LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Weather Factors and *E. coli* Concentration in Barnviken, Malmö

- A Linear Mixed Model Approach

Elias Gullberg

Supervisor: Johan Larsson, Catherine Paul, Jonas Wallin

Lund University - School of Economics and Management

Department of Statistics

STAH11 - Statistics: Bachelor's Thesis – 15 ECTS

January 2023

Abstract

Barnviken in Malmö has persistently encountered challenges associated with elevated concentrations of *Escherichia coli* (*E. coli*) without a discernible source of contamination. A notable concentration peak has been identified at the outflow of the stream traversing Hammer's Park, prompting speculation regarding the stream's potential role as the origin of the contamination. In order to investigate this hypothesis, sampling of *E. coli* concentrations was conducted at various points along the stream during the spring and summer of 2023. Two models were developed to analyze the spring and summer samples separately, employing the linear mixed model. Water temperature and rainfall were incorporated as explanatory variables to ascertain whether elevated concentrations could be attributed to weather factors. The spring model yielded statistically significant estimates for both water temperature and rainfall, affirming their impact on *E. coli* concentrations. Conversely, the summer model, optimized for improved fit, excluded rainfall as a parameter, with water temperature failing to attain statistical significance. The discerned disparity in results led to the conclusion that factors influencing contamination during the summer differ from those present in the spring. Postulated explanations encompass potential leakages from proximate sanitary facilities, heightened pollution from nearby camping sites, or increased presence of seagulls in the vicinity.

Table of Contents

1. Introduction

Big bathing places in Sweden are registered as EU-bathing sites and must be controlled regularly to protect public health according to the EU Bathing Water Directive. Bathing sites that have an average of more than 200 bathers per day are considered as EU-bathing site, but places with less bathers still have the option to be registered as such. During the bathing seasons, the county administrative boards are responsible to test and inform the public about the water qualities according to EU-directives (Swedish Agency Marine and Water Management, 2018). The current EU-directive 2006/7/EC does not specify any guideline on individual samples, instead the water should be evaluated by samples taken over the last four years. When analysing individual samples, the Swedish Agency for Marine and Water Management indicates that assessments should be done according to the EU-directive 1976/160/EEG. The concentration is measured in colony-forming unit (cfu) per 100 mL, which means that the number of bacteria that is viable and able to form colonies in a 100 mL sample is measured. EU-directive 1976/160/EEG states that a *Escherichia coli* (*E. coli*) concentration under or equal to 100 cfu/100 mL is suitable, between 200 and 1000 cfu/100 mL is suitable with remarks, and above 1000 cfu/100 mL is not suitable (Swedish Agency for Marine and Water Management, 2013). If water is deemed suitable with remarks, the county is not required to advise against bathing. However, water quality should be monitored to detect any worsening of contamination. When bathing water is deemed not suitable, the county administrative boards are must inform the public about the dissuasion, and the cause of the contamination should be investigated (Swedish Agency for Marine and Water Management, 2022).

E. coli is a common bacterium, found in the intestines of both humans and animals. Even though *E. coli* is mostly considered harmless, it is a large and diverse group of bacteria with some strains causing illness. Common symptoms of *E. coli* infection are stomach cramps, diarrhea and vomiting (public health agency of sweden, 2015) (Center for Disease Control and Prevention, 2014). The testing for the presence of E. coli in bathing water aims to determine whether the water has been contaminated with faecal matter, originating from sources such as animal waste, sewage water, or runoff following rainfall. The presence of *E. coli* is also an indication that other harmful bacteria, viruses, and parasites might be present (beaches.ie, 2023).

High concentrations of *E. coli* have been a problem in Barnviken, Malmö over decades. While other bathing places in Malmö do not seem to have a contamination problem, this has been an occurring problem in Barnviken without any lead to why that might be (Westerberg, 2018). In a study from 2022 it was found that there was an unusually high *E. coli* concentration in Barnviken at the outlet of the Hammer's stream. Runoff from the stream could be the possible source of contamination. The contamination of the stream could be a consequence of overflow incidents, sewage pipe leakage, and animals habituating the vicinity. For this study, samples taken from the stream during the spring and the summer were analysed with weather factors to statistically test if there is a connection between the two. To test this, a linear mixed model was created where weather factors were used as explanatory variables for the *E. coli* concentration. We aimed to determine whether weather played a role as a contributing factor in contamination. In cases where the model lacks significance, it hints at the involvement of an alternative source contributing to the contamination (Dwite, 2023).

Previous models depicting how the growth and survival of *E. coli* are related to different factors have been developed. Wolska et al. (2022) found that some survival factors include temperatures, solar insolation, hydrologic conditions, water chemistry, nutrient conditions, suspended and settled solids, and land-use practices. Wolska et al. also list animal usage of the water as a contributing factor. Pets, such as dogs and birds, especially seagulls, are carriers of *E. coli* and can be a cause of contamination. This was also stated by Palazón et al (2017). The ground around the water area can be contaminated by birds living in the area or by precipitation and rainfall can then be expected to contribute to the *E. coli* concentration through runoff of the faeces. Both these articles found that rainfall was a major contributor to high *E. coli* contamination in bathing water.

2. Sampling and Data

2.1 Sampling

All data, except the rainfall, were obtained from the rapport *Internship Project: Investigation of Fecal Contamination in the Hammar´s stream in Malmö, Barnviken* (Dwite, 2023). Samples were collected once a week during week 17 to 22 and between week 27 and 32 in the year 2023. All samples were taken in the stream running through Hammer's Park in Malmö and every week the samples were collected from the same locations in the stream. The exact locations and the labelling numbers can be seen in *Figure* 1. Originally, only locations 1 through 7, excluding 3.5 and 5.5, were sampled. Location 5.5 was added during week 18 and location 3.5 and 8 was added during week 19. Therefore, these locations are missing some measurements. The water temperature was also measured at each location during the sampling. During the summer, there were no samples collected in locations 2 and 3.5 (Dwite, 2023).

Figure 1: A satellite picture over Barnviken and Hammar's stream in Malmö. The red markers show the locations, labelled with the location number, where the samples were collected.

At each location, triplicates of a 100 mL water sample were collected. The samples were then analysed using the Colilert-18/Quanti-Tray method, which is a well-known method for determining the cfu concentration of *E. coli*. This method follows the ISO standard specified as ISO 9308-2:2012. The principle is that most *E. coli* strains can produce an enzyme called β-glucuronidase. This enzyme is used to metabolize a substance called MUG, and the result of the metabolization is a fluorescent end-product. That the product is fluorescent means that it will emit light when exposed to UV-light. The emitted light can then be measured and is proportional to amount of *E. coli* in the sample. The enzyme β-glucuronidase is almost exclusively produced by *E. coli,* making the method very specific for detecting the bacterium (Dwite, 2023) (Moberg, 1985). The average concentrations of the triplicates were calculated and used as a dependent variable in the model.

The data for rainfall was collected from the Swedish Meteorological and Hydrological Institute's (SMHI) website. The data used provided the amount of rain that had occurred during a specific day. The measurements were taken at SMHI's weather station called Malmö A and it was assumed that the same amount of rainfall had occurred at every location in the stream. The total rainfall that had occurred seven days before the sample and up to the sampling day was added up and used as a parameter. The seven days were used since the samples were usually taken every seventh day (SMHI, 2023).

2.2 Data

Figure 2 shows the average *E. coli* concentration that was measured at each location on every sampling date during the spring. The black lines are the error bars that were calculated from the triplicates that were selected for each sampling. *Figure* 3 displays the same type of data but for the samples collected during the summer.

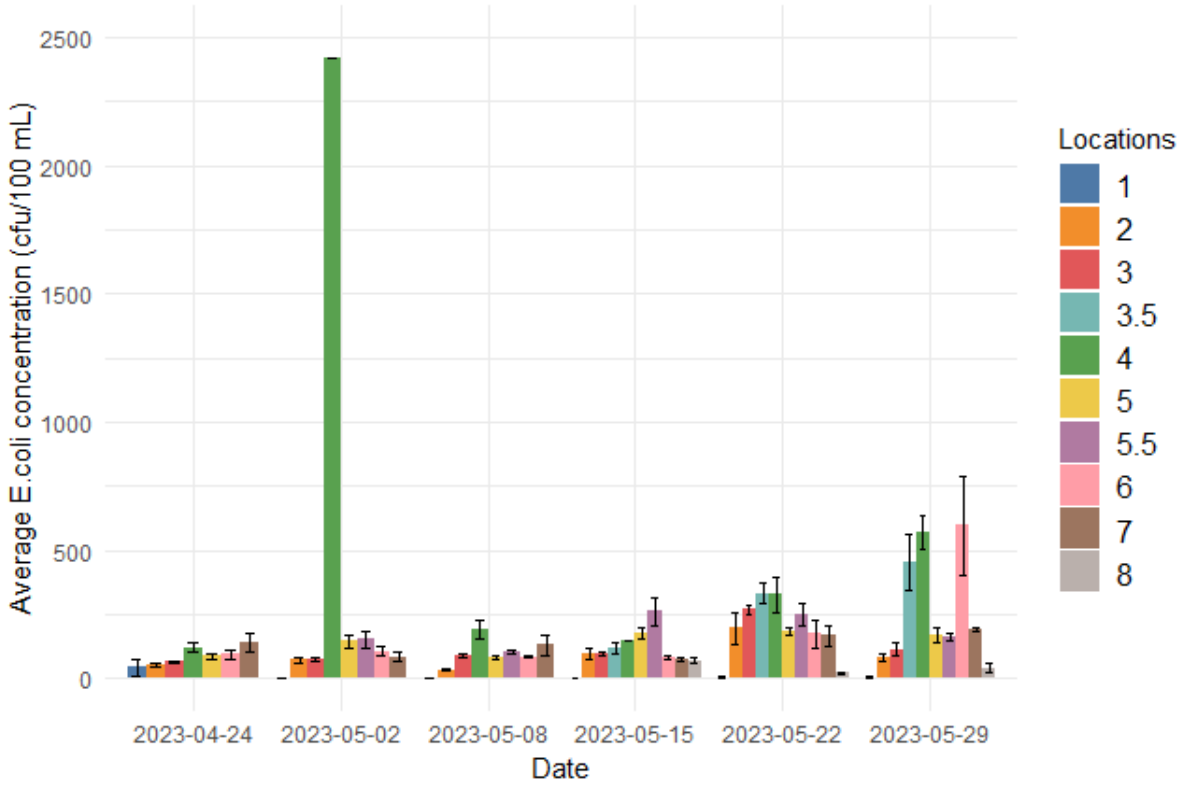

Figure 2: Bar chart over the average *E. coli* concentration that was measured at each location during the spring. On the y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the date the sample was collected. The bars' colours represent from what location the sample was collected. What location each colour represents are shown to the right. The black lines in the graph symbolize the error bars, indicating one standard deviation from the average value. The average and the standard deviation were calculated from the sampling triplicates that were collected.

