requirements are valid for the first 24 months of operating the WWTP. The maximum concentrations are mean concentrations, for BOD_7 the mean is for a period of 6 months, for Tot-P and Tot-N the means are for a period of 12 months (Länsstyrelsen Skåne, 2020).

Table 1 The permit parameters and their maximum concentrations, the first 24 months of operation.

Permit parameter	Maximum concentration (mg/l)
BOD_7	8
Tot-P	0.3
Tot-N	10

After that the requirements regarding the time for the mean concentration of BOD_7 is changed to 3 months and the maximum concentration of Tot-P is lowered however the mean period is unchanged. The requirements that apply after the first 24 months of operating are presented in Table 2 (Länsstyrelsen Skåne, 2020).

Table 2 The permit parameters and their maximum concentrations, after the first 24 months of operation.

Permit parameters	Maximum concentration (mg/l)
BOD_7	8
Tot-P	0.2
Tot-N	10

It is also stated in the permit that measurements of BOD_7, Tot-P, and Tot-N are to be taken at least 12 times a year. Besides the maximum concentration of containments the treated water also has to have a pH of 6.5-10 and measurements are to be done continuously (Länsstyrelsen Skåne, 2020).

2.3.2 Design parameters

Some of the maximum design loads and concentrations are presented in Table 3, so not exceeding these limits the WWT is designed to fulfill the requirements of the permit.

Table 3 The design loading rates and concentrations of the WWT, all values are maximum values except for pH which is a monthly average.

Parameter	Unit	Design capacity
Flow	m^3/day	1500
Biological Oxygen Demand (BOD_7)	kg/day	5460
Chemical Oxygen Demand (COD)	kg/day	8730
Total nitrogen (Tot-N)	kg/day	210
Total phosphorus (Tot-P)	kg/day	19
Sulphate (SO_4)	mg/l	20
Suspended Solids (SS)	kg/day	1960
pH	-	5-11
Temperature	°C	30-43

2.3.3 Wastewater parameters

The wastewater parameters of the primary influent are very varying. It has not been established what might be the cause, but one belief is that it might be because the content of the wastewater generated from the production of different products varies. A possible consequence of this would then be the variation in wastewater parameters as the products and the amounts produce vary from week to week. Figures 2 to 6 illustrate how the influent concentrations of COD and BOD_7, the influent concentration of tot-P, the loading rate of COD and BOD_7, the flow rate, and the influent pH all have varied during the period 2021-11-01 to 2022-10-02. The same variations can also be found for the influent concentration of SO_4 and the temperature.

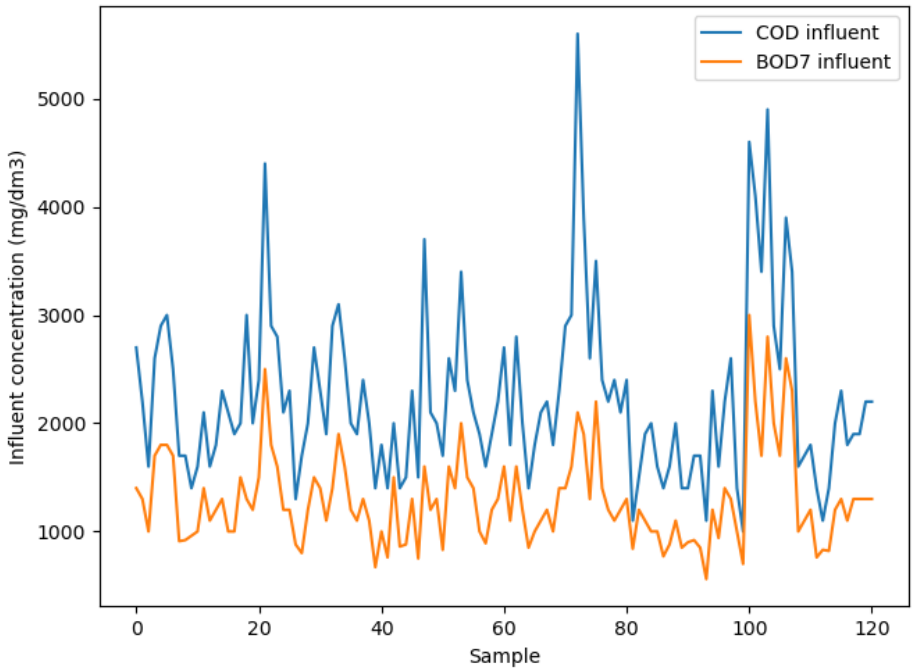

Figure 2 The concentration of COD and BOD_7 (mg/L) in the influent during the period 2021-11-01 to 2022-10-02.

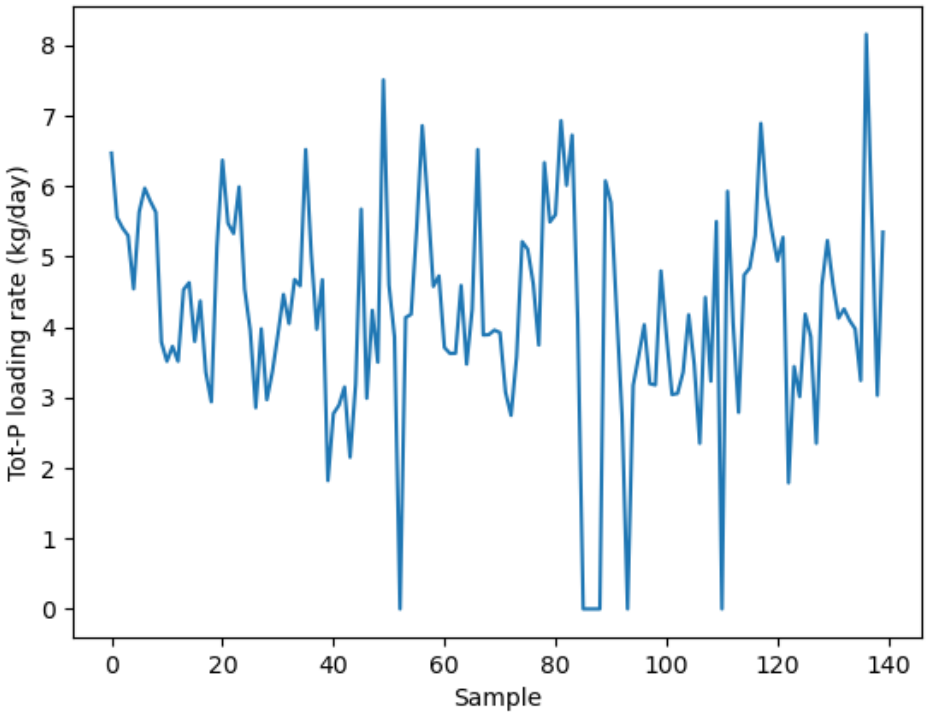

Figure 3 The loading rate of Tot-P (kg/day) during the period 2021-11-01 to 2022-10-02.

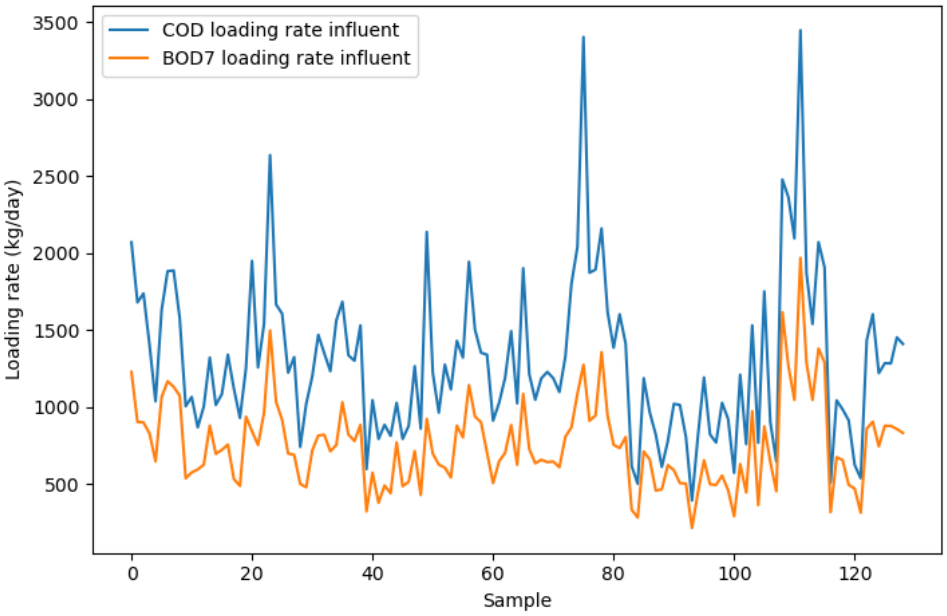

Figure 4 The loading rate of COD and BOD_7 (kg/day) during the period 2021-11-01 to 2022-10-02.

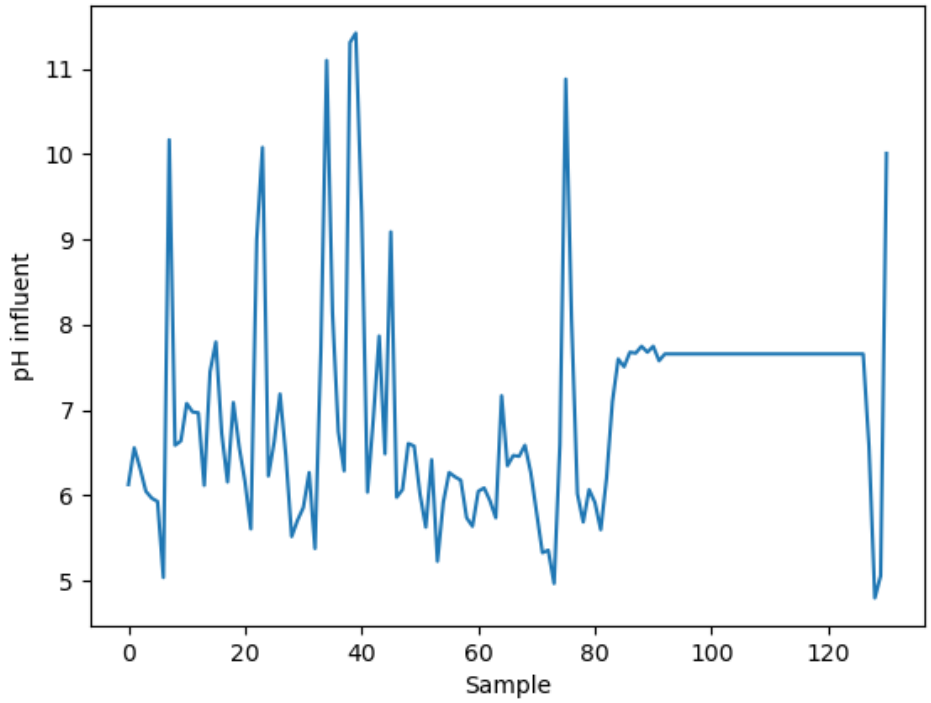

Figure 5 The influent pH during the period 2021-11-01 to 2022-10-02.