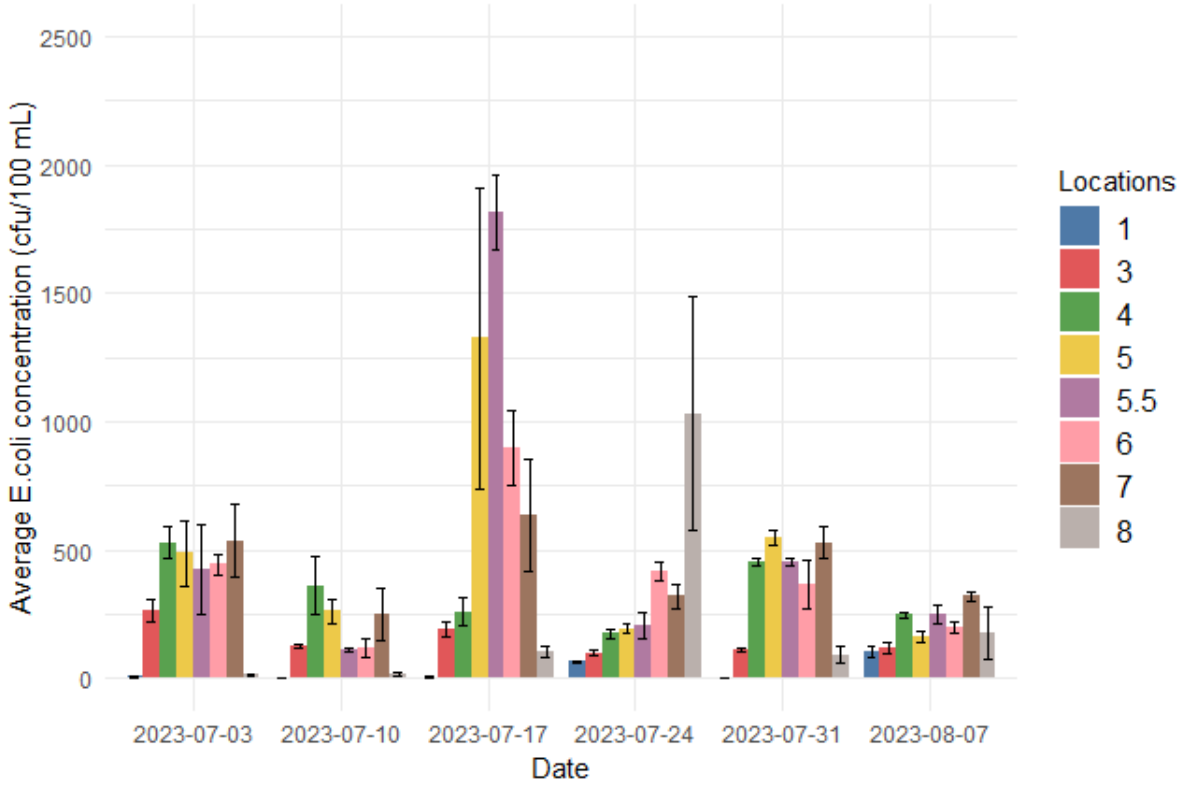

Figure 3: Bar chart over the average *E. coli* concentration that was measured at each location during the summer. On the y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the date the sample was collected. The bars' colours represent from what location the sample was collected. What location each colour represents are shown to the right. The black lines in

the graph symbolize the error bars, indicating one standard deviation from the average value. The average and the standard deviation were calculated from the sampling triplicates that were collected.

As mentioned in *Handling outliers*, the values outside of the IQR were deemed as outliers and were modified by Winsorization to the highest value inside the IQR. *Figure* 4 and *Figure* 5 show the average *E. coli* concentration in the same way as in *Figure* 2 and *Figure* 3, but after Winsorization was performed. *Figure* 4 shows the data for the spring and *Figure* 5 for the summer. The error bars are no longer displayed since the standard deviation could not be calculated for the Winsorizated values.

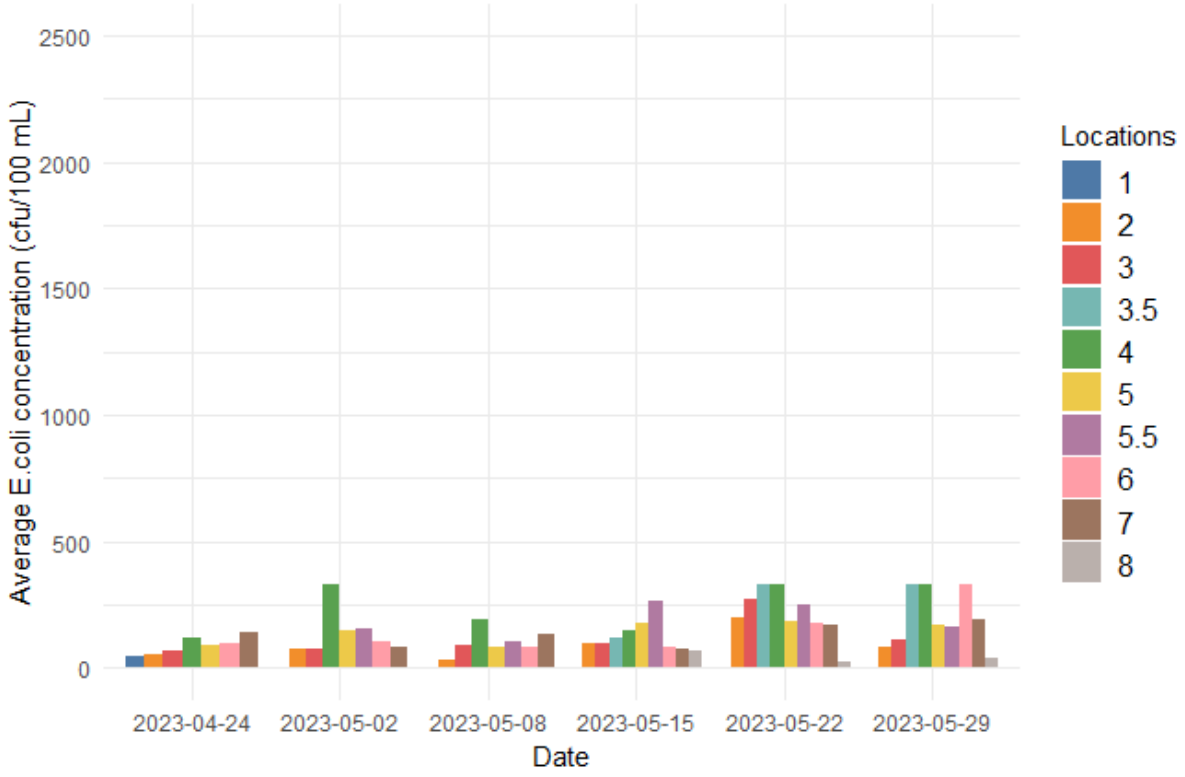

Figure 4: Bar chart over the average *E. coli* concentration that was measured at each location during the spring and after Winsorization. On the y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the date the sample was collected. The bars' colour represents from what location the sample was collected. What location each colour represents are shown to the right.

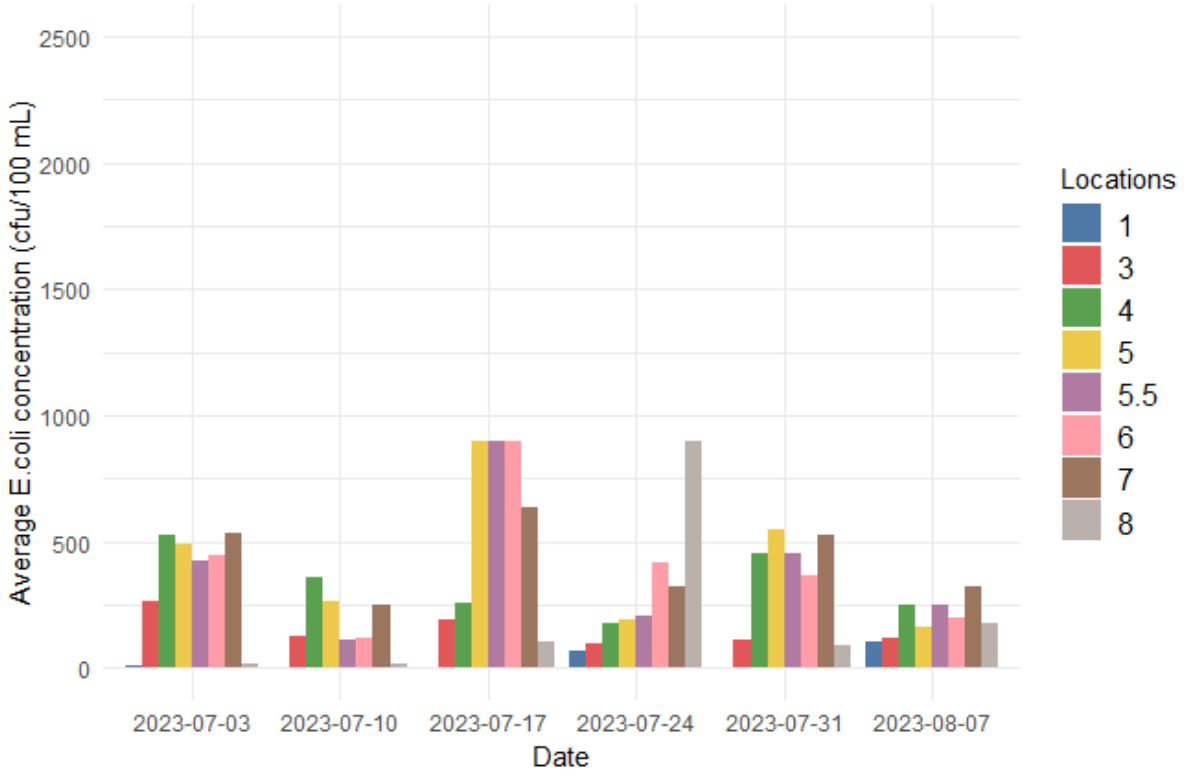

Figure 5: Bar chart over the average *E. coli* concentration that was measured at each location during the summer and after Winsorization. On the y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the date the sample was collected. The bars' colour represents from what location the sample was collected. What location each colour represents are shown to the right.

On 2023-04-24, the first sampling day during the spring, no data for the water temperature was collected. The values for this data were instead imputed by using the mice package in R. The values that were imputed by the model is displayed in *Table* 1.

Table 1: The values of the imputed water temperatures. The table shows what value was imputed in each location.

Location	1	2	3	4	5	6	7
Imputed Water Temperature (°C)	11.8	12.5	11.6	12.1	11.6	12.7	12.7

In *Figure* 6, the measured water temperatures at each location are shown for every sampling date during the spring. *Figure* 7 instead, shows the data for the water temperature that was collected during the summer in the same way.

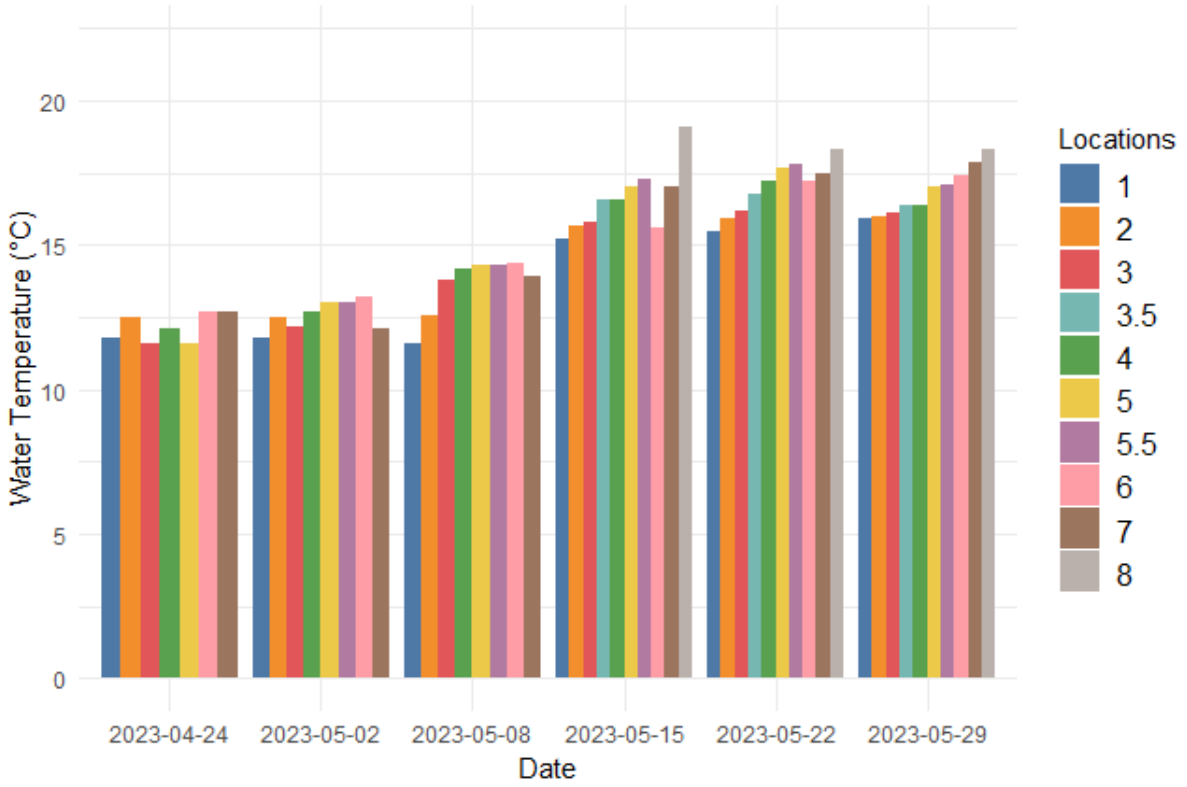

Figure 6: Bar chart that shows the measured water temperature at each location on every sampling date during the spring. On the y-axis is the water temperature in ℃ and the x-axis is the date the sample was collected. The bars' colours represent from what location the sample was collected. What location each colour represents are shown to the right.

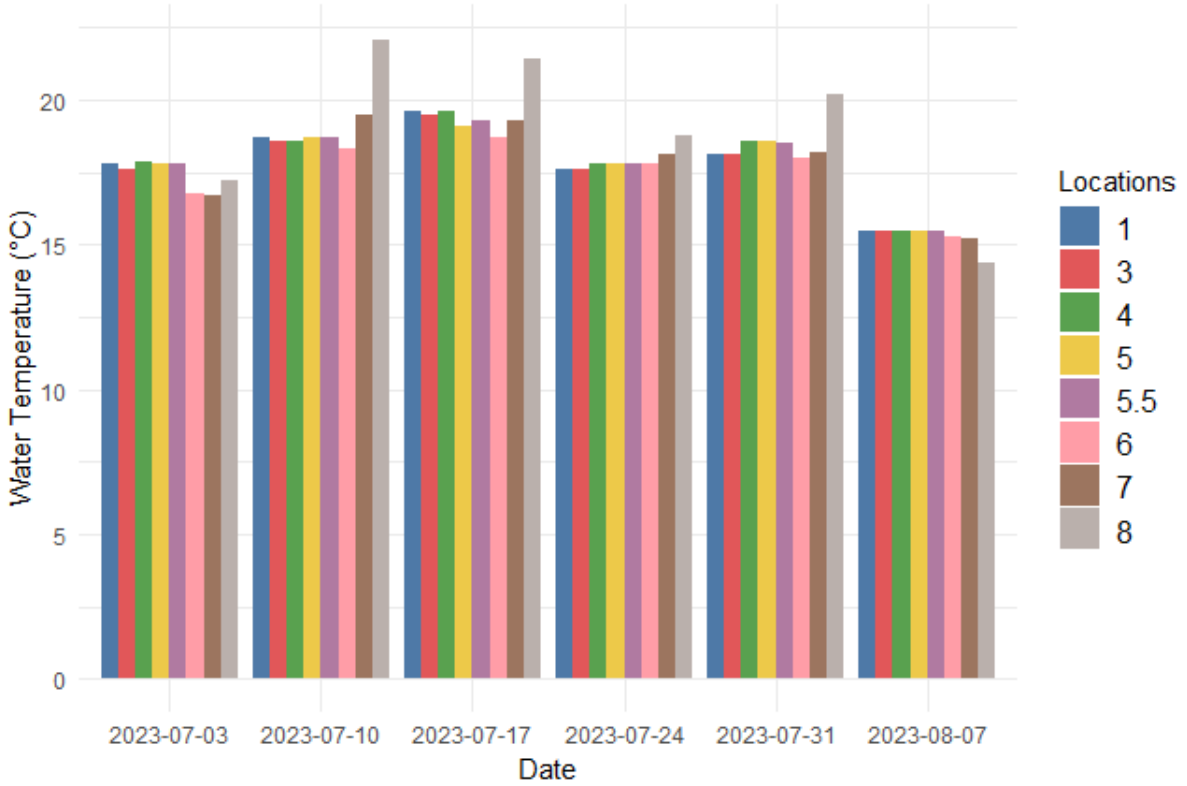

Figure 7: Bar chart that shows the measured water temperature at each location on every sampling date during the summer. On the y-axis is the water temperature in ℃ and the x-axis is the date the sample was collected. The bars' colours represent from what location the sample was collected. What location each colour represents are shown to the right.

Figure 8 and *Figure* 9 show the total rainfall that occurred during the seven days prior to when the samples were taken. The dots in the graph represent the amount of rain. *Figure* 8 shows the data for the spring and *Figure* 9 the data for the summer. Note that the rainfall is never 0 during the summer. This will later be discussed as a reason to why the spring model were showed rainfall as a significant parameter and the summer model did not.

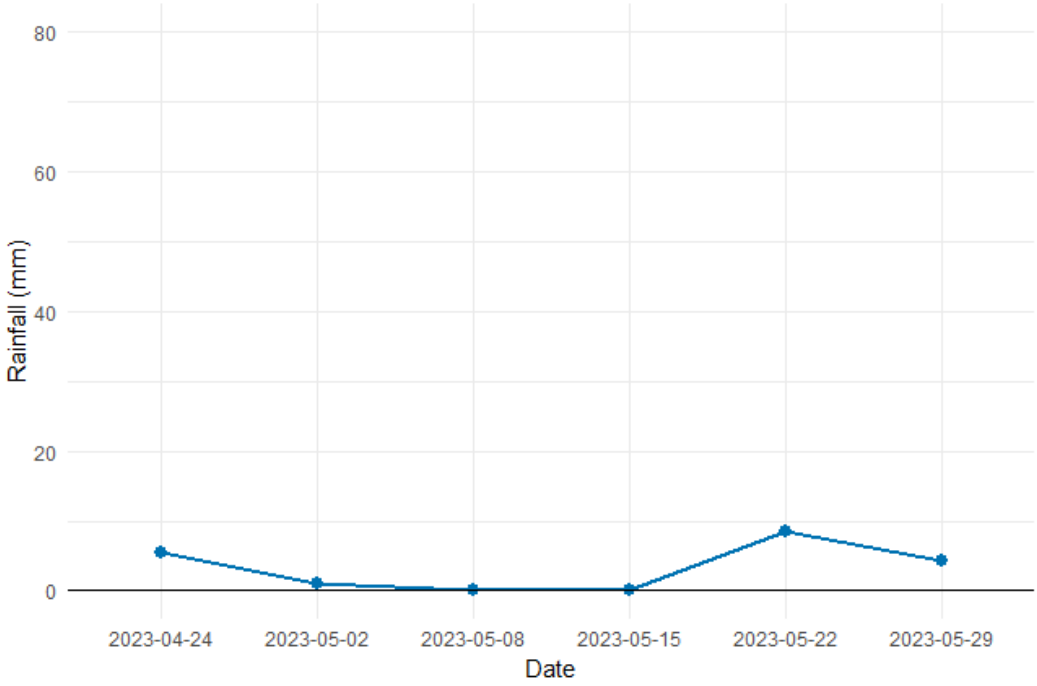

Figure 8: Time series plot over the rainfall during the spring. The y-axis is the rainfall in mm and the x-axis are the dates the samples were collected. The blue dots are the total rainfall that occurred on the seven days prior to when the samples were collected. The lines are not actual values but are instead there to help see how the rainfall changed between weeks.

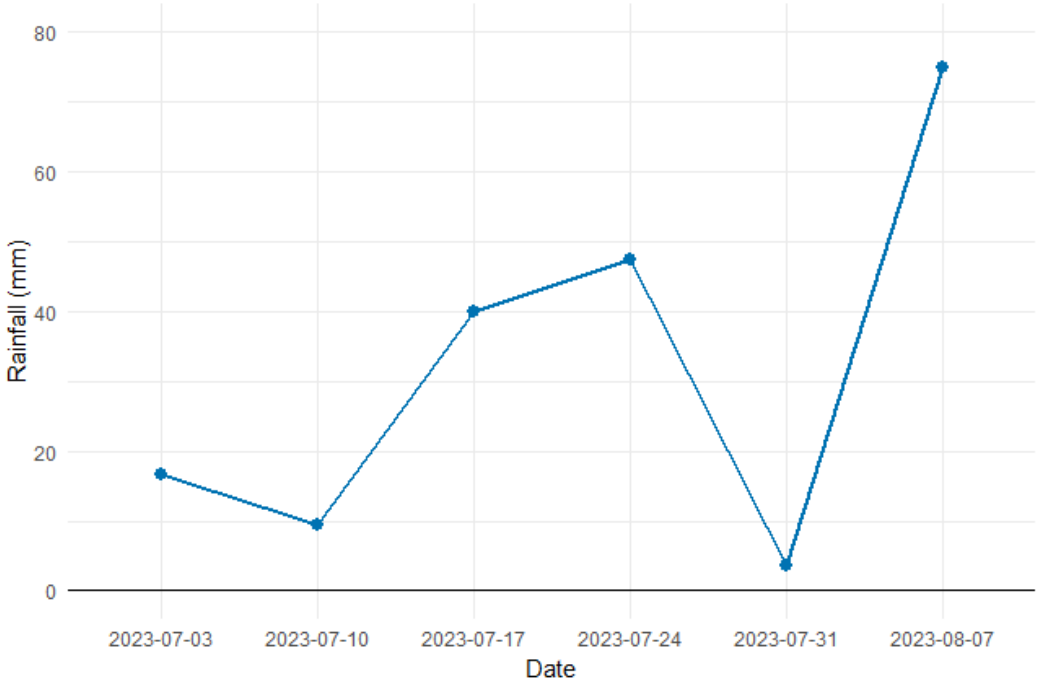

Figure 9: Time series plot over the rainfall during the summer. The y-axis is the rainfall in mm and the x-axis are the dates the samples were collected. The blue dots are the total rainfall that occurred on the seven days prior to when the samples were collected. The lines are not actual values but are instead there to help see how the rainfall changed between weeks.