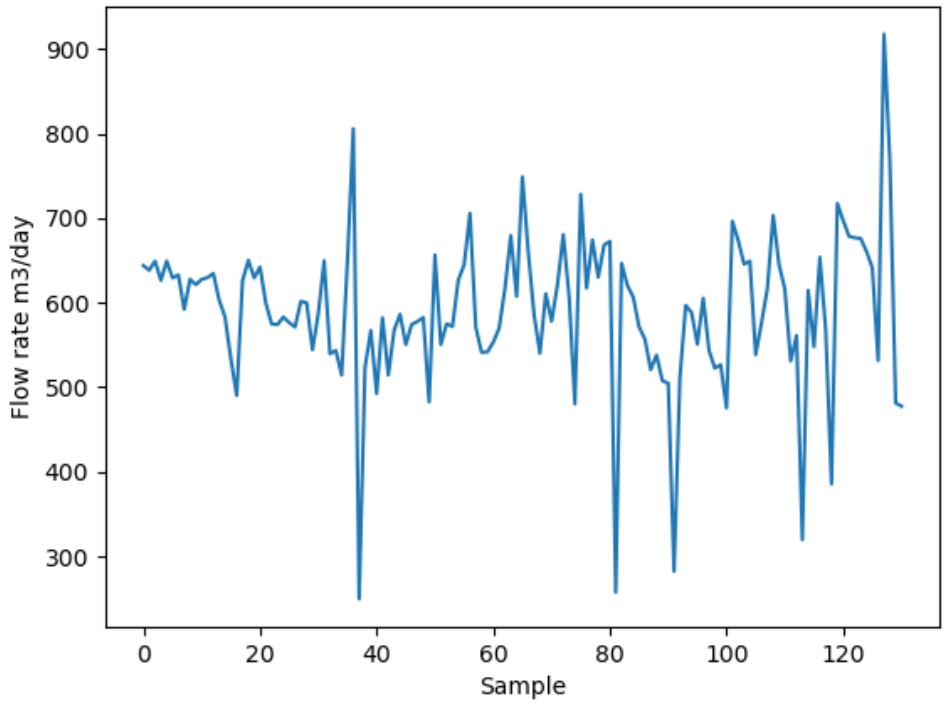

Figure 6 The flow rate (m^3/day) during the period 2021-11-01 to 2022-10-02.

Although there are large variations, all values are well within those used as design parameters from the WWT as Table 4 shows.

Table 4 The average values and design values for some of the wastewater parameters recorded from the period 2021-11-01 to 2022-10-02.

Wastewater parameter	Unit	Average value	Design value
COD loading rate	kg/day	1300	8730
BOD_7 loading rate	kg/day	760	5460
Tot-P loading rate	kg/day	4.18	19
pH		7.06	5-11
Flow rate	m^3/day	580	1500

This means that none of the values can be considered abnormal and the WWT is designed to fulfill the requirements of the environmental permit at this composition. This is not always the case however as Figure 7 shows.

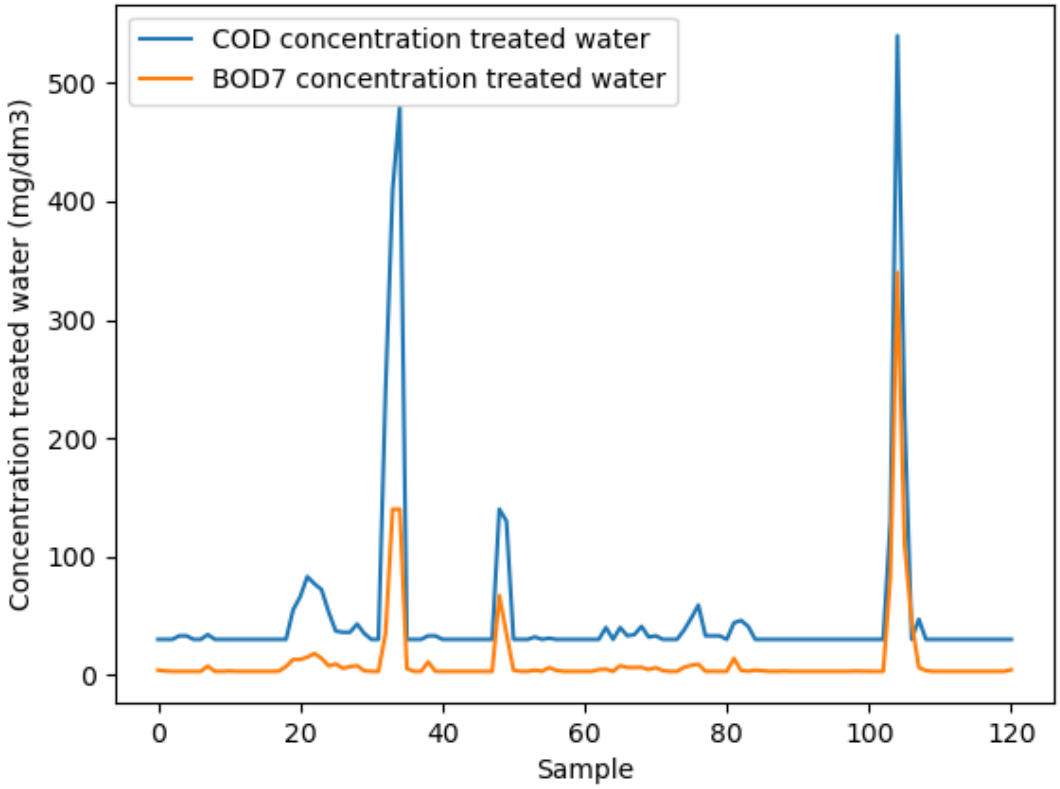

Figure 7 The concentration of COD and BOD_7 in the treated water during the period 2021-11-01 to 2022-10-02.

The concentration of COD in the treated water is not regulated in the permit however the concentration of BOD_7 is and as previously mentioned, the average concentration for 6 months can be no higher than 8 mg/L. The graph shows the concentrations for the period

2021-11-01 to 2022-10-02 with the average concentration of BOD_7 being 7.7 mg/L, so just slightly under the limit. However, as the figure also shows, the main reason for the high average concentration is a few concentration peaks. Most of these peaks are caused by various disturbances, for example, on 2022-04-07 the phosphoric acid was used up and accidentally replaced with biocide until 2022-04-12 when it was once again replaced by phosphoric acid. During these couple of days, the concentration of BOD_7 in the treated water varied between 45 and 340 mg/L. There are other examples of disturbances but this one appears to have had the greatest impact on the quality of the treated water. So excluding the samples from those days instead results in an average BOD_7 of 5.55 mg/L. After this period the number of disturbances has decreased and the average BOD_7 concentration is currently 4 mg/L (calculated for the period 2022-10-12 to 2023-04-12) indicating that the treatment process, under normal circumstances, is performing satisfactorily.

Achieving these results however considering the challenges caused by the varying composition of the influent, requires multiple adjustments of the PAC dosage used for both the first and

second coagulation. Figure 8 shows how the dosages have been changed during the period and as can be seen, in particular, the second dosage has changed a lot.

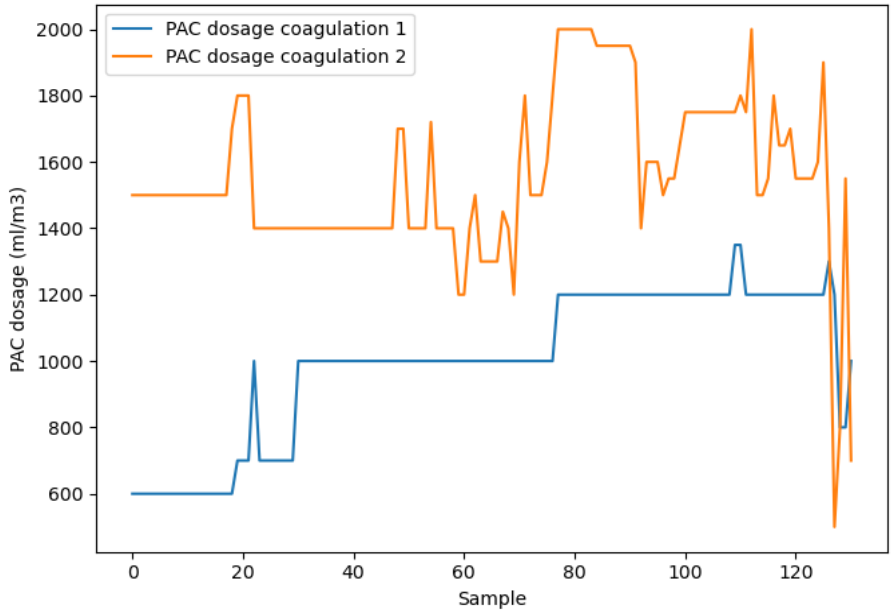

Figure 8 The PAC dosages used for both the first and second coagulation during the period 2021-11-01 to 2022-10-02.

Each adjustment requires the operators to conduct several jar tests to approximate a suitable dosage. Based on those results and the operators' experience the dosage is then changed and the results are evaluated by performing a visual inspection of the quality of effluent. For the second chemical treatment step there is also the possibility of using the results from the on-line turbidity monitor located shortly after the flotation unit. This approach requires that the operators are experienced enough to perform the visual inspection and can also be a source of frustration as multiple jar tests might be required before finding a suitable dosage. There have also been occasions where the results from the jar tests have been difficult to apply as a much large amount has been required for the actual process compared to what the tests showed. In addition, this approach makes it difficult to determine how much coagulant that is required to fulfill the requirements of the treated water. The very low concentrations of BOD_7 in the treated water also indicate that there could be a possibility of decreasing the PAC dosage without exceeding the permit. This is however difficult to do just based on the results from jar tests due to the varying water quality parameters.

3 Theory

The theory section is focused on different wastewater parameters, coagulation-flocculation including important factors, PAC, flotation as a separation technique, and different methods to estimate coagulant dosage.

3.1 Wastewater parameters

This section covers some common wastewater parameters and their relationship to each other.

3.1.1 BOD

Biological oxygen demand, BOD, is a measurement that can be used to approximate the amount of biochemically degradable organic matter in the wastewater and is an important indicator of the quality of wastewater. It is defined as the amount of oxygen that is required for an aerobic microorganism to oxidize the organic matter to a stable inorganic form. BOD_5 and BOD_7 are measurements of the amount of oxygen consumed for a period of 5 respective 7 days. (Jain and Singh, 2003)

3.1.2 COD

Chemical oxygen demand, COD, is related to BOD and also describes the bioavailability of carbon in a sample. It is a measurement of the amount of oxygen equivalents that is consumed by a strong oxidant during the oxidation of organic matter. During this reaction both organic and inorganic components are oxidized as compared to BOD where only the biologically reactive carbon is oxidized. It can be used as a surrogate measurement of BOD by establishing a correlation between BOD and COD, assuming that the proportions and types of materials in the wastewater remain relatively constant. Normally COD is in the range of 1.3 to 1.5 times the BOD. Besides the possibility of acting as a surrogate measurement, it is also often used as a measurement of pollutants in wastewater (Hu and Grasso, 2005; Jain and Singh, 2003; Woodard & Curran, 2006).

3.1.3 TOC

Total organic carbon, TOC, is a measurement of the concentration of organic matter in the wastewater and is therefore related to both COD and BOD_7 (Assmann et al., 2017). Studies highlight the possibility of establishing a ratio of TOC and COD for both influent and effluent respectively (Dubber and Gray, 2010). However, TOC concentrations cannot replace BOD_7 in issues related to environmental permits but can be used for process control purposes. TOC measurements are especially suited for process control as in-line monitoring is possible by providing real-time data. These in-line detectors can use different methods to oxidize the organic content in a sample to generate CO_2 which is then detected. The time required for this is often just 5-10 minutes as compared to COD tests which usually take a few hours to complete (Assmann et al., 2017).

3.1.4 Turbidity

Turbidity is a parameter that can affect the quality of water and can be used to describe how smoky or hazy the wastewater appears. It is defined as a measurement of the light-scattering properties caused by fine particles in a sample (Woodard & Curran, 2006; Khan and Ali, 2018).

These light-scattering properties also affect light penetration and treated wastewater with high turbidity can have a negative impact on the habitat quality for aquatic life (Water Science School, 2018).

3.2 Coagulation and flocculation

In this section, coagulation is described in more detail, including reactions and mechanisms, as well as factors that are important for the efficiency of coagulation.

3.2.1 Chemical reactions

Coagulants consist of positively charged metal ions and form positively charged hydroxides that can crosslink and aggregate impurities thereby neutralizing their charge (Parsons et al., 2014). The ability to form aggregates also depends on the turbulence and mixing speed as it determines the frequency and number of particle collisions. This dictates the likelihood of the hydroxides coming into contact, with and reacting with, the contaminants.

Coagulants of metal salts, often iron or aluminum-containing trivalent ions, Al^{3+} or Fe^{3+}, react with water forming hydroxide ions along with hydrogen according to the reversible reactions shown in equations 1-4 (Bottero et al., 1980; Kemira, n.d.).

$$Al^{3+} + H_2O \rightarrow Al(OH)^{2+} + H^+ \quad (1)$$

$$Al^{3+} + 2H_2O \rightarrow Al(OH)_2^+ + 2H^+ \quad (2)$$

$$2\,Al^{3+} + 2\,H_2O \rightarrow Al_2(OH)_2^{4+} + 2H^+ \quad (3)$$

$$13Al^{3+} + 28H_2O \rightarrow Al_{13}O_4(OH)_{24}^{7+} + 32H^+ \quad (4)$$

The formed hydroxide ion can be further hydrolyzed as described by the two equations below (Kemira, n.d.).

$$AlOH^{2+} + H_2O \rightarrow Al(OH)_2^+ + H^+ \quad (5)$$

$$Al(OH)_2^+ + H_2O \rightarrow Al(OH)_3 + H^+ \quad (6)$$

In partially neutralized aluminum solutions, the dominating polymeric compound is $Al_{13}O_4(OH)_{24}(H_2O)_{12}^{7+}$ commonly abbreviated as Al_{13}^{7+} (Pernitsky and Edzwald, 2003).

3.2.2 Coagulation mechanisms

There are different mechanisms influencing the coagulation process however the two most important ones are charge neutralization and sweep coagulation and which mechanism that will dominate depend mainly on the pH and concentration of the coagulant (Wei et al., 2015). Charge neutralization is where the cationic Al (or Fe depending on the coagulant) adsorbs to the surface of the negatively charged contaminants thereby neutralizing their zeta potential. The zeta potential is a measurement of the strength of the electrostatic charge of a particle, contaminants usually have a zeta potential of -14 to -30 mV if the pH is in the range between 5 to 8. The more negative zeta potential the more strongly negatively charged the colloidal (Ecolab company, 2018, pp.203). Coagulants can be used to neutralize the charge and once the zeta potential is in the range of -8 mV to +3 mV coagulation can occur (DeNigris, 2020). During

sweep coagulation ions adsorb to the surface of the containments enabling the growth of layers of aluminum hydroxides eventually causing enmeshment of the particles (Bache et al., 1999).

The different mechanisms have different advantages, sweep coagulation for example requires a higher coagulant dosage which results in higher sludge production. However, the mechanism is less sensitive to suboptimal dosage conditions, especially overdosing, which can make process control easier. It is also advantageous when treating wastewater with low turbidity as a large amount of coagulant increases the probability of contact between containments and precipitate particles (Bache et al., 1999). The quality of the formed flocs is also influenced by the dominating mechanism where sweep coagulation is associated with large flocs and charge neutralization with stronger microflocs (Cruz et al., 2020).

3.2.3 pH

The pH of the wastewater has a substantial influence on the removal of COD and by extent, the amount of coagulant required for a sufficient removal efficiency however this effect decreases with high amounts used (Ahmad et al., 2008; Ecolab company, 2018, pp. 210).

As previously mentioned, charge neutralization dominates at relatively low pH and dosage while sweep coagulation occurs mainly at high pH and coagulant dosage (Wei et al., 2015). This is due to charge neutralization dominating for small aggregates which is the case at low pH and dosage (Ahmad et al., 2008). It is therefore possible to influence the coagulation process by adjusting the pH before the addition of a coagulant (Johnson et al., 2019). For the removal of natural organic matter in raw water using PAC as coagulant, the best removal efficiency was achieved at a pH close to the minimum solubility of the coagulant. This can be achieved by matching the basicity of the coagulant to the alkalinity of the wastewater (Pernitsky and Edzwald, 2006).

3.2.4 Mixing and motion

There are different types of motion affecting the particles depending on the size of the particles. The motion of small particles is dominated by Brownian motion and flocculation occurring due to this is referred to as perikinetic flocculation, flocculation due to the natural movement of the particles (Kemira, n.d.). Flocculation occurring due to the addition of energy is referred to as orthokinetic flocculation and energy is most commonly added through mixing.

The rate of floc formation is determined by the rate of collisions between the particles. Increasing the mixing rate leads to an increase in velocity gradient thereby increasing both the frequency and the number of collisions. increases the movement causing an increase in the frequency and number of collisions (Huck and Sozański, 2011).

However, one cause of the breakage of flocs is too large shear forces caused by both the floc growth and too rapid mixing. The rate is therefore the result of the balance between the floc formation and breakage as described by the equation below where R_{Br} is the rate of floc breakage, R_{Col} the rate of particle collision, $\propto$ a collision factor and R_{Floc} the rate of floc formation (Jarvis et al., 2005).

$$R_{Floc} = \propto R_{Col} - R_{Br} \quad (7)$$

The collision factor is dependent on the effective shear rate and particle size and indicates the reversibility of floc breakage with a value of 0 if the breakage is fully irreversible. As a

consequence of what is described in the equation above, if the shear rate, and thereby the rate of breakage, is kept constant, steady state will eventually reign (Jarvis et al., 2005).

If steady state is not the goal but instead to maximize the rate of floc formation, the speed of the mixing can be adjusted to ensure high velocity during the aggregation stage, when the rate of collision is greater than the rate of breakage, and then decreased as the flocs grow to cause a decrease in the rate of floc breakage. This can be done using the velocity gradient, *G*, and retention time, *t*, and progressively reducing the *Gt* value. Optimal values of *G* and t are often established in practice and differ depending on the separation method following the flocculation (Huck and Sozański, 2011).

Flocs breaking risk having a direct effect on the efficiency of the sequent separation methods, e.g., the removal efficiency of the flotation. For flotation, this is due to the smaller broken flocs being less efficiently adsorbed by the air bubbles. The flocs do have the ability to re-grow but exposed to too large shear forces that ability is diminished (Jarvis et al., 2005). The possibility of re-flocculation is influenced by the coagulation mechanism, which is dependent on, inter alia, the coagulation dosage, with charge neutralization having the highest re-flocculation factor (Miranda et al., 2020).

3.3 Flotation

Flotation is a separation method where the wastewater is mixed with pressurized water before entering the flotation tank. Due to the change in pressure bubbles are formed to which the flocs adsorb and then float to the surface. After reaching the surface the flocs can either be allowed to overflow or the containments can be scraped off as sludge. (Huck and Sozański, 2011). The efficiency of flotation as a separation method depends on several factors such as stirring velocity, residence time, size of flocs, the ratio of pressurized water, the addition of polymers, and pH (Ødegaard, 1995).

3.4 PAC

Poly-aluminum chloride, or PAC, is a pre-hydrolyzed metal-ion coagulant that is efficient at COD removal with the capacity of achieving a COD reduction of over 90% (Ahmad et al., 2008). The molecular formula of PAC is $[Al_m(H_2O)_x] \cdot Cl_{3m-n}$ (n ≤ 3m) and can also contain different amounts of hydroxyl (Li et al., 2010). These hydroxyl ions will react with water causing the release of hydrogen ions, this is referred to as the basicity of the coagulant which controls the pH-lowering effect of the coagulant. The higher the basicity, the smaller the pH-lowering effects, and generally, commercial PAC has a basicity of between 15-85%. The basicity can also be used as a measurement of the fraction of polymeric and colloidal Al in the coagulant which increases with the basicity (Pernitsky and Edzwald, 2003). This is especially useful as it also indicates which monomeric are the most abundant which is important when considering the coagulation mechanism. In PAC with a basicity of below 33% the most plentiful species are monomers such as Al^{3+} and $AlOH^{2+}$, both of which can only partake in charge neutralization (Ahmad et al., 2008). When comparing different aluminum species it was found that this monomeric aluminum was the most efficient at the removal of soluble COD and Suspended Solids (Miranda et al., 2020).

The advantage of PAC compared to conventional coagulants such as aluminum or ferric sulfate, is that it is efficient over a larger pH range, and less coagulant is generally required to achieve the same removal efficiency (Li et al., 2010). What pH is optimal depends on the composition

of the wastewater treated and can differ, it can for example be around 7.5-8, or around 6, or between 5-6 (Ahmad et al., 2008; Kemira, n.d; Nti et al., 2021).

The PAC used in the WWT at Oatly is Ekoflock 90 which has a basicity of 45 ± 3 weight%, an Al content of 9.0 ± 0.3 weight%, and a density of 1370 ± 25 kg/m^3 (Feralco, 2021).

3.5 Optimizing coagulant dosage

In this section, traditional methods, such as jar tests, are discussed along with more novel approaches such as using mathematical models or machine learning to determine a suitable coagulant dosage.

3.5.1 Jar tests

Jar tests are performed by taking samples of the wastewater, adding coagulant, and then analyzing the concentration of the compound of interest, the turbidity, or inspecting the quality of the flocs. Rapid mixing is used as the coagulant is added but once the floc formation has begun it is switched to slow stirring. After the flocculation, the flocs sediment before any analysis is performed. The tests usually take approximately 30 minutes to perform not including analysis of the results (Kemira, n.d.). As previously mentioned, the efficiency of the coagulant is dependent on several factors which means that with very varying water quality parameters a great number of tests are required which makes the method very time-consuming (Heddam et al., 2011).

3.5.2 Mathematical models

As an alternative to this, there have been made several attempts to use data collected at wastewater treatment plants to create models to be used for control of coagulant dosage based on parameters such as pH, temperature, turbidity, suspended solids, total organic content, alkalinity, total phosphorus concentration and conductivity among others. The data collected can be analyzed using different mathematical methods such as multivariate regression analysis to find an equation describing the relationship between the raw water parameters and coagulant dosage. At Hazard's Green Waterworks, multiple regression was used to find the equation shown below which could be used to calculate the dosage of iron sulfate based on the turbidity, conductivity, temperature, and flow of raw water (Ratnaweera and Fettig, 2015).

$$Iron\ sulfate\ dose\ (mg/l) = C_0 + C_1 \times Turb + C_2 \times \log(Turb) + C_3 \times Cond + C_4 \times \log(Cond) + C_5 \times Temp + C_6 \times \log(Temp) + C_7 \times Flow + C_8 \times \log(Flow) \quad (8)$$

A similar correlation could be established for surface water from Norway where the concentration of residual NOM (NOMr) could be calculated from pH, initial NOM (NOMi), and coagulant dosage (Al-dose), as seen below (Ratnaweera and Fettig, 2015).

$$NOM_r = (C_1 + C_2 \times pH + C_3 \times pH^2 + C_4 \times pH^3) \times Al - dose^{C5}(mg/l) \times NOM_i^{C6} \quad (9)$$

An evaluation of data from jar tests performed using 21 raw water sources using either ferric salts or alum as coagulant was used to develop a model that could be used to predict non-adsorbable DOC based on influent DOC (DOCi) and its specific ultraviolet absorbance (SUVAi), see equation below (Ratnaweera and Fettig, 2015).

$$DOC_{na} = DOC_i \times (K_1 \times SUVA_i + K_2) \quad 10$$

These models are all based on mathematical analysis of gathered data and are usually created using either multivariate regression analysis, artificial neural network models (ANN), or fuzzy logic models (Ratnaweera and Fettig, 2015).

3.5.3 Machine learning

An alternative is to use various Machine Learning (ML) algorithms that have shown to be a promising substitute for jar tests and relying only on the experience of the operators (Zhang et al., 2013). ML refers to a system's ability to use data to find patterns making it possible to make predictions based on unknown data. This can be done by training the system using historical data, and supervised machine learning, to learn the relationship between a large number of input and output variables. One of the advantages is that a linear relationship is not required, and different algorithms use different learning techniques making it possible to use for several different types of applications (Woolf, 2009).