2. Method

2.1 The Linear Mixed Model

In the data for this study, the samples are divided into groups depending on which location they were taken. These types of datasets are called hierarchical data and is characterized by that it involves organizing observations into nested groups or levels, enabling the analysis of patterns within and between these levels. There are many ways that one can deal with analysing hierarchical data. One way would be to aggregate the data. Instead of using all the dependent data in a group one could aggregate it by taking the average of the data. This would result in data that now is independent. The problem with this approach is that we lose a lot of valuable datapoints. At an aggregated level, there will only be as many data points as there are groups. Another approach would be to analyse each unit by itself. For example, one could run a separate linear regression for each unit. Although this approach is possible, the problem is that it creates a lot of models and none of them take advantage of the data in the other groups. The linear mixed model can be seen as a model that is in between the two approaches stated above. The individual linear regressions of units have many estimates but is "noisy" and the aggregate approach is less noisy, but we lose a lot of data points (UCLA: Statistical Consulting Group, 2021)

In this study, the linear mixed model was used to find a relationship between *E. coli* concentration and weather factors. The linear mixed model is named after the fact that the model is linear in the parameters, and that there is a mix of fixed and random effects in the explanatory variables. The fixed effects are associated with the explanatory variables of the model. These explanatory variables can be either values of continuous range or categorical factors. The fixed effects are viewed as unknown constant parameters that are related to the explanatory variables. By estimating the parameters, the relationship between the explanatory variables and the observations can be obtained. When the levels of a factor are not of intrinsic interest, the effect of that factor is modelled as a random effect. It is evident that there might be a variability between the different levels but the effect that the levels have on the observations is not of interest. The effects from the levels are then, instead of treated as parameters, represented by random unobserved variables which are assumed to follow a normal distribution. (West, et al., 2007).

The expression for a linear regression model can be seen in *Equation* 1, where y is a vector of observations, X is a matrix of known explanatory variables, β is a vector of unknown regression coefficients and ϵ is a vector of errors (Jiang & Nguyen, 2021).

$$y = X\beta + \epsilon \tag{1}$$

In this model the regression coefficients are said to be fixed, meaning they are constant for the entire population. However, when observations correlate with each other it makes more sense to assume that the coefficients are random parameters (Jiang & Nguyen, 2021). In this study, the different observations from the same locations can be assumed to correlate therefore making the linear mixed model a better fit for the data than linear regression.

The general model for a linear mixed model for the given unit i can be expressed as in *Equation* 2. In the equation, y_i is a vector of observations, X_i is a matrix of known explanatory variables, β is a column vector of unknown regression coefficients, also called fixed effects, Z_i is a design matrix for the random effects and the units, α_i is a vector of random effects and ϵ_i is a vector of residuals for y_i that are not explained by $X_i\beta + Z_i\alpha_i$ (UCLA: Statistical Consulting Group, 2021) (Jiang & Nguyen, 2021).

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i \tag{2}$$

Let us define an index t that represents the time point an observation was taken and a second index i that represents the unit the observation was taken from. Then t (t = 1, …, n_i), where n_i is the number of observations in unit i (i=1, …, m), where m is the number of units. The data contains p number of

explanatory variables X ($X^{(1)}$, …$X^{(p)}$) each of which is associated with the fixed effects β_1, …, β_p. Each β parameter represents the effect a one-unit change in the associated X explanatory variables has on y, assuming that all other explanatory variables are constant. For the explanatory variables, $X^{(1)}$, …, $X^{(p)}$ the terms $X_{t,i}^{(1)}$, …, $X_{t,i}^{(p)}$ each represent the t:th observed explanatory variables for unit i. The data also contains q number of explanatory variables Z ($Z^{(1)}$, …, $Z^{(q)}$) which are associated with the random effects $u_{1,i}$, …, $u_{q,i}$ that are specific to the unit i. $\epsilon_{t,i}$ represents the residual for the t observation from unit i (West, et al., 2007).

In *Equation* 2, y_i is a $n_i \times 1$ vector of all observed values in unit i, where N is the total number of observations, that is, all observations from every unit. The y_i vector can be written as in *Equation* 3.

$$y_i = \begin{pmatrix} y_{1,i} \\ y_{2,i} \\ \vdots \\ y_{n_i,i} \end{pmatrix} \tag{3}$$

X_i in *Equation* 2 is an $n_i \times p$ matrix of all known values of the explanatory variables in group i. The matrix can be written as in *Equation* 4. If the model includes an intercept term, then all values in the first column would be equal to 1.

$$X_i = \begin{pmatrix} X_{1,i}^{(1)} & X_{1,i}^{(2)} & \cdots & X_{1,i}^{(p)} \\ X_{2,i}^{(1)} & X_{2,i}^{(2)} & \cdots & X_{2,i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i,i}^{(1)} & X_{n_i,i}^{(2)} & \cdots & X_{n_i,i}^{(p)} \end{pmatrix} \tag{4}$$

In *Equation* 2, β is a $p \times 1$ column vector of the unknown fixed effect regression coefficients. It can be expressed as in Equation *5*.

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{5}$$

Much like the X_i matrix, the Z_i matrix also represent known values of the observed explanatory variables in group i. The difference is that they are the explanatory variables for the random effects. The Z_i matrix is an $n_i \times q$ matrix and can be expressed as in *Equation* 6.

$$Z_i = \begin{pmatrix} Z_{1,i}^{(1)} & Z_{1,i}^{(2)} & \cdots & Z_{1,i}^{(q)} \\ Z_{2,i}^{(1)} & Z_{2,i}^{(2)} & \cdots & Z_{2,i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i,i}^{(1)} & Z_{n_i,i}^{(2)} & \cdots & Z_{n_i,i}^{(q)} \end{pmatrix} \tag{6}$$

The α_i vector in *Equation* 2 is a $q \times 1$ vector that represent the random effects for each covariate in Z_i. The u_i vector can be expressed as in *Equation* 7. According to the model assumptions, the random effects should follow a normal distribution with mean 0 and a variance-covariance matrix D, $\alpha_i \sim N(0, D)$. The D matrix is represented in *Equation* 8. The diagonal from right to left is the variance of each random effect in ui. The rest of the elements are the explanatory variables between two random effects. Since there are q number of random effects, D is a $q \times q$ matrix.

$$\alpha_i = \begin{pmatrix} \alpha_{1,i} \\ \alpha_{2,i} \\ \vdots \\ \alpha_{q,i} \end{pmatrix} \tag{7}$$

$$D = Var(\alpha_i) = \begin{pmatrix} Var(\alpha_{1i}) & cov(\alpha_{1i}, \alpha_{2i}) & \cdots & cov(\alpha_{1i}, \alpha_{qi}) \\ cov(\alpha_{1i}, \alpha_{2i}) & Var(\alpha_{2i}) & \cdots & cov(\alpha_{2i}, \alpha_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\alpha_{1i}, \alpha_{qi}) & cov(\alpha_{2i}, \alpha_{qi}) & \cdots & Var(\alpha_{qi}) \end{pmatrix} \tag{8}$$

ϵ_i in *Equation* 2 is an $n_i \times 1$ vector of residuals associated with a specific observation and is shown in *Equation* 9. The residuals of observations within the same group can be correlated with each other. This is a trait of the linear mixed model that is not allowed the standard linear model. Like α_i, the residuals are assumed to follow a normal distribution. Again, the mean is 0 but has a positive definite symmetric covariance matrix, R_i. The R_i matrix is represented in *Equation* 10. It also assumed that residuals of different groups are independent from each other. The α_i vectors and ϵ_i vectors are also assumed to be independent (West, et al., 2007) (UCLA: Statistical Consulting Group, 2021) (Jiang & Nguyen, 2021).

$$\epsilon_i = \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \\ \vdots \\ \epsilon_{n_i,i} \end{pmatrix} \tag{9}$$

$$R_i = Var(\epsilon_i) = \begin{pmatrix} Var(\epsilon_{1i}) & cov(\epsilon_{1i}, \epsilon_{2i}) & \cdots & cov(\epsilon_{1i}, \epsilon_{n_i i}) \\ cov(\epsilon_{1i}, \epsilon_{2i}) & Var(\epsilon_{2i}) & \cdots & cov(\epsilon_{2i}, \epsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\epsilon_{1i}, \epsilon_{n_i i}) & cov(\epsilon_{2i}, \epsilon_{n_i i}) & \cdots & Var(\epsilon_{n_i i}) \end{pmatrix} \tag{10}$$

In the linear mixed model, it is assumed that the random effects and errors have means of zero and finite variances. Both α_i and ϵ_i are unobservable and it is assumed that they are uncorrelated. If we assume that all random effects, α, are independent and identically distributed (i.i.d.) with mean zero and a variance of σ^2 and that the residuals also are i.i.d. and have a mean zero and a variance of τ^2, then, since the random effects and the errors are uncorrelated, the correlation between two observations between the same unit can be expressed as in *Equation* 11 and two observations from different units are uncorrelated (Jiang & Nguyen, 2021).

$$r = \frac{\sigma^2}{(\sigma^2 + \tau^2)} \tag{11}$$

There are different ways the data can be structured when making a linear mixed model. The three types are clustered data, repeated-measures data, and longitudinal data. In clustered data, the dependent measure is measured only one time for each unit and the units are grouped together into clusters. For repeated-measures data, the dependent variable is measured more than once within the same group across levels of a repeated-measures factor. The factor could be time or an observation level within the unit. In longitudinal data the same unit is measured several times over a certain period. It can be hard to distinguish a data set from a repeated-measures data set and a longitudinal data set. However, when doing a linear mixed model this distinction is not essential. The only important feature is that the dependent variable is measured more than once. This is because the measurements within the same unit are likely to be correlated (West, et al., 2007).