A decision tree is an algorithm that has a structure like a flowchart where the data is divided into decision nodes and leaves. The decision nodes are how the system splits the data and the leaf nodes are the result of the splitting. The algorithm then continuously splits the data and compares the results with the dependent variable thereby trying to find the splitting that results in the leaf nodes closest to the real result (Schonlau and Zou, 2020). Figure X below shows an example of what a decision tree might look like based on the dataset from Oatly's WWT.

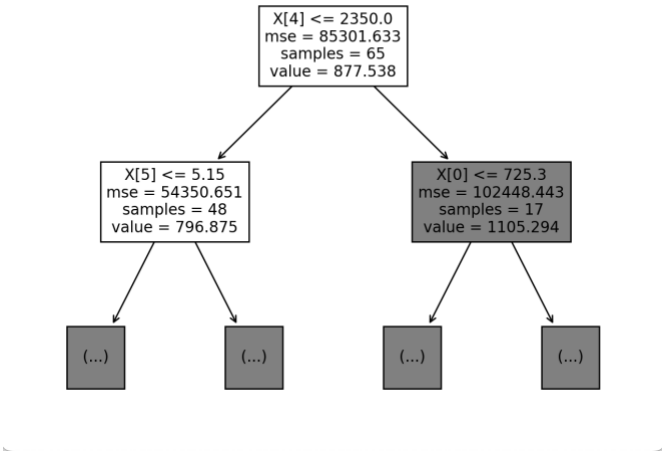

Figure 9 An example of a decision tree, for plotting purposes the height has been set to 1.

In the figure, the root node (the first "box" in the figure) is the starting point of the tree. X[i] denotes the variable that the split is based on, here the dataset is split so that all COD concentrations in the influent lower or equal to 2350 is to the left, the rest to the right. MSE is the mean squared error of the real concentration of COD in the effluent compared to the predicted. Samples state the number of data points (here the number of rows from the dataset) and the value is the predicted COD concentration in the effluent. As can be seen, the base of the splitting can be different independent variables and the algorithm will continuously split the tree to try to find the tree that results in the smallest MSE value. In this example the solutions

are evaluated using MSE the aim can also be to minimize the mean absolute error (MAE) or other functions can be used for evaluation to find the “best” decision tree.

Random Forest (RF) is an algorithm that uses a combination of several decision trees to make predictions. The algorithm can be used for both classification and regression and can handle nonlinear relationships. By using several decision trees, the risk of overfitting the model decreases when compared to using a single decision tree (Schonlau and Zou, 2020). This causes RF to be a suitable algorithm when working with small datasets(Aliashrafi et al., 2021). RF can be used for example to optimize a coagulation-flocculation process to ensure optimal pH, amount of coagulant, and settling time, to minimize the residual turbidity. There is even a possibility of this using results from jar tests which makes it possible to evaluate changes in operating conditions without risking disturbing the treatment process (Ugonabo et al., 2022). Another study compared different machine learning algorithms also used to predict turbidity and found that RF along with something called Adaptive Neuro-Fuzzy Inference System showed the highest accuracy levels of removing water
turbidity (Arab et al., 2022). In conclusion, RF has proved to be an attractive alternative to jar tests and can be used as a tool for operators to predict concentrations in the effluent, optimize operating conditions, estimate coagulant dosage, and can be used with data extracted from, for example, a SCADA system like the one used at Oatly (Dürrenmatt and Gujer, 2011).

4 Methodology

As illustrated in section *2.3.3 Wastewater parameters* the influent concentration of BOD_7, COD, Tot-P, and SO_4 varied greatly along with the pH, temperature, and flow rate. Due to these fluctuations, jar tests are an unsuitable method as the results are highly dependent on the wastewater parameters making any conclusions only valid for the specific composition of the sample used for testing. An alternative approach is to use machine learning which has the potential to take these variations into account when making predictions based on what it has learned from training on data from previous samples. When using machine learning algorithms the results will be very dependent on the data that has been used for training the model. This section, therefore, describes how the data has been collected by Oatly, how the selection of variables was done from all of the available wastewater parameters, what preprocessing of the data was required, and finally, what algorithm and libraries have been used for the creation of the model.

4.1 Data collection

The data used is historical data covering the period 2021-11-04 to 2022-03-25 that has been gathered from internal documentation and the software used for process monitoring, VA-operatör, and then compiled. VA-operatör is a SCADA system developed by VA-ingenjörerna where data from the on-line sensor is gathered and it is also used to control the treatment process by, for example, changing the dosage of coagulant. The software can also be used to extract information, including the flow rate, temperature, and pH, from these on-line sensors. There is also a possibility of recording both the results from analyses done by accredited laboratories and by the operators. The accredited laboratory ALS is responsible for analyzing samples from the primary influent, the treated water, and samples taken from the process by the operators. An automatic sampler is used to take samples from the primary influent and from the treated water, both of which are taken flow-proportionally and show the average concentration over 24 hours. The samples from the treated water are taken automatically by the Liquistation CSF48 is flow-proportional, and each sample is taken for 24 hours.

Previously it has also been possible for the operators to use LCK314 Hach-Lange test kits to measure the concentrations of COD using a DR3900 spectrophotometer also from Hach-Lange. The LCK314 contains potassium dichromate, the usage of which is now highly regulated in Sweden and the rest of Europe (Mantech, n.d.). Therefore, today other test kits are used to measure the concentration of TOC instead.

So the data from the primary influent and treated water are average concentrations, and the sample concentrations in the effluent from the first and second chemical treatment steps are from discrete samples. It has not been established if, and in that case, how large the variations in concentration are during a day.

4.2 The dataset

The most intense period of measuring was between 2021-11-04 to 2022-06-24 when samples were taken 2-5 times a week, resulting in a total of 142 samples. The samples were taken from different parts of the process and different parameters were measured at different points. Some

of the wastewater parameters measured are presented in Table 5 along with where the sample was taken. The table also includes data from the on-line sensors.

Table 5 Some of the wastewater parameters measured during the period 2021-11-04 to 2022-06-24 and from where in the process the samples were taken. There are also comments in those cases that the routine of taking samples changed during the period.

Parameter	Primary influent	Effluent 1st coagulation	Influent 2nd coagulation	Effluent 2nd flotation	Treated water	Comment
Flow rate	x				x	
BOD_7	x		x		x	
COD	x	x	x	x	x	Effluent concentration only measured until 2022-03-26
TOC	x	x		x	x	Effluent 1st coag. and 2nd flot. measured occasionally 2022-06-20 to 2022-10-03
Tot-N	x				x	
Tot-P	x				x	
PO_4-P	x		x		x	Measured occasionally
SO_4	x				x	
SS	x	x	x	x	x	
pH	x				x	
Temperature	x				x	

As the table shows a lot of the samples from both the first and second chemical treatment steps are taken either the influent or effluent. However, there is more data related to the first chemical treatment step, mostly due to data from the on-line sensors. Another issue regarding the second chemical treatment step is that the level of detection for COD is 30 mg/L so all concentrations of 30 mg/L or less are just recorded as "<30 mg/L" making much of the data difficult to use. In

addition to what is presented in Table 5, there was also data available regarding the coagulant and polymer dosage for both the first and second coagulation.

Due to this, it was decided to focus on the first coagulation and flotation of the process instead of the second to maximize the amount of data available for use. Also, as the focus of the project is to investigate if it is possible to use the data available to create a model that can be used to estimate the COD concentration in the effluent, only data from 2021-11-04 to 2022-03-26 was included in the dataset as measurements of the COD concentration in the effluent was discontinued after this point. Only including data containing the COD concentration in the effluent decreases the size of the dataset from 142 rows to 65 rows. A number of the wastewater parameters included in this dataset are presented in Table 6 along with the minimum, maximum, and average values. It is this dataset that the modeling described later on in this rapport is based on.

Table 6 The minimum, maximum, and average flow rate, temperature primary influent, pH primary influent, PAC dosage, loading rate, and concentration of COD, Tot-P, and SO_4 in the primary influent. The data is from the period 2021-11-04 to 2022-03-26 and no rows containing missing data have been included in these calculations.

Water quality parameter	Minimum	Maximum	Average
Flow rate (m^3/day)	170.5	748.8	583.3
Temperature influent (°C)	11.9	38.8	30.6
pH influent	5.33	11.5	7.22
PAC dosage (ml/m^3)	600	1200	977
COD influent (mg/L)	1100	3500	2070
Tot-P (mg/L)	3.7	16	2.0
SO_4 (mg/L)	1	29	3.4

As the coagulant used, Ekoflock90, has an aluminum content of 9 weight% these PAC dosages of between 600ml/m^3 and 1200ml/m^3 correspond to an aluminum concentration of approximately 148 mg/L and 74 mg/L respectively. It should be noted that the concentrations of the primary influent are average concentrations while the concentrations from the effluent are discrete samples.

4.3 Selection of variables

The dataset is made up of the wastewater parameters, flow rate, temperature, pH, coagulant dosage, polymer dosage, the influent concentration of COD, Suspended Solids, BOD_7, Tot-P, and SO_4, and the effluent concentration of COD and Suspended Solids. There was also a possibility of including the primary influent concentrations of ammonia, total nitrogen, and fat but this data was believed to be redundant for the model and was therefore excluded.

The choice of independent variables was chosen based on factors that have a large impact on the efficiency of a coagulant as described in section *3.2 Coagulation and Flocculation*. The final selection was done by testing different constellations of independent variables and then evaluating the resulting model.

4.4 Data preprocessing

During the data preprocessing any rows containing missing data were removed from the dataset as the algorithm requires that no data is missing from the independent variables. In addition, data with a pH of 9 or above was removed from the set, further decreasing the size of the dataset. This is because if the pH is too high it will impact the coagulation mechanism which will have consequences for the coagulation efficiency. There is a risk that this might change the "pattern" that the algorithm establishes and from which it makes predictions as the relationship suddenly behaves differently. To avoid this risk, this data was therefore removed.

4.5 Modeling

The modeling was done in Python using a RandomForest regression algorithm. For the creation, the libraries pandas and scikit-learn were imported. Pandas were mainly used to import all data from Excel to Python and make it readable. Pandas DataFrame was used to store the data for the variables used in the model.

The method test_split was used to divide the dataset into a training and a test set. The test set consisted of 10 % of the total dataset and, as discussed later on in this report, 10 % was chosen to maximize the size of the training set.

To evaluate the results from the model the built-in functions of the module sklearn.metrics was used and matplotlib was used to create all the figures and graph used in the report.

5 Results and discussion

In this section, the final selection of variables from all the available wastewater parameters is discussed along with the results from the model based on these. It is also evaluated how accurate the predictions of the COD concentrations in the effluent from the first flotation are using statistical methods and by discussion of the model's limitations. Finally, the model is used to predict the COD concentration at different dosages of coagulant to test if it could have been possible to, according to the model, decrease the dosage and by how much. To illustrate the possible gains of using the model, it is calculated what Oatly possibly could save based on these results in terms of CO_2 emissions and SEK.

5.1 Final selection of variables and preprocessing

As mentioned in section *2.3.3 Wastewater parameters*, the relevant wastewater parameters in the dataset were flow rate, temperature, pH, the influent concentration of COD, Tot-P and SO_4, the PAC dosage, and the COD concentration of the effluent. In addition, there were also data regarding the concentration of suspended solids, total nitrogen, ammonia, and alkalinity, along with other water quality parameters.

During the literature overview, possible variables from other studies appeared to be flow rate, temperature, pH, influent COD concentration, influent Tot-P concentration, coagulant dosage, effluent COD concentration, and turbidity. These variables were either used along machine learning or when trying to establish mathematical correlations.