When creating a linear mixed model, a common turnout is singular fits. Singular fits, or singularity, means that the estimated variance-covariance matrices have reduced rank, indicating that certain aspects

of the matrix were precisely estimated as zero. In simpler terms, some dimensions are treated as if they have no impact in the estimation. This is especially common when dealing with a linear mixed model that uses a large number of explanatory variables (rdrr.io, 2023).

2.2 Assessing Model Assumptions

In a linear mixed model, it is assumed that both the random effects and the residuals follow a normal distribution (West, et al., 2007). A way of verifying a normal distribution is therefore required. In this study, residual plots, Q-Q plots and the Shapiro-Wilk normality test will be used.

A residual plot is a plot where fitted values are plotted on the x-axis against residuals. The plot is used to identify problems with the model by examining patters, trends, and anomalies in the distribution of the data points. Problems that can be observed using a residual plot are heteroscedastic data, non-linearity, and outliers within the data. The distribution of the data points should preferably be equally and randomly spaced around the *x*-axis (Glen, 2023a)

A Q-Q plot, or Quantile-Quantile plot, is a scatterplot of two quantiles to test if the data comes from a certain distribution. If both sets used comes from the same distribution, the data points form a 45° line. If the data points deviate from the straight line, it can be assumed that they don't follow the distribution. However, a Q-Q plot is only a visual tool and not a certainty of a data distribution (Ford, 2015).

The Shapiro-Wilk normality test is used to test if a sample follows a normal distribution, with mean μ and variance σ^2. If we have a random sample X with n data points, $X = \{x_1, x_2, \dots x_n\}$, we test if the sample follows a normal distribution, namely $X \sim N(\mu, \sigma^2)$. The two hypotheses that we want to test are as follows:

H_0: X is normally distributed, $X \sim N(\mu, \sigma^2)$.

H_1: X is not normally distributed.

To perform the test, W is calculated according to *Equation* 12, where, $\bar{x}$ is the mean value of X and a_i are constants that are calculated from *Equation* 13. V represents the covariance matrix of the order statistics and $m = (m_1, m_2, \dots, m_n)^T$ represents the expected values of the order statistics, characterizing independent and identically distributed random variables that follow the standard normal distribution $N(0, 1)$ (Ramachandran & Tsokos, 2021).

$$W = \frac{\left(\sum_{i=1}^{n} a_i \, x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

$$(a_1, a_2, \dots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} m}} \tag{13}$$

2.3 Information Criteria

When it comes to comparing the fits of different models, two widely used criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria are grounded in different assumptions and serve distinct purposes. AIC, rooted in information theory, aims to produce a probability distribution that exhibits the least divergence from the true distribution, making it well-suited for model selection. On the other hand, BIC relies on large sample asymptotic approximation, making it especially useful in situations where sample size is substantial. In both cases, the goal is to identify the model with the lowest index, reflecting its ability to balance model complexity and fit to the data. AIC is defined as in *Equation* 14 and BIC is defined as in *Equation* 15. L is the likelihood, K the number of parameters in the model, and n is the number of observations ($n = \sum_i^m n_i$) (Yang & Yang, 2014) (Busemeyer & Diederich, 2014).

$$\mathrm{AIC} = -2\log L + 2K \tag{14}$$

$$BIC = -2\log L + K\log n \tag{15}$$

2.4 Handling Influential Values

All observations used to estimate a model have some kind of influence on the regression parameters. The power of influence can indeed vary between observations. Some observations may have a stronger impact on the regression model than others depending on its relation to its fitted value. Some observations might have such an overly influence on the regression parameters that the choice to include or exclude them can alter the outcome of the model's estimates. Influential observations cannot always be detected by analysing residuals. Outliers can be overly influential on the regression but that is not always the case. Since the influential observations pulls the regression line closer to itself, in some cases the influence is so strong that the regression line settles close enough to the observation for it to no longer be recognised as an outlier. A method for detecting influential observations is of the essence (Nieuwenhuis, et al., 2012). For this study, Cook's distance was used to detect overly influential observations.

Cook's distance is calculated by assessing the impact on all the values in a regression model when the j:th observation is systematically removed from the model. The formula for calculating Cook's distance is shown in *Equation* 16 where C_j is the Cook's distance, $\hat{y}$ is the fitted values, $\hat{y}_{(-j)}$ is the fitted values without observation j, K is the number of design parameters, excluding the intercept, and R_i^{-1} is the inverse of the variance vector (Nieuwenhuis, et al., 2012).

$$C_j = \frac{1}{K+1}\left(\hat{y} - \hat{y}_{(-j)}\right)' R_i^{-1}\left(\hat{y} - \hat{y}_{(-j)}\right) \tag{16}$$

There is no threshold value for when an observation is considered too influential, but a rule of thumb is that for any value above $4/n$, where n is the number of observations used in the model, the observations should be investigated (Nieuwenhuis, et al., 2012) (Altman & Krzywinski, 2016). After the influential observations have been identified, it is to be decided how the values should be dealt with. Altman and Krzywiniski (2016) suggest that for data with limited observations, more samples within the same population as the influential observation should be taken. When that is not a possibility, like in this study, the observations can instead be deleted, and the model re-evaluated.

2.5 Handling Outliers

Because the data have relatively few observations, it was important to handle the outliers since they would have a large influence on the model. For this study the outliers were handled by Winsorization, which is a strategy that can improve effectiveness and robustness of statistical interference. The downside to this is that is introduces a bias to the model. However, this bias is still less than it would be to just delete the outliers (Glen, 2023b). The outliers were identified by looking at values outside of the interquartile range (IQR) and then replacing them with the highest value of the observations inside of the IQR.

2.6 Handling missing data

During the first sampling period, there were seven samples, taken during the first sampling day of 2023-04-24, that were missing a measurement for water temperature. New data for the temperature was generated by imputation by using the mice: Multivariate Imputation by Chained Equations (mice) package in R.

The mice package is used to create multiple imputations for multivariate missing data. The incomplete variables are imputed by a method which is based on Fully Conditional Specification, where each

missing value is imputed by a separate imputation model (RDocumentation, 2023). The mice package can perform many different imputation methods, but for this study Predictive Mean Matching (PMM) was used. PMM is similar to the regression model in its approach, but instead of predicting missing values based solely on regression, it introduces randomness. For each missing value, it selects a value randomly from observed donor values. These donors are chosen from observations whose regression-predicted values closely match the regression-predicted value for the missing data in the simulated regression model. PMM makes sure that the imputed values are reasonable which the regression method does not if the normality assumption is violated (UCLA: Statistical Consulting Group, 2021). When imputing the water temperatures, the donor values used were the air temperature that was measured, not only on the sampling day but, around the same timepoint the sample was collected.

4. Modelling

4.1 Spring Model

The explanatory variables that will be used in the model are water temperature, rainfall, sampling date and location. It's important to consider what explanatory variables that should be modelled as fixed effects and random effects. The observations are grouped by location which will therefore be modelled as a random effect. Both water temperature and rainfall were modelled as fixed effects. A fixed effect is assumed to have the same effect in all units. *Figure* 10 shows that the effect water temperature had on the concentration is similar in all locations. *Figure* 11 shows a similar trend for the rainfall. Modelling them as fixed effects is also necessary to be able to do predictions for the model. Because the data measured is longitudinal, the data explanatory variables can be modelled as either a fixed or random effect depending on the objective of the model. To address the variability of concentration between dates without specifically exploring the influence of time on concentration levels, the date was modelled as a random effect.

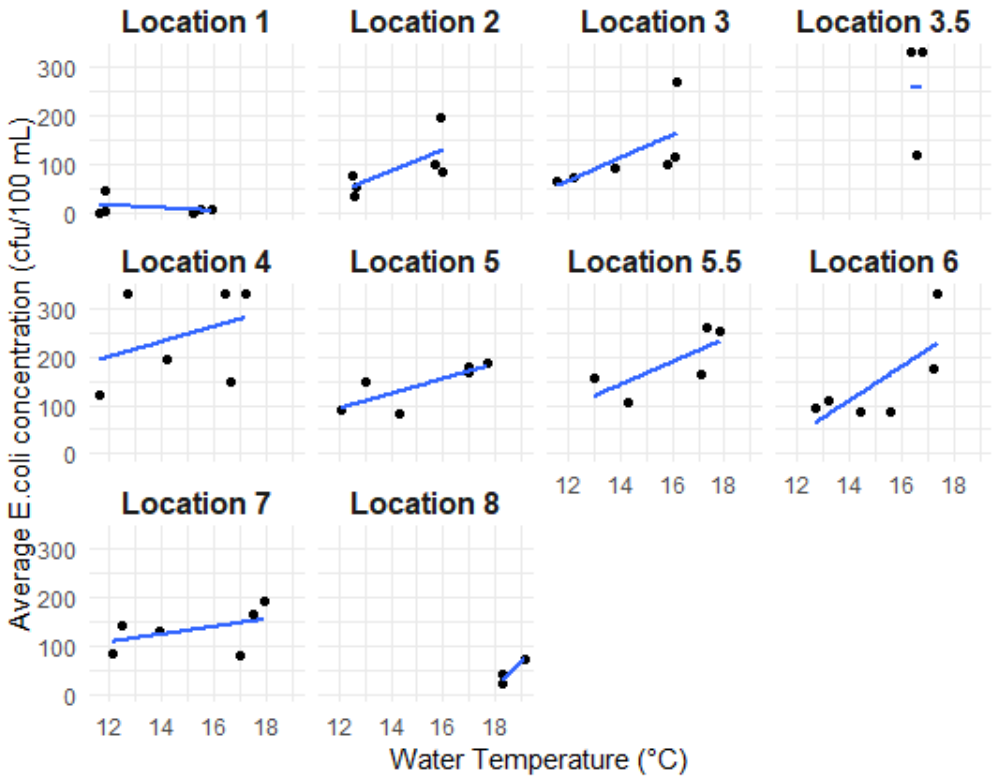

Figure 10: Scatter plots on how the water temperature affects the *E. coli* concentration. The y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the water temperature in °C. The figure is divided into ten different graphs where each represents one location. The black dots are data points for samples collected during the spring, and the blue lines are linear regression fits between water temperature and *E. coli* concentration.