When discussing the possibility of removal of COD and BOD_7 it is important to take into consideration that these measurements include both soluble and particulate COD/BOD_7 (Mackenzie Davis, 2020). There is also a possibility of removal of some soluble matter however, during the chemical treatment, the majority of the COD reduction, for example, is due to the removal of particulate COD (Haydar and Aziz, 2009). It can therefore be misguiding to evaluate the removal efficiency based on the reduction in COD and BOD_7 concentration if the fraction of soluble COD/BOD_7 varies. So different fractions of soluble and particle COD/BOD_7 could mean that the removal efficiency does vary although the difference in concentration before and after the treatment is the same (Iwapublishing.com, 2010). The concentration of particulate COD/BOD_7 can be measured by filtering the sample prior to analysis and assuming that the fraction of particulate vs soluble does not vary, makes it possible to estimate how much of the total COD that can be considered possible to remove. Turbidity measurements can also be used to get an indication of the amount of particulate COD as it is a measurement of the amount of suspended particles in the wastewater which is related to the amount of particulate COD (Mucha and Kułakowski, 2016). Although the turbidity does not only depend on the amount of particulate COD or BOD, but it provides a good complement when trying to evaluate a chemical treatment step (El bied et al., 2021). The dataset used here does not contain any measurements of the turbidity, however, the concentration of suspended solids has been recorded. The concentration of suspended solids is related to the turbidity, so it was therefore considered a potential variable as it is also related to the amount of particular COD (Rügner et al., 2013). Another factor that can affect the efficiency of a coagulant is the concentration of SO_4 at high concentrations (Kemira, n.d.). This was therefore also considered a possible candidate when selecting variables for the model.

Based on this, the flow rate, pH, temperature, COD influent, Tot-P influent, and dosage of PAC were chosen as independent variables and the COD concentration in the effluent as the

dependent variable. Trials exchanging Tot-P for either SO_4 or influent SS were made however this resulted in less accurate predictions. This may be because even though SO_4 can affect the efficiency of a coagulant, the concentrations are too low to have an impact. Regarding the usage of SS as a variable, it is possible that it is unsuitable to use a combination of SS and COD since there is a correlation between the parameters (Guida et al., 2007).

Overall, the selection of variables is in line with what has been used in previous research where different machine learning algorithms have been used for predictions of COD concentrations in the effluent or coagulant dosages. Following the selection of variables the data was, as mentioned in *4.3 Data preprocessing*, preprocessed meaning that any rows containing missing data were removed from the set along with data where the pH was 9 or above.

The minimum, maximum, and average values for each variable can be seen in Table 7, the main difference from the entire dataset (as described in section *2.2.3 Wastewater parameters*) is that the dataset after preprocessing only consists of 56 rows (as compared to 65 rows) and the concentration of SO_4 in the influent has been excluded from the table.

Table 7. The minimum, mean, and max values of the independent variables when calculated from the entire dataset.

Independent Variable	Minimum	Maximum	Average
Influent flow rate (m^3/d)	171	749	589
COD concentration influent (mg/L)	1100	3500	2040
Ekoflock 90 (ml/m^3)	600	1200	980
Temperature influent (°C)	11.9	38.8	30.1
pH influent	5.33	8.14	6.72
Tot-P in (mg/L)	3.7	16.0	7.36

The maximum and minimum dosage of Ekoflock90 which has an aluminum content of 9 weight%, corresponds to an Al concentration of approximately 148 mg/L and 74 mg/L respectively.

Regarding the pH, there are on-line sensors that continuously measure the pH at three different locations related to the first chemical treatment step. One sensor is located just before the coagulation tank, one inside of the coagulation tank, and one in the flocculation tank meaning that there is a possibility to either include the pH at the start, during the coagulation, or after the coagulation. There have been attempts to include each of the three in the model however the starting pH caused the best model performance. It is possible that the starting pH potentially has a greater impact on the coagulation efficiency, or it is that this is the pH before adjustments have been done. The pH during the coagulation and after varies very little which might negotiate any effects that the pH might have had on the coagulant.

As mentioned in the section *Project Boundaries* this selection of variables has only been evaluated for the first coagulation-flotation step. It is therefore possible that this constellation is suboptimal for the second coagulation-flotation step, in particular, the concentration Tot-P as it only affects the coagulation efficiency at high concentrations which might not be the case for the second chemical treatment step.

5.2 Model estimations

The model was used to predict the COD concentration in the effluent from the first flotation, which was then compared with the real COD concentration, both of which are presented in Figure 10. The data used for this was the test set which consisted of 10% of the entire dataset, resulting in only 6 points. An alternative could have been to use a larger section of the dataset for testing however this would have decreased the amount of data available for training the model. It was decided that it was more important to maximize the amount of training data to improve the model than to use more data for testing to increase the validity of the results.

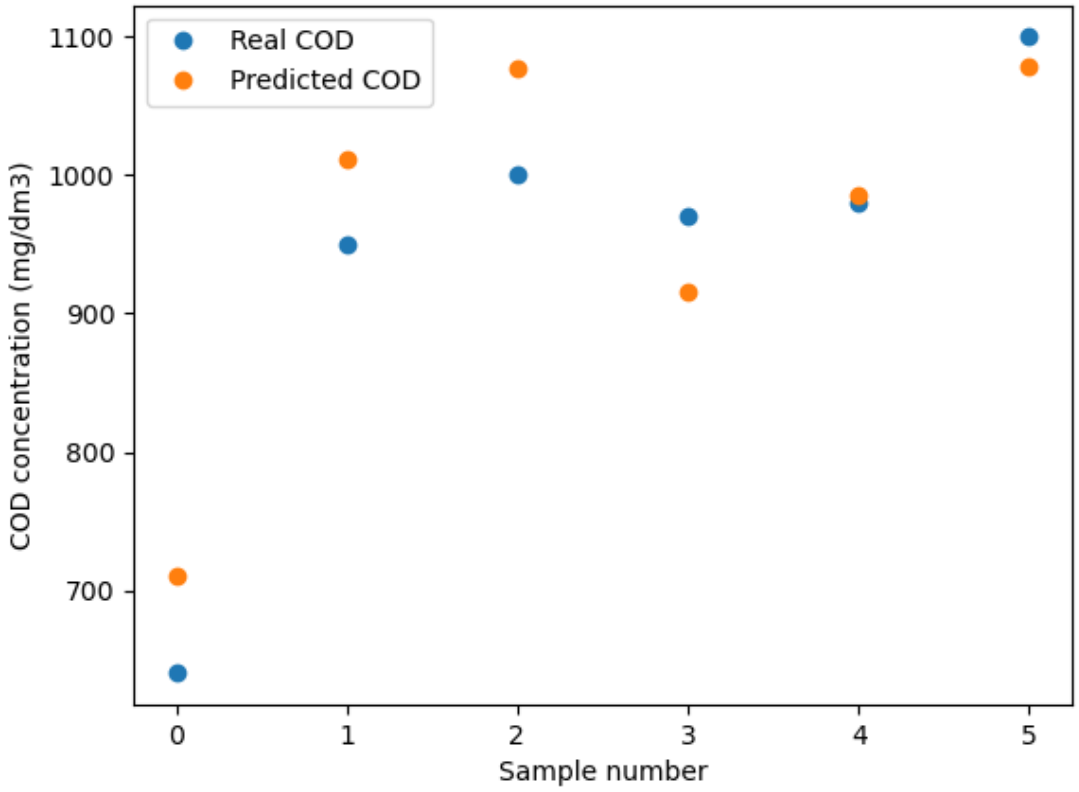

Figure 10. Real COD concentration vs the predicted COD concentration using the model, the comparison is done using the 6 data points making up the test set.

The figure illustrates how well the model is at making predictions based on unknown data by applying patterns learned from the training data. As can be seen, the shape of the curves is very similar although the accuracy of the predictions is not great. This could mean that the size and spread of the training data are enough for the model to find a pattern but not enough for very

precise predictions. A possible solution for this could either be to increase the size of the dataset or to optimize the algorithm through the tuning of hyperparameters. It is difficult to determine exactly which is the most limiting to the performance of the model, but the most likely obstacle is the size of the dataset.

It is usually recommended that the number of rows in a training set should be approximately ten times the number of variables used for the model (Alwosheel et al., 2018). This training set only consists of 50 rows even though the model is based on a total of 7 variables. However, although a smaller number of variables could have been beneficial from a machine learning point of view it would also have made the model unable to take important factors for the coagulation efficiency into account. Had there been smaller variations in water parameters decreasing the number of variables might not have posed a problem however in this case any attempts caused the model to perform worse.

Another solution could have been to increase the size of the dataset, however, as the model is based on the COD concentration both in the influent and effluent it was not possible to increase the amount of data as COD is no longer measured due to the environmental hazard of the reagent used (Mantech, n.d.). Instead, TOC is measured and, according to the theory presented in section *3.1.4 TOC*, it is plausible to assume that the TOC concentration could be used instead. However, both the investigation of the possibility of replacing COD with TOC and optimizing the algorithm is outside of the scope of this project.

The same results are illustrated in Table 8 and as shown the model could be used for making rough estimations but in its current form, it should be complemented with jar tests. But depending on the accuracy required, it could provide a plausible tool for the operators to decrease the amount of jar testing required.

Table 8. Real COD concentration and the COD concentration predicted by the model are based on data from the test set.

Sample number	Real COD concentration (mg/L)	Predicted COD concentration (mg/L)
0	640	711
1	950	1011
2	1000	1077
3	970	916
4	980	985
5	1100	1079

The model was also used to make predictions based on the entire dataset to see how good it was overall at finding the patterns that describe the coagulation efficiency. Figure 11 shows the results for the predictions done for the entire dataset, but it should be noted that no conclusions regarding the validity of the model should be drawn from this as the model has been fitted based on the majority of the data.

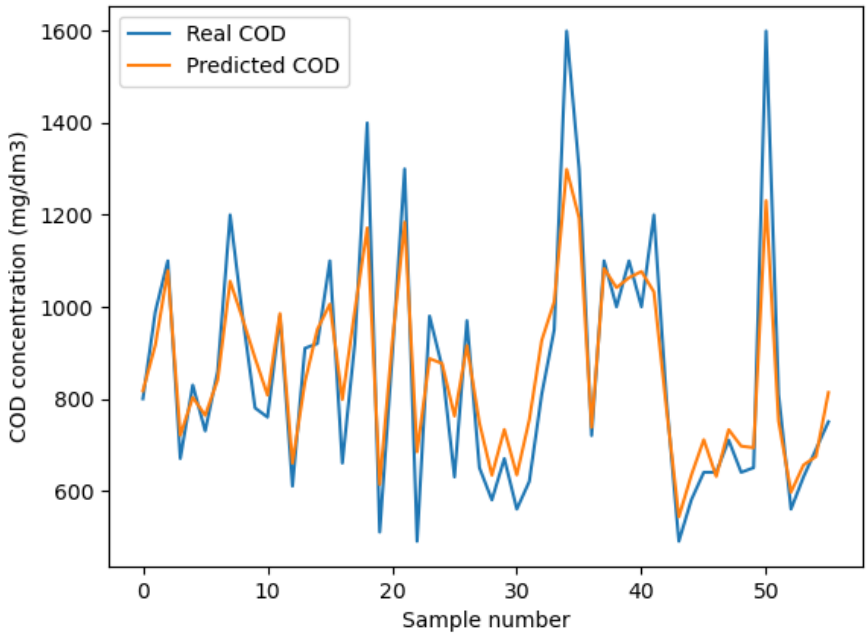

Figure 11. The real COD concentration and the predicted COD concentration for the entire dataset

However the figure does present the limitations of using the RF algorithm, it has difficulties making accurate predictions in cases where the COD concentration in the effluent is either very high or low. This is linked to the algorithm's inability to extrapolate; it is limited to the data that has been used for the training and so, if there is a need for extrapolation, other machine learning algorithms should be considered (Tyralis et al., 2019). It is also possible that the performance could be improved by simply increasing the size and range of the training set or it could be that because of the large variations in the dataset, other algorithms might be better suited, both solutions which are out of scope for this project.

5.3 Model evaluation

The performance of the model was evaluated by calculating the root-mean-squared-error (RMSE) and the goodness-of-fit (r^2) using sklearn.metrics in Python. The RMSE value described how far off the model predictions were compared to the actual values and the lower the value the more accurate predictions. The r^2 value describes the strength of the relationship between the independent and dependent variables and can be used to evaluate how well the pattern found by the model fits reality. To avoid overfitting where the model performs great using known data but not when faced with unknown data, the evaluation was done using unknown data, the test set.

The results from this evaluation are shown in Table 9 below and were mainly used to compare the model performance when selecting variables.

Table 9. The RMSE and r^2 values can be used to evaluate the accuracy of the model.

RMSE	r^2
48.98	0.85

An r^2 value of above 0.75 indicates a strong relationship between the variables as it means that 75% of the variance of the dependent variable can be explained by the variance of the independent variables (Allwright, 2022). Therefore this was the major method of evaluation when the independent variables were selected for the model. As mentioned in section *5.1 Final selection of variables* there is no guarantee that these variables are optimal for modeling the second chemical treatment step, mainly because of the presumed large difference in concentration of both Tot-P and SO_4 in the influent.

The RMSE value describes how far off the model predictions were from the actual concentrations and is mainly useful for comparing different models or discussing the accuracy of approximation required for the usage of the model. By improving the model it is possible to decrease the RMSE value and therefore decrease the need for validation of the predictions, for example by performing jar tests. A model with a low enough RMSE might even be used to determine the dosage of coagulant automatically as a future application.

5.4 Model limitations

One important model limitation, besides the inability to extrapolate, is the possible inability to take changes of the dominating coagulation mechanism into account. A change of the coagulation mechanism could result in a deviation from the pattern learned by the model thereby increasing the inaccuracy of the predictions. The risk of this could be avoided by ensuring that the model is only used under conditions where the same mechanism dominates. But it is also possible that with enough data the model can take this change into account when finding the pattern thereby decreasing this risk.

Another important limitation is the range of the data used for training the model. As previously mentioned, it is unable to extrapolate, and as a direct consequence, it is only possible to perform optimization within the span of wastewater parameters used during the training. This means that, for example, if all coagulant doses in the dataset are unnecessarily high, the estimated

coagulant dosage will be too high as well. So if the aim is to use the model to evaluate rather than predict it is important to take this limitation into account or ensure good quality of the data.

The model is only based on the historical data available and therefore a too small dataset. This could lead to overfitting where the model is good at making accurate predictions based on known data but is unable to handle unknown data (Ellis, 2021).

5.5 Model consequences

According to internal documentation, the first chemical treatment step is designed to have a mean COD removal efficiency of 57%. It can therefore be presumed that if a removal efficiency of 57 % in the first chemical treatment step is achieved it should be enough to fulfill the requirements of the permit assuming that there are no abnormal circumstances affecting the rest of the treatment process. Therefore the COD concentrations with a removal efficiency of 57% are calculated and presented in Table 10 along with the real COD concentrations in the effluent.

Table 10. Data from the test set including the predicted COD concentration, and the concentration at a 57% removal efficiency.

Sample number	COD concentration influent (mg/L)	Real COD concentration effluent (mg/L)	Predicted COD concentration (mg/L)	Calculated COD concentration (mg/L) with 57 % removal efficiency
0	1700	640	711	731
1	3000	950	1011	1290
2	2800	1000	1077	1204
3	2200	970	916	946
4	2100	980	985	903
5	3000	1100	1079	1290

As seen in the table the majority of real COD concentrations are lower than what would be required if a 57% removal efficiency is sufficient. This indicates that there could be a possibility to decrease the coagulant dosage but still fulfill the process requirements. As one of the independent variables that the model is based on is the coagulant dosage it can be used to predict what effect different coagulant dosages have on the concentration in the effluent. It is however limited by the data used for training and for the majority of the data, the coagulant dosage is the same as seen in Figure 12.

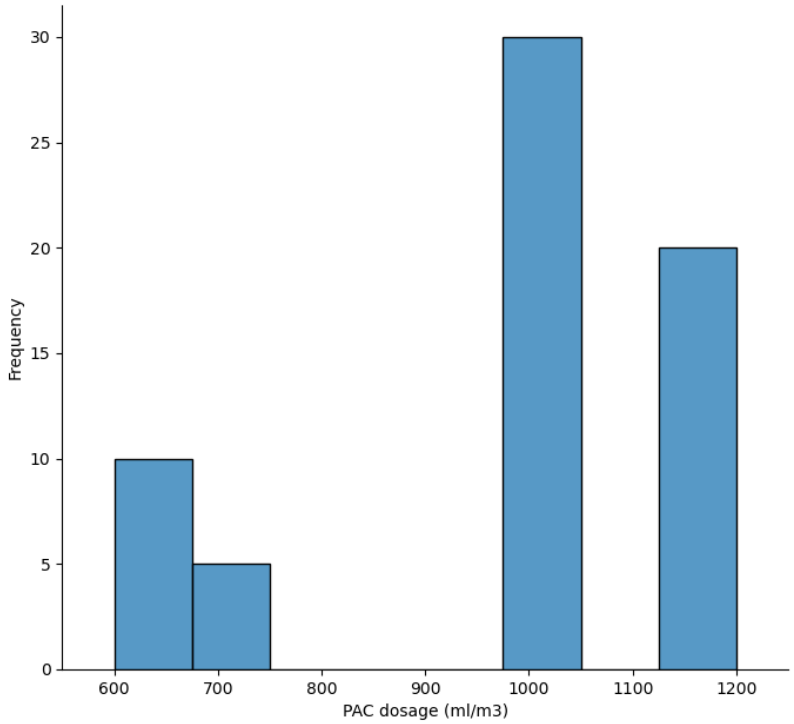

Figure 12 Histogram showing the frequency of the different PAC dosages.

As a result it is difficult to evaluate how the COD concentration in the effluent changes with the coagulant dosage from the dataset as it stands. For example evaluating a dosage between 700 to 950 ml/m^3 would require extrapolation decreasing the validity of the predictions.

It is still possible to use the model to evaluate the effect of changing the coagulant dosage but only using the same dosages that can be found in the training set making the estimations very rough. The data from the test set used for evaluating the coagulant amounts can be seen in Table 11.

Table 11. The data from the test set

Sample number	Flow rate m^3/d	Temperature °C	pH influent	PAC ml/m^3	COD influent (mg/L)	Tot-P influent (mg/L)	COD effluent (mg/L)
0	522	20.6	7.6	1200	1700	8.7	640
1	681	34.6	5.36	1000	3000	7.5	950
2	646	27	6.21	1200	2800	9.4	1000
3	542	34.5	5.64	1000	2200	6.7	970
4	583	34.1	6.23	700	2100	4.9	980
5	629	36.3	5.93	600	3000	9.2	1100

In Table 12 the effect of varying the amount of PAC can be seen. To approximate the required dosage the coagulant dosage was varied for each sample and the predicted COD concentration was compared to the COD concentration at 57% removal efficiency.

Table 12. Predicted COD concentrations in mg/L at different dosages of Ekoflock 90 (PAC), the values are with three significant figures

Sample	Predicted COD at Ekoflock 600 ml/m^3	Predicted COD at Ekoflock 700 ml/m^3	Predicted COD at Ekoflock 1000 ml/m^3	Predicted COD at Ekoflock 1200 ml/m^3	Goal COD concentration (mg/L)	Actual dosage
0	813	780	700	711	731	1200
1	1010	1010	1010	1050	1290	1000
2	1060	1050	1070	1080	1200	1200
3	930	930	916	937	946	1000
4	1000	985	967	1010	903	700
5	1080	1080	1080	1140	1290	1000

In accordance with the theory, there does not seem to be a linear relationship between the amount of coagulant and the COD concentration. However, this could also be a consequence of the small variation in dosage in the dataset making it difficult for the model to determine the impact of the amount of coagulant.
According to these results, it would have been possible to decrease the dosage for all samples except for sample 4 while still fulfilling the treatment requirements. The results also indicate that a lower dose could lead to increased removal efficiency.
The model thereby shows the potential of being a helpful tool for the operators although the model in its current state should be complemented with jar tests.

As Table 13 shows, the average amount of PAC that could be saved, according to the model, is approximately 206 L/d. As the density of Ekoflock 90 is 1370 kg/m^3 this is equal to approximately 282 kg/d.

Table 13 Result of estimation of coagulant dosage according to the model.

Sample number	Flow rate m^3/d	Actual dosage ml/m^3	Sufficient dosage according to model ml/m^3	Difference of amount of coagulant L/d
0	521,7	1200	1000	104
1	680,5	1000	600	272
2	646,4	1200	600	388
3	541,7	1000	600	217
4	583	700	700	0
5	629,2	1000	600	252

The Aluminum content of Ekoflock90 is 9.0 ± 0.3 weight%. In Table 14 the aluminum concentration at current as well as the estimated sufficient dosages is presented.

Table 14 The aluminum concentration at actual and estimated sufficient dosage

Sample number	Actual dosage ml/m^3	Aluminum concentration at actual dosage (mg/L)	Sufficient dosage according to model ml/m^3	Aluminum concentration at the estimated sufficient dosage (mg/L)
0	1200	148	1000	123
1	1000	123	600	74
2	1200	148	600	74
3	1000	123	600	74
4	700	86	700	86
5	1000	123	600	74

It is important to note that no tests or measurements have been used to validate these results so there is a very large uncertainty. So no conclusions should be drawn from these numbers,

instead, these results should be considered a suggestion of how the model might be used and what the advantages could be.

5.5.1 From an economical perspective

The cost of Ekoflock 90 according to internal documentation is approximately 3.48 SEK/kg coagulant. The possible average amount of PAC saved would mean that it could be possible to decrease the cost of coagulant by approximately 1000 SEK/day.

5.5.2 From an environmental perspective

The carbon footprint of PAC is 0.537 CO_2-eq/kg PAC including the emissions related to the transport of the coagulant (Homa and Hoffmann, 2014). According to the results from the model, saving an average of approximately 280 kg/day would then mean a 152 kg reduction of CO_2 emissions per day. The average CO_2-eq emissions from coagulants used in the WWT at Oatly is around 28 tons a month so the results from the model indicate that there could be a possibility of reducing the environmental impact from coagulants by 16% a month.

6 Conclusion

According to the results presented in the previous sections, it is possible to use the random forest algorithm to create a model that can be used for approximation of the concentration of COD in the effluent based on the influent concentration of COD, and Tot-P, temperature, pH, flow rate and dosage of coagulant.
The high r^2 value indicates that there is a strong relationship between the independent and dependent variables. This means that the wastewater parameters used for the model are well-suited for predicting the concentration of COD in the effluent. However the high RMSE value indicates that the accuracy of the predictions could be improved but depending on the requirements, it could be deemed accurate enough. In total, the results indicate that the aim of creating a model and training it using historical data to make predictions is fulfilled.

The second research question was regarding the possible consequences of using the model. By using the model to make predictions based on different coagulant dosages it was possible to estimate how much the amount of PAC could be decreased while still meeting the requirements of the WWT. The results indicate the possibility of saving approximately 280 kg of PAC/day which corresponds to saving 980 SEK/day or 152 kg CO_2-eq/day by decreasing the coagulant dosage according to the model. Although there is a high degree of uncertainty, it still illustrates how the model has the potential to be used and what there is to gain but more research is required before any conclusions should be drawn. Nonetheless, if this is accurate it would mean that it is possible to decrease the environmental impact caused by PAC by approximately 16% a month.

The final research question was what is required to improve the model. This has been previously mentioned but will be discussed more in depth in the next section of the rapport.

To conclude, the research questions presented at the beginning of the rapport are considered fulfilled, and the overall aim of the project is achieved.