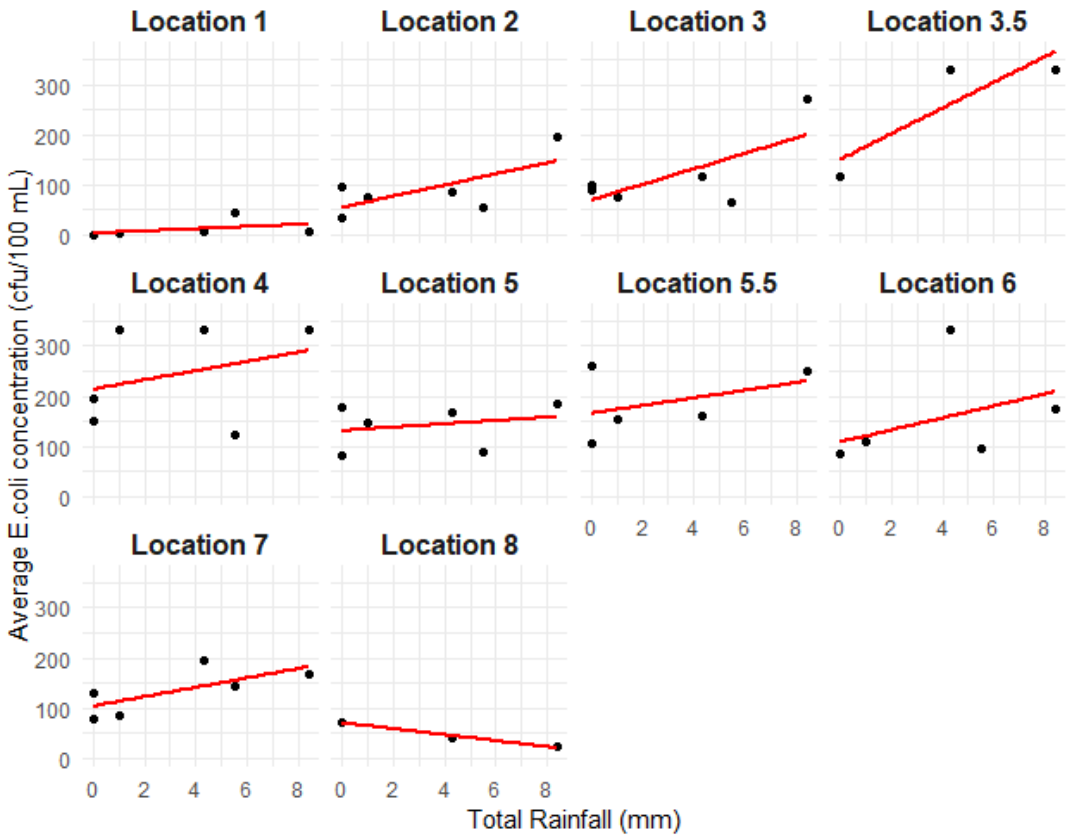

Figure 11: Scatter plots on how the rainfall affects the *E. coli* concentration. The y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the rainfall in mm. The figure is divided into ten different graphs where each represents one location. The black dots are data points for samples collected during the spring, and the red lines are linear regression fits between rainfall and *E. coli* concentration.

The two models that were considered are shown in *Table* 2. The difference between them is that Model A uses date as a random effect and Model B does not. Looking at AIC and BIC it shows that Model B is considered a better fit. This is probably because there is little variation between the locations for each date. Model B was used for later analyses.

Table 2: The different models that were tested and what the explanatory variables was used in each model and if they were used as a fixed effect or random effect. The AIC and BIC values guide model selection by balancing fit and complexity. Model B was selected as the model with best fit since the AIC and BIC values were the lowest.

Model	Fixed effects	Random effects	AIC	BIC
A	Water temperature + Rainfall	Location + Date	612.6	624.4
B	Water temperature + Rainfall	Location	610.6	620.4

4.2 Summer Model

The same explanatory variables were used for the summer model as in the spring model. Water temperature and rainfall were modelled as fixed effects and location and date as random effects. *Figure*

12 shows the effect water temperature had on the concentration and *Figure* 13 shows the effect of rainfall. Looking at both *Figure* 12 and 13, the parameters do not clearly have the same effect in each location. Compare this to *Figure* 10 and 11, where the effect was clear for both parameters. This may suggest that water temperature and rainfall would be better fitted as random effects. This is later indicated to be correct in *Table* 3. However, since the random effects are only representing variations between locations, they cannot be used to make predictions from the model. Therefore, is would be meaningless to model all explanatory variables as random effects since it would not yield a useful result. Water temperature and rainfall were therefore modelled as fixed effects, with the consideration that this might not be the best model for the data. Modelling one of them as a random effect was tested but gave a model that had a singular fit.

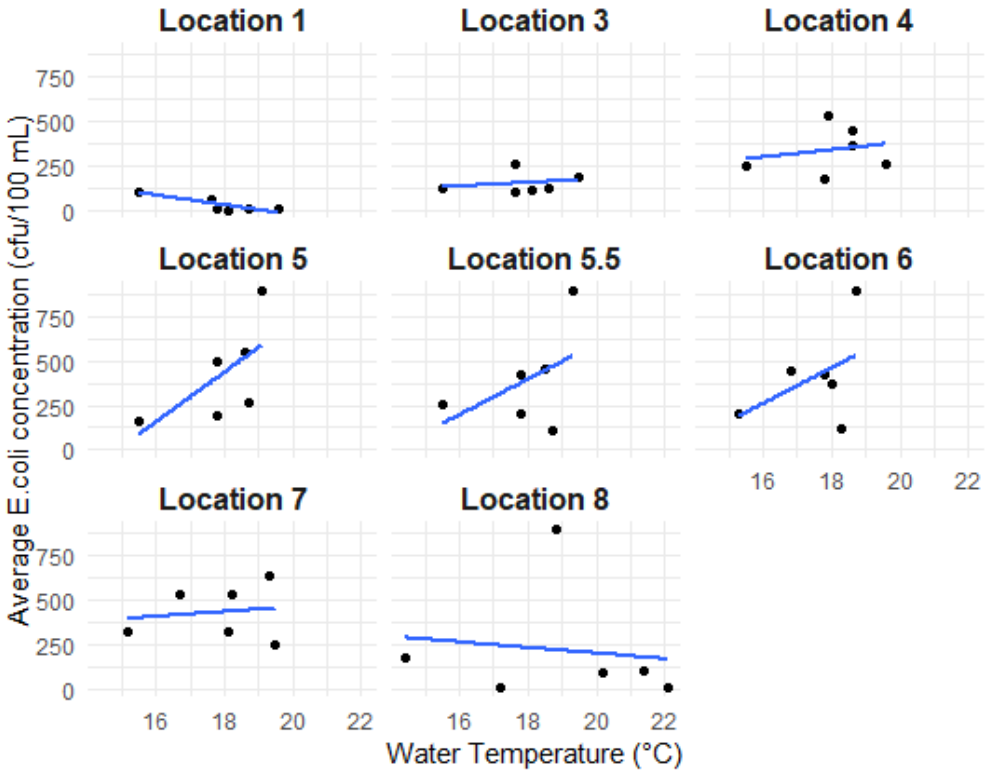

Figure 12: Scatter plots on how the water temperature affects the *E. coli* concentration. The y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the water temperature in °C. The figure is divided into eight different graphs where each represents one location. The black dots are data points for samples collected during the summer, and the blue lines are linear regression fits between water temperature and *E. coli* concentration.

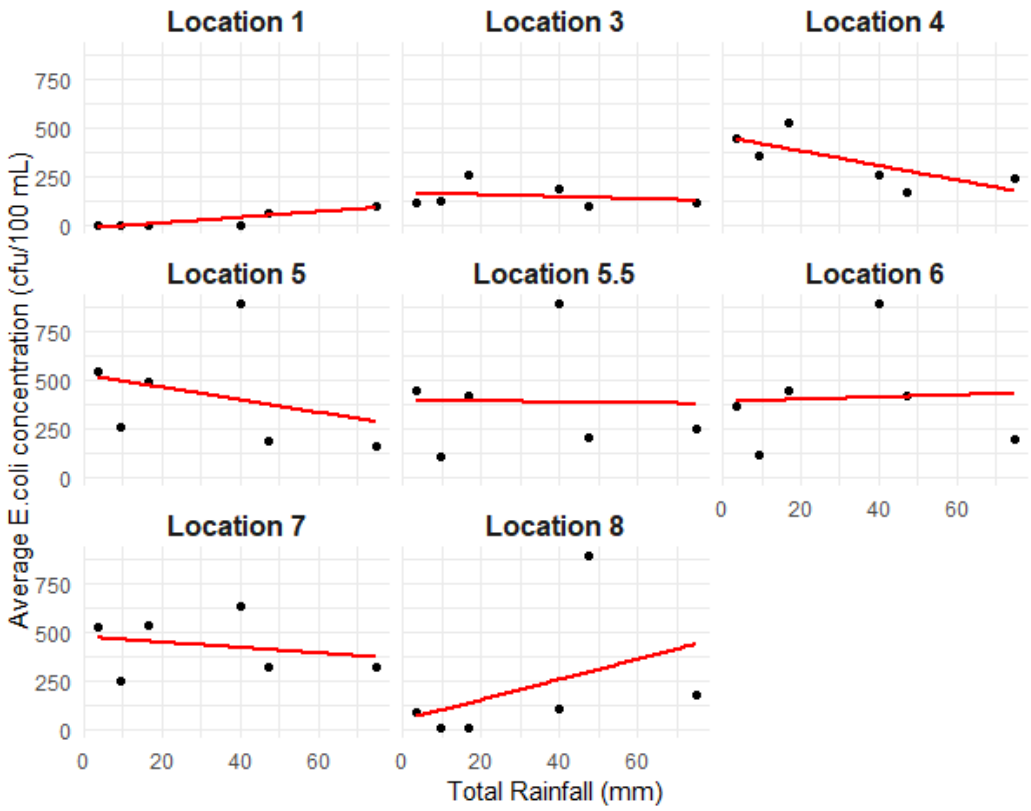

Figure 13: Scatter plots on how the rainfall affects the *E. coli* concentration. The y-axis is the average *E. coli* concentration in cfu/100 mL and the x-axis is the rainfall in mm. The figure is divided into eight different graphs where each represents one location. The black dots are data points for samples collected during the summer, and the red lines are linear regression fits between rainfall and *E. coli* concentration.

Table 3 shows the result of the summer models. Looking at the values for AIC and BIC, Model C was the best model. However, Model C is modelled without fixed effects and has every explanatory variable as a random effect but as stated before, this model would not yield a useful model. Therefore, Model D was considered the best model according to the AIC and BIC values. The model does not have rainfall as a fixed effect like the spring model has. The reason for this model having a better fit than Model B, that was used for the spring model, might be that the rainfall seems to have a very little effect on the concentration during the summer (see *Figure* 13).

Table 3: The different models that were tested and what the explanatory variables was used in each model and if they were used as a fixed effect or random effect. The AIC and BIC values guide model selection by balancing fit and complexity. Model C was identified as having the best fit, however Model D was later used since Model C contains no fixed effect.

Model	Fixed effects	Random effects	AIC	BIC
Model A	Water temperature + Rainfall	Location + Date	671.3	682.5
Model B	Water temperature + Rainfall	Location	670.5	679.9
Model C	-	Water temperature + Rainfall + Location	663.0	672.4
Model D	Water temperature	Location	668.8	676.6

5. Results

5.1 Spring Model

In the spring model, water temperature and rainfall were used as fixed effects. The results for estimating the coefficients can be seen in *Table* 4. The estimated value for water temperature was 12.62. The t-value was 2.963 with a p-value of 0.00478, suggesting that water temperature is a significant predictor for the concentration. The estimated value for rainfall was 6.65, and the t-value was 2.579 with a p-value of 0.0134 pointing to statistical significance.