7 Future recommendations

7.1 Model improvements

The largest model limitation is the size of the dataset. With more data available it would be possible to increase the training of the model which should lead to better predictions. It would also make it possible to evaluate the results to a greater extent since it would mean more test data to use when comparing predictions to real values. One way of gathering a very large amount of data could be to install more on-line sensors to automatically gather measurements of the different concentrations. As long as the data is then compiled in an excel-file it is then relatively easy to import it into Python and make it ready to use.

It is suggested that it is investigated how concentrations might vary during a day to determine if the data from the flow-proportional samples are suitable for modeling purposes or if data from single samples is required instead. This would be an important aspect to take into consideration when planning how more data can be gathered. When gathering data it is also important that it is done systematically and preferably well-documented. By establishing a structure for the systematical gathering of samples from the process it could be possible to also establish relationships between the wastewater parameters such as TOC, COD, and BOD_7. It could also make it possible to investigate if there are also great variations in the composition of the primary influent during the day. In the future this might make it possible to relate the variations to different products being produced and the results could thereby serve as a tool for production planning. In an evaluation of the WWT, it was found that the biological treatment step has trouble handling large and sudden increases in the loading rate of BOD. It could therefore be beneficial to link single products to these peaks so that the operation of the wastewater treatment process may be adjusted.

The algorithm that has been used in this project is a RandomForest regressor, but several other algorithms could have been used instead. For example, a support vector machine algorithm has the advantage of being able to extrapolate, unlike the random forest. Similarly, the random forest algorithm has several hyperparameters that for example, control the number of decision trees, how decisions should be evaluated, and many other things. These hyperparameters have not been optimized in this project but it is suggested for future work. Along with any optimization, a comparison of different algorithms should be done as well.

7.2 Areas for development of the model

As mentioned in section *6.5 Model consequences* the results from the example of how the model could be used to test different dosages of coagulant, have not been validated. It is believed that this is a plausible area of application for the model but as there is no data for how the COD concentration of the effluent varies with the coagulant dosage, there are no results to compare to. If there is an ambition to use the model in this way, it is therefore recommended that a series of jar tests are performed so that the results can be compared to the model predictions. This is currently done except that no concentrations are measured. The operators have reported that the results from the jar tests, regarding the amount of coagulant required, sometimes differ greatly from what is needed for the actual treatment process. So there is a risk that this might not be an optimal method to collect data for these comparisons however it might be worth investigating.

As mentioned in section *6.1 Final selection of variables*, the pH is currently measured by three on-line sensors located before the first coagulation, in the area for coagulation and the area for flocculation. The pH measurement that has been used for the model has been chosen just based on model performance, but another pH may have a greater impact on the coagulation efficiency. It would also be interesting to investigate the impact of the pH in a similar way as done with the coagulant dosage. This would require that another pH measurement is used as a variable as the data from the current sensor is before the pH adjustment.

According to previous research, it also seems possible to use the model to evaluate operating conditions such as stirring intensity. If this is included in the model it could also be possible to evaluate if different stirring intensities are more or less suitable for different coagulant dosages. As shortly mentioned in section *3.2.4 Mixing and motion*, the flocs' ability to re-grow is influenced by the coagulation mechanism used for the formation. This could potentially mean that if the flocs were formed using charge neutralization, which has the highest re-flocculation factor, and the change in dosage causes sweep coagulation to dominate instead, then the flocs might be more breakable. If the flocs became more breakable, then a decrease in stirring intensity might be required. This is just one example of how it could be possible to include operating conditions that are known to be important factors affecting the efficiency of the chemical treatment steps.

The only dependent variable used in this model is, as previously mentioned, the concentration of COD in the effluent. This is not the only important factor when concluding is a chemical treatment step is performing satisfactorily or not. Another important parameter that has not been included in the model is the residual turbidity which is especially important if the model is to be used for the second chemical treatment step as well. As mentioned in section *2.2.6 The post-treatment*, a coagulant is added to the drum filter if the turbidity of the filter influent is 4.5 NTU or above. This could mean that if the model is used in its current state, thereby making decisions regarding the coagulant based on the COD concentration only, there might be an increase in residual turbidity in the effluent from the second chemical treatment step. A consequence of this would then be more coagulant is used in the post-treatment step diminishing possible gains from decreasing the dosage in the second coagulation. It is therefore important to either try to include the turbidity in the model or to perform other tests to evaluate the consequence of any coagulant change.

One advantage of decreasing the coagulant dosage is a decreased amount of sludge produced during the chemical treatment steps. According to the operators, there is a noticeable difference when the dosage is changed but this is not reflected in the results from the model. To get a more comprehensive view of the consequences of using the model, it would therefore be interesting to include this aspect, either by testing if it is possible to adapt the model for this purpose, or by performing manual calculations.

Finally, as mentioned in section *3.5.3 Machine learning*, there are several ways to use machine learning to improve a wastewater treatment process. It is therefore not implausible that the model could be further developed, perhaps by combining results from different algorithms, to potentially be used to evaluate multiple steps of the process and how changes in one part affect the rest.

8 Reference list

ADEMOROTI, C.M.A. (1986). Model to predict BOD from COD values. *Model to predict BOD from COD values*, [online] 26(3-4), pp.80–84. Available at: http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=8644681 [Accessed 10 Nov. 2022].

Ahmad, A.L., Wong, S.S., Teng, T.T. and Zuhairi, A. (2008). Improvement of alum and PACl coagulation by polyacrylamides (PAMs) for the treatment of pulp and paper mill wastewater. *Chemical Engineering Journal*, [online] 137(3), pp.510–517. doi:https://doi.org/10.1016/j.cej.2007.03.088.

Aliashrafi, A., Zhang, Y., Groenewegen, H. and Peleato, N.M. (2021). A review of data-driven modelling in drinking water treatment. *Reviews in Environmental Science and Bio/Technology*, 20(4), pp.985–1009. doi:https://doi.org/10.1007/s11157-021-09592-y.

Allwright, S. (2022). *What is a good R-Squared value? (simply explained)*. [online] Stephen Allwright. Available at: https://stephenallwright.com/good-r-squared-value/.

Alwosheel, A., van Cranenburgh, S. and Chorus, C.G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, 28, pp.167–182. doi:https://doi.org/10.1016/j.jocm.2018.07.002.

Ammary, B. (2004). Nutrients requirements in biological industrial wastewater treatment. *African Journal of Biotechnology*, [online] 3(4), pp.236–238. Available at: https://tspace.library.utoronto.ca/bitstream/1807/4122/1/jb04045.pdf.

Arab, M., Akbarian, H., Gheibi, M., Akrami, M., Fathollahi-Fard, A.M., Hajiaghaei-Keshteli, M. and Tian, G. (2022). A soft-sensor for sustainable operation of coagulation and flocculation units. *Engineering Applications of Artificial Intelligence*, [online] 115, p.105315. doi:https://doi.org/10.1016/j.engappai.2022.105315.

Armstrong, M. (2022). *Which passenger cars pollute the most?* [online] World Economic Forum. Available at: https://www.weforum.org/agenda/2022/04/most-polluting-passenger-cars/.

Assmann, C., Scott, A. and Biller, D. (2017). Online total organic carbon (TOC) monitoring for water and wastewater treatment plants processes and operations optimization. *Drinking*

Water Engineering and Science, 10(2), pp.61–68. doi:https://doi.org/10.5194/dwes-10-61-2017.

Aziz, J.A. and Tebbutt, T.H.Y. (1980). Significance of COD, BOD and TOC correlations in kinetic models of biological oxidation. *Water Research*, 14(4), pp.319–324. doi:https://doi.org/10.1016/0043-1354(80)90077-9.

Bache, D.H., Johnson, C., Papavasilopoulos, E., Rasool, E. and McGilligan, F.J. (1999). Sweep coagulation: structures, mechanisms and practice. *Journal of Water Supply: Research and Technology-Aqua*, [online] 48(5), pp.201–210. doi:https://doi.org/10.2166/aqua.1999.0022.

Bajpai, P. (2018). *Chapter 15 - Environmental Impact*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780128142387000155.

Barwal, A. and Chaudhary, R. (2014). To study the performance of biocarriers in moving bed biofilm reactor (MBBR) technology and kinetics of biofilm for retrofitting the existing aerobic treatment systems: a review. *Reviews in Environmental Science and Bio/Technology*, 13(3), pp.285–299. doi:https://doi.org/10.1007/s11157-014-9333-7.

Bottero, J.Y., Cases, J.M., Fiessinger, F. and Poirier, J.E. (1980). Studies of hydrolyzed aluminum chloride solutions. 1. Nature of aluminum species and composition of aqueous solutions. *The Journal of Physical Chemistry*, 84(22), pp.2933–2939. doi:https://doi.org/10.1021/j100459a021.

Cruz, D., Pimentel, M., Russo, A. and Cabral, W. (2020). Charge Neutralization Mechanism Efficiency in Water with High Color Turbidity Ratio Using Aluminium Sulfate and Flocculation Index. *Water*, [online] 12(2), p.572. doi:https://doi.org/10.3390/w12020572.

da Rosa, J.J. and Rubio, J. (2005). The FF (flocculation–flotation) process. *Minerals Engineering*, 18(7), pp.701–707. doi:https://doi.org/10.1016/j.mineng.2004.10.010.

DeNigris, J. (2020). *Top 7 Reasons To Consider Zeta-Potential For Coagulant Dose Control - Materials Talks*. [online] Malvern Panalytical. Available at: https://www.materials-talks.com/top-7-reasons-to-consider-zeta-potential-for-coagulant-dose-control/ [Accessed 23 Nov. 2022].

Dubber, D. and Gray, N.F. (2010). Replacement of chemical oxygen demand (COD) with total organic carbon (TOC) for monitoring wastewater treatment performance to minimize

disposal of toxic analytical waste. *Journal of Environmental Science and Health, Part A*, 45(12), pp.1595–1600. doi:https://doi.org/10.1080/10934529.2010.506116.

Dürrenmatt, D.J. and Gujer, W. (2011). Data-driven modeling approaches to support wastewater treatment plant operation. *Environmental Modelling & Software*. doi:https://doi.org/10.1016/j.envsoft.2011.11.007.

Ecolab company (2018). *The NALCO Water Handbook, Fourth Edition /*. 4th ed. [online] New York, N.Y.: Mcgraw-Hill Education, pp.197, 203–204, 210–212. Available at: http://www.abfaeng.ir/Content/media/article/The%20Nalco%20Water%20Handbook%201-500_0.pdf [Accessed 18 Nov. 2022].

El bied, O., Kessler, M., Terrero, M.A., Fechtali, T., Cano, A.F. and Acosta, J.A. (2021). Turbidity and Chemical Oxygen Demand Reduction from Pig Slurry through a Coagulation Flocculation Process. *Agronomy*, 11(11), p.2158. doi:https://doi.org/10.3390/agronomy11112158.

Ellis, C. (2021). *Random forest overfitting*. [online] Crunching the Data. Available at: https://crunchingthedata.com/random-forest-overfitting/.

Franceschi, M., Girou, A., Carro-Diaz, A.M., Maurette, M.T. and Puech-Costes, E. (2002). Optimisation of the coagulation–flocculation process of raw water by optimal design method. *Water Research*, 36(14), pp.3561–3572. doi:https://doi.org/10.1016/s0043-1354(02)00066-0.

Ganesh, S. (2010). *Multivariate Linear Regression*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780080448947013506.

Gebbie, P. (2001). *64 th Annual Water Industry Engineers and Operators' Conference All Seasons International Hotel -Bendigo*. [online] Available at: https://www.wioa.org.au/conference_papers/2001/pdf/paper6.pdf.

Guida, M., Mattei, M., Della Rocca, C., Melluso, G. and Meriç, S. (2007). Optimization of alum-coagulation/flocculation for COD and TSS removal from five municipal wastewater. *Desalination*, 211(1-3), pp.113–127. doi:https://doi.org/10.1016/j.desal.2006.02.086.

Hancock, N. (2017). *Safe Drinking Water Foundation*. [online] Safe Drinking Water Foundation. Available at: https://www.safewater.org/fact-sheets-1/2017/1/23/conventional-water-treatment.

Haydar, S. and Aziz, J.A. (2009). Coagulation–flocculation studies of tannery wastewater using combination of alum with cationic and anionic polymers. *Journal of Hazardous Materials*, 168(2-3), pp.1035–1040. doi:https://doi.org/10.1016/j.jhazmat.2009.02.140.

Heddam, S., Bermad, A. and Dechemi, N. (2011). ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. *Environmental Monitoring and Assessment*, 184(4), pp.1953–1971. doi:https://doi.org/10.1007/s10661-011-2091-x.

Homa, J. and Hoffmann, E. (2014). *Life Cycle Analysis of Leading Coagulants: Executive Summary*. [online] Available at: https://www.incopa.org/wp-content/uploads/2019/02/INCOPA_LCA_Executive_Summary_web.pdf [Accessed 3 Apr. 2023].

Hu, Z. and Grasso, D. (2005). *WATER ANALYSIS | Chemical Oxygen Demand*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B0123693977006634.

Huck, P. and Sozański, M. (2011). *3.16 - Chemical Basis for Water Technology*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B978044453199500069 5#bib116 [Accessed 17 Nov. 2022].

Iwapublishing.com. (2010). *Coagulation and Flocculation in Water and Wastewater Treatment | IWA Publishing*. [online] Available at: https://www.iwapublishing.com/news/coagulation-and-flocculation-water-and-wastewater-treatment.

Jain, S.K. and Singh, V.P. (2003). *Chapter 13 - Water Quality Modeling*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/S0167564803800679 [Accessed 9 Nov. 2022].

Jarvis, P., Jefferson, B., Gregory, J. and Parsons, S.A. (2005). A review of floc strength and breakage. *Water Research*, [online] 39(14), pp.3121–3137. doi:https://doi.org/10.1016/j.watres.2005.05.022.

Johnson, I., Sithik Ali, M.A. and Kumar, M. (2019). *Chapter 10 - Cyanobacteria/Microalgae for Distillery Wastewater Treatment- Past, Present and the Future*. [online] ScienceDirect. Available at:

https://www.sciencedirect.com/science/article/pii/B9780128168097000105#bib9 [Accessed 23 Nov. 2022].

Kemira. (n.d.). *Download the updated Water Handbook!* [online] Available at: https://www.kemira.com/insights/water-handbook-2020/ [Accessed 16 Nov. 2022].

Kemmer, F.N. and Ill, B. (1988). *The NALCO water handbook*. New York: Mcgraw-Hill.

Khan, S. and Ali, J. (2018). Chemical analysis of air and water. *Bioassays*, pp.21–39. doi:https://doi.org/10.1016/b978-0-12-811861-0.00002-4.

Li, F., Jiang, J.-Q., Wu, S. and Zhang, B. (2010). Preparation and performance of a high purity poly-aluminum chloride. *Chemical Engineering Journal*, [online] 156(1), pp.64–69. doi:https://doi.org/10.1016/j.cej.2009.09.034.

Länsstyrelsen Skåne (2020). *Tillstånd enligt miljöbalken till livsmedelsproduktion, Landskrona kommun*. [online] Available at: https://www.naturvardsverket.se/contentassets/e54a650ab51a46dbbe614cfdcb077d6a/2020-12-10-mpd-skane.pdf [Accessed 7 Feb. 2023].

Mackenzie Leo Davis (2020). *Water and wastewater engineering : design principles and practice*. New York: Mcgraw-Hill.

Madan, S., Madan, R. and Hussain, A. (2022). Advancement in biological wastewater treatment using hybrid moving bed biofilm reactor (MBBR): a review. *Applied Water Science*, 12(6). doi:https://doi.org/10.1007/s13201-022-01662-y.

Majid, A. (2019). Application of Lab-Scale MBBR to Treat Industrial Wastewater using K3 Carriers: Effects of HRT, High COD Influent, and Temperature. *International Journal of Environmental Sciences & Natural Resources*, 20(2). doi:https://doi.org/10.19080/ijesnr.2019.20.556031.

Mantech (n.d.). *Potassium Dichromate, COD Analysis and the European Chemical Agency: What You Need to Know*. [online] Available at: https://mantech-inc.com/wp-content/uploads/2016/07/REACH-Dichomate-Ban-Release.pdf.

Miranda, R., Latour, I. and Blanco, A. (2020). Understanding the efficiency of aluminum coagulants used in dissolved air flotation (DAF). *Frontiers in Chemistry*, [online] 8. doi:https://doi.org/10.3389/fchem.2020.00027.

Mucha, Z. and Kułakowski, P. (2016). Turbidity measurements as a tool of monitoring and control of the SBR effluent at the small wastewater treatment plant – preliminary study. *Archives of Environmental Protection*, 42(3), pp.33–36. doi:https://doi.org/10.1515/aep-2016-0030.

Naturvårdsverket (n.d.). *Reduced Climate Impact*. [online] www.naturvardsverket.se. Available at: https://www.naturvardsverket.se/en/environmental-work/swedish-environmental-objectives/reduced-climate-impact/ [Accessed 13 Apr. 2023].

Naturvårdsverket (n.d.). *Industri, utsläpp av växthusgaser*. [online] www.naturvardsverket.se. Available at: https://www.naturvardsverket.se/data-och-statistik/klimat/vaxthusgaser-utslapp-fran-industrin/.

Naturvårdsverket (n.d.). *Sveriges utsläpp och upptag av växthusgaser*. [online] www.naturvardsverket.se. Available at: https://www.naturvardsverket.se/data-och-statistik/klimat/sveriges-utslapp-och-upptag-av-vaxthusgaser/.

Nti, S.O., Buamah, R. and Atebiya, J. (2021). Polyaluminium chloride dosing effects on coagulation performance: case study, Barekese, Ghana. *Water Practice and Technology*. doi:https://doi.org/10.2166/wpt.2021.069.

Oatly (n.d.). *Our process | Oatly*. [online] www.youtube.com. Available at: https://youtu.be/OdHRe6kgf4c [Accessed 7 Feb. 2023].

Ødegaard, H. (1995). Optimization of flocculation/flotation in chemical wastewater treatment. *Water Science and Technology*, 31(3-4), pp.73–82. doi:https://doi.org/10.2166/wst.1995.0518.

Parsons, S.A., Goslan, E.H., McGrath, S., Jarvis, P. and Jefferson, B. (2014). *2.7 - Disinfection Byproduct Control*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780123821829000840 [Accessed 17 Nov. 2022].

Pernitsky, D.J. and Edzwald, J.K. (2003). Solubility of polyaluminium coagulants. *Journal of Water Supply: Research and Technology-Aqua*, [online] 52(6), pp.395–406. doi:https://doi.org/10.2166/aqua.2003.0036.

Pernitsky, D.J. and Edzwald, J.K. (2006). Selection of alum and polyaluminum coagulants: principles and applications. *Journal of Water Supply: Research and Technology - Aqua*, 55(2), pp.121–141. doi:https://doi.org/10.2166/aqua.2006.062.

Ratnaweera, H. and Fettig, J. (2015). State of the Art of Online Monitoring and Control of the Coagulation Process. *Water*, 7(11), pp.6574–6597. doi:https://doi.org/10.3390/w7116574.

Rattanapan, C., Sawain, A., Suksaroj, T. and Suksaroj, C. (2011). Enhanced efficiency of dissolved air flotation for biodiesel wastewater treatment by acidification and coagulation processes. *Desalination*, 280(1-3), pp.370–377. doi:https://doi.org/10.1016/j.desal.2011.07.018.

Rügner, H., Schwientek, M., Beckingham, B., Kuch, B. and Grathwohl, P. (2013). Turbidity as a proxy for total suspended solids (TSS) and particle facilitated pollutant transport in catchments. *Environmental Earth Sciences*, 69(2), pp.373–380. doi:https://doi.org/10.1007/s12665-013-2307-1.

Schonlau, M. and Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*, 20(1), pp.3–29. doi:https://doi.org/10.1177/1536867x20909688.

Shi, Z., Chow, C.W.K., Fabris, R., Liu, J., Sawade, E. and Jin, B. (2022). Determination of coagulant dosages for process control using online UV-vis spectra of raw water. *Journal of Water Process Engineering*, 45, p.102526. doi:https://doi.org/10.1016/j.jwpe.2021.102526.

Tyralis, H., Papacharalampous, G. and Langousis, A. (2019). A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water*, 11(5), p.910. doi:https://doi.org/10.3390/w11050910.

Ugonabo, V.I., Ovuoraye, P.E., Chowdhury, A. and Fetahi, E. (2022). Machine learning model for the optimization and kinetics of petroleum industry effluent treatment using aluminum sulfate. *Journal of Engineering and Applied Science*, 69(1). doi:https://doi.org/10.1186/s44147-022-00164-7.

VA Ingenjörerna (2022). *Godkänd anläggning/Prestandaprov Oatly reningsverk SR200*.

Veolia Water Technologies AB (2020). *Oatly reningsverk utbildning 2020-12-03*.

Veolia Water Technologies AB (2021). *Bruksanvisning Preliminär*.

Villa, A., Fölster, J. and Kyllmar, K. (2019). Determining suspended solids and total phosphorus from turbidity: comparison of high-frequency sampling with conventional monitoring methods. *Environmental Monitoring and Assessment*, 191(10). doi:https://doi.org/10.1007/s10661-019-7775-7.

Water Science School (2018). *Turbidity and Water | U.S. Geological Survey*. [online] www.usgs.gov. Available at: https://www.usgs.gov/special-topics/water-science-school/science/turbidity-and-water.

Wei, N., Zhang, Z., Liu, D., Wu, Y., Wang, J. and Wang, Q. (2015). Coagulation behavior of polyaluminum chloride: Effects of pH and coagulant dosage. *Chinese Journal of Chemical Engineering*, 23(6), pp.1041–1046. doi:https://doi.org/10.1016/j.cjche.2015.02.003.

Woodard & Curran, Inc. , (2006). *5 - Waste Characterization*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780750679633500072.

Woolf, B.P. (2009). *Chapter 7 - Machine Learning*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780123735942000071.

World Health Organization (2017). *Guidelines for drinking-water quality, 4th edition, incorporating the 1st addendum*. [online] www.who.int. Available at: https://www.who.int/publications/i/item/9789241549950.

Yang, Z., Gao, B. and Yue, Q. (2010). Coagulation performance and residual aluminum speciation of Al2(SO4)3 and polyaluminum chloride (PAC) in Yellow River water treatment. *Chemical Engineering Journal*, 165(1), pp.122–132. doi:https://doi.org/10.1016/j.cej.2010.08.076.

Zarei Mahmudabadi, T., Ebrahimi, A.A., Eslami, H., Mokhtari, M., Salmani, M.H., Ghaneian, M.T., Mohamadzadeh, M. and Pakdaman, M. (2018). Optimization and economic evaluation of modified coagulation–flocculation process for enhanced treatment of ceramic-tile industry wastewater. *AMB Express*, 8(1). doi:https://doi.org/10.1186/s13568-018-0702-4.

Zhang, K., Achari, G., Li, H., Zargar, A. and Sadiq, R. (2013). Machine learning approaches to predict coagulant dosage in water treatment plants. *International Journal of System Assurance Engineering and Management*, 4(2), pp.205–214. doi:https://doi.org/10.1007/s13198-013-0166-5.

Zhang, Z., Jing, R., He, S., Qian, J., Zhang, K., Ma, G., Chang, X., Zhang, M. and Li, Y. (2018). Coagulation of low temperature and low turbidity water: Adjusting basicity of polyaluminum chloride (PAC) and using chitosan as coagulant aid. *Separation and Purification Technology*, [online] 206, pp.131–139. doi:https://doi.org/10.1016/j.seppur.2018.05.051.

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY
Faculty of Engineering
Department of Chemical Engineering
Water and Environmental Engineering