Table 4: Estimates of the fixed effects in the spring model. The table shows the estimation of each fixed effect in the model, as well as a 95% confidence interval of that variable. The t-value and p-value come from performing a t-test on the fixed effects significance. Both variables were found to be statistically significant.

Parameter	Estimate	95% confidence interval	t-value	p-value
Water Temperature	12.62	4.07 – 21.12	2.963	0.00478
Rainfall	6.65	1.45 – 11.83	2.579	0.0134

The plots in *Figure* 14 show the marginal predictions for water temperature and in *Figure* 15, the marginal predictions for rainfall. The graphs are divided to display the model in each location. Note that the intercept is different in each location which is a result from the random effect in the model. Performing the Shapiro-Wilks normality test on the random effect gave $W = 0.947$ and a p-value of 0.637, suggesting that the random effects follow a normal distribution which is a model assumption.

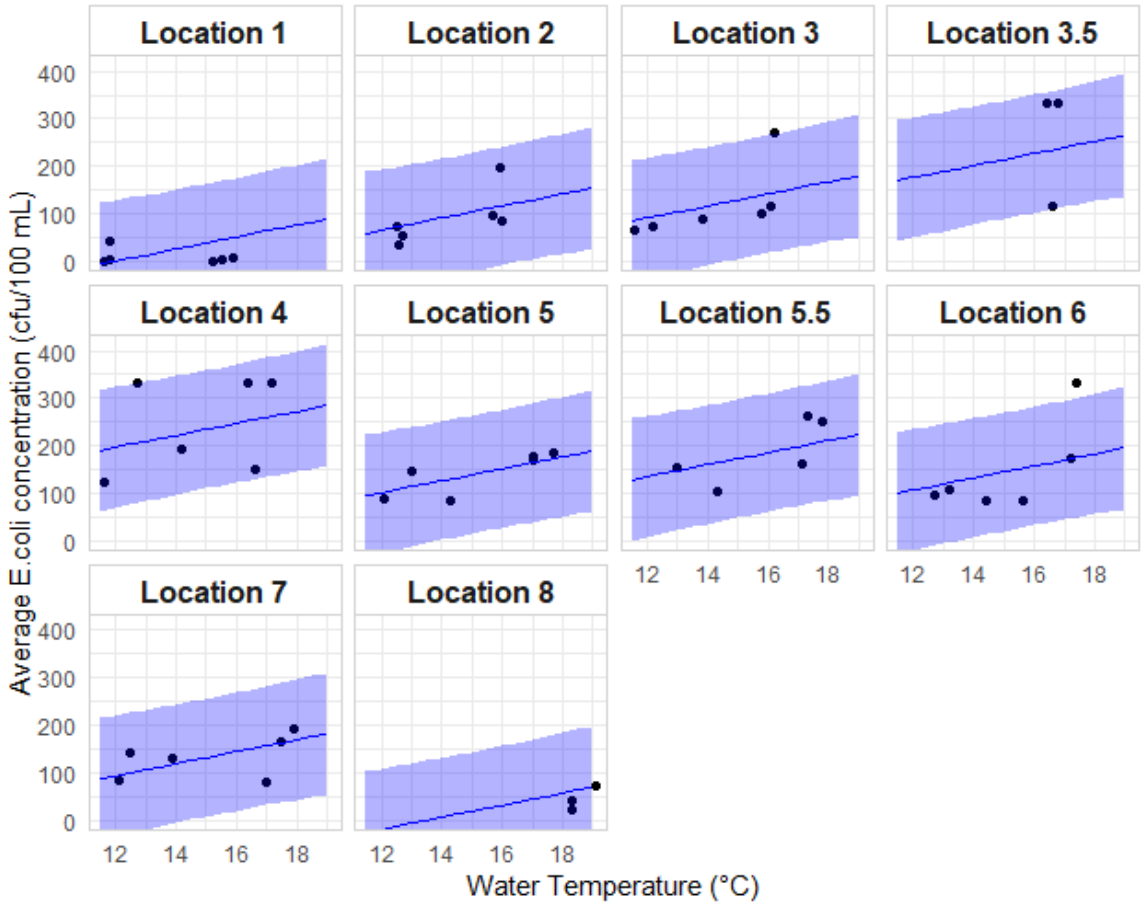

Figure 14: The predictions for water temperature during the spring. The plots show the relationship between water temperature and *E. coli* concentration in each location. The y-axis is the *E. coli* concentration in cfu/100 mL and the x-axis is the water temperature °C. The black dots are data points. The blue lines are the model predictions, used to explain the concentration from water temperature. The blue ribbons signify the confidence intervals around the predicted values. Each plot has different intercepts, which is a result of random effects.

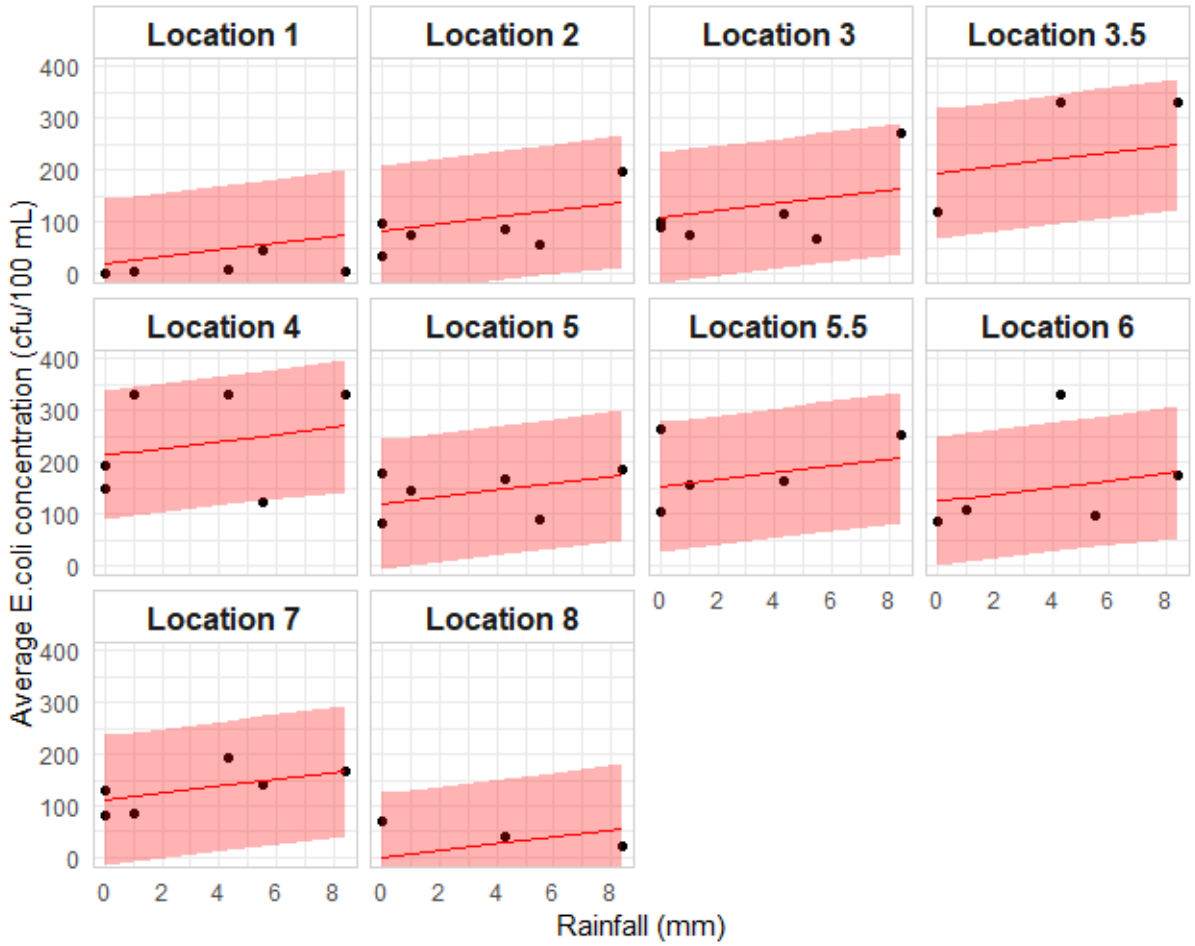

Figure 15: The predictions for rainfall during the spring. The plots show the relationship between rainfall and *E. coli* concentration in each location. The y-axis is the *E. coli* concentration in cfu/100 mL and the x-axis is the rainfall in mm. The black dots are data points. The red lines are the model predictions, used to explain the concentration from rainfall. The red ribbons signify the confidence intervals around the predicted values. Each plot has different intercepts, which is a result of random effects.

In *Figure* 16a, the residual plot indicates the presence of heteroskedasticity since the residuals have a bigger variance for higher fitted values. This goes against one of the model assumptions that says that the residuals should be evenly and randomly distributed. Since the variance is bigger at higher values, it suggests that the model might not be less effective at predicting higher *E. coli* concentrations. The confidence intervals in *Figure* 14 and 15 might therefore not have correct coverage over the prediction. The same assumption can be identified when looking at the Q-Q plot in *Figure* 16b. The points in the plot follows the theoretical quantiles except, but higher values seems to deviate from the trend. The Shapiro-Wilk normality test provided W = 0.9645 and a p-value of 0.1158. The Null hypothesis can be

accepted, affirming that the residuals come from a normal distribution, even despite the heteroskedasticity.

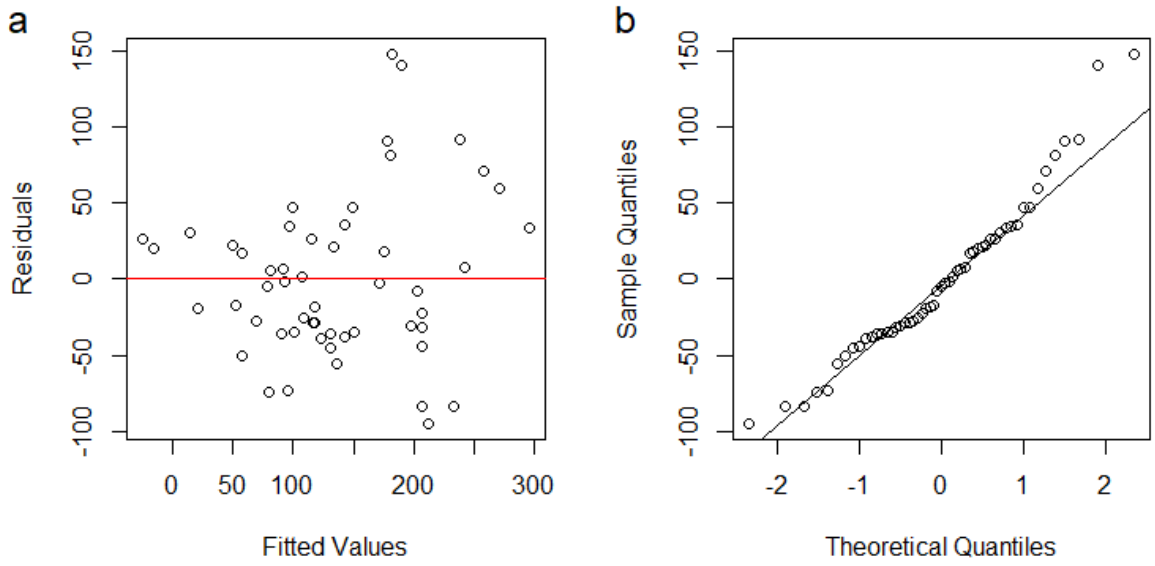

Figure 16: The residual plot and Q-Q-plot for the model. **(a)** The fitted values are plotted against the residuals to analyse the goodness of fit by noticing trends in the residual distribution. The residual appears to be heteroscedastic, having a bigger variance for higher values. The y-axis is the residuals, and the x-axis is the fitted value. The red line represents where the residual is zero and a well-fitted model would ideally have residuals evenly distributed around this line. **(b)** The Q-Q quantiles of the dataset is plotted against the expected quantiles of a theoretical distribution. The y-axis is the sample quantiles, and the x-axis the theoretical quantiles. The diagonal line in a Q-Q plot represents an ideal match between observed and expected quantiles.

Figure 17 shows Cook's distances for all samples in the first model. The threshold value was calculated to 0.0755, and every value above that was identified as being overly influential and was analysed further. The sample ID:s of the observations identified as overly influential was 5, 10, 26 and 52. The values of these observations are displayed in *Table* 5.

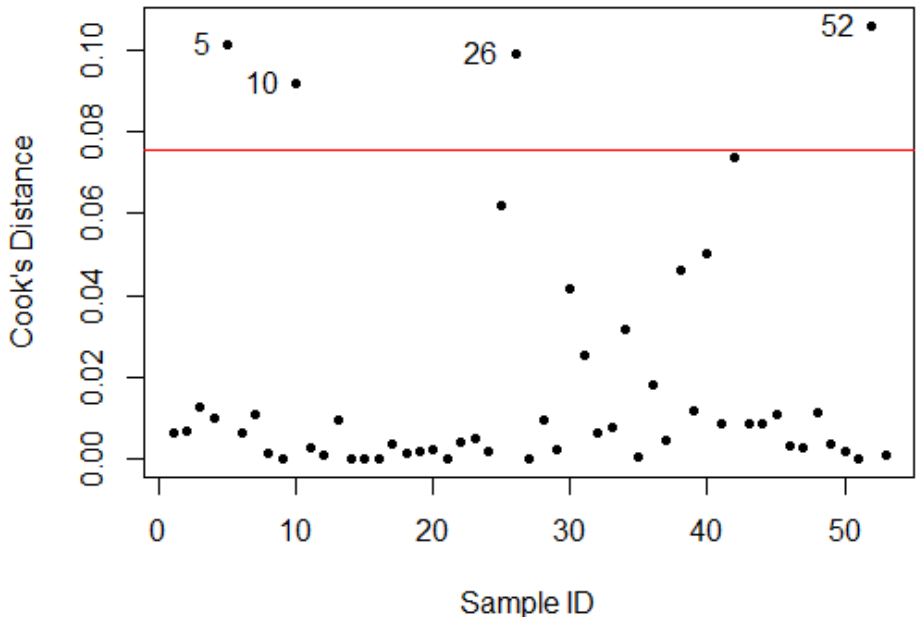

Figure 17: The graph shows every sample's calculated Cook's distance. The y-axis is the Cook's distance and on the x-axis are the sample ID:s. The black dots are the Cook's distance for respective sample. The red line is the threshold value, where the values above are considered overly influential. All samples that have a value higher than the threshold value are labelled with their respective sample ID.

Table 5: The values of the samples that were found to be overly influential. The table shows each sample's ID, date, location, average *E. coli* concentration, water temperature and rainfall.

Sample ID	Sampling Date	Location	Average *E. coli* concentration (cfu/100 ml)	Water temperature (C°)	Rainfall (mm)
5	2023-04-24	4	122.8	11.6	5.5
10	2023-05-02	4	330.8	12.7	1.0
26	2023-05-15	3.5	117.3	16.6	0.0
52	2023-05-29	6	330.8	17.4	4.3

5.2 Summer Model

In the summer model, water temperature had an estimate of 18.86 which is higher than the estimate of spring model where it was 12.34. However, unlike the spring model, the t-value was 0.908 and the p-value 0.369, meaning that the parameter is not statistically significant. This can also be seen in *Figure* 12 where the water temperature seems to have different effects in each location, pointing to that there probably are not a true parameter for water temperature that can explain the concentration.

The plot in *Figure* 18 shows the model predictions for the *E. coli* concentration at different water temperature. The random effects were found to exhibit a normal distribution after doing a Shapiro-Wilk normality test, where W = 0.856 and the p-value was 0.109.

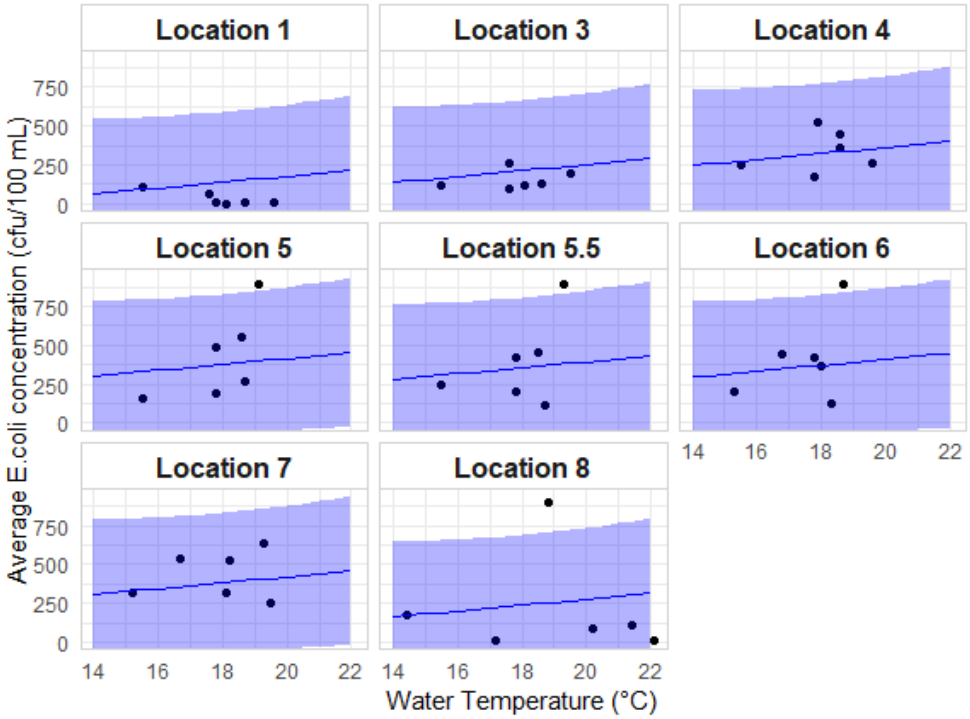

Figure 18: The prediction marginal for water temperature during the summer. The plots show the relationship between water temperature and *E. coli* concentration in each location. The y-axis is the *E. coli* concentration in cfu/100 mL and the x-axis is the water temperature °C. The black dots are data points. The blue lines are the model predictions, used to explain the concentration from water temperature. The blue ribbons signify the confidence intervals around the predicted values, showing the likely range of actual values. Each plots have different intercepts, which is a result of random effects.

In *Figure* 19a, is the residual plot for the model. The residuals show a random and even distribution of residuals expect for four residuals. In *Figure* 19b the same observation can be seen again with the four residuals sticking out. The Shapiro-Wilk normality test gave $W = 0.8796$ and a p-value = 0.00017. The null hypothesis was rejected and there is no statistical evidence that the residuals followed a normal distribution. This was also evident when looking at *Figure* 19A and *Figure* 19B, where the four residual clearly disrupts the depart from the distribution. The values for the identified samples can be seen in *Table* 6. The values of these samples are discussed later in Model Assumptions and linearity assumptions.

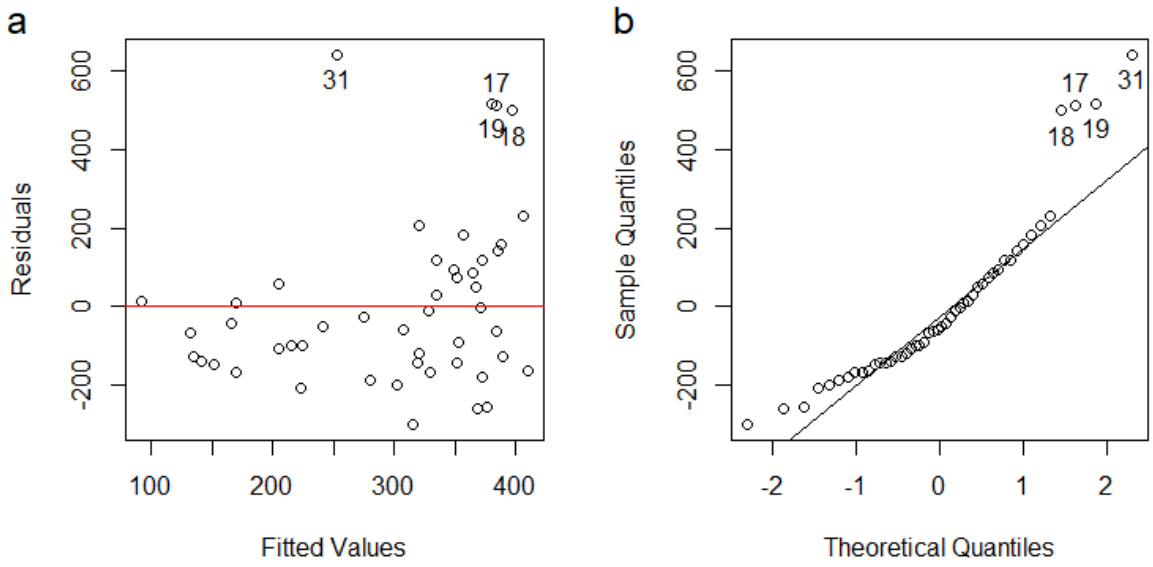

Figure 19: The residual plot and Q-Q-plot for the model. **(a)** The fitted values are plotted against the residuals to analyse the goodness of fit by noticing trends in the residual distribution. There are four residuals that stand out well from the other points. These residuals are labelled with their sample ID:s. The y-axis is the residuals, and the x-axis is the fitted value. The red line represents where the residual is zero and a well-fitted model would ideally have residuals evenly distributed around this line. **(b)** The Q-Q quantiles of the dataset are plotted against the expected quantiles of a theoretical distribution. The y-axis is the sample quantiles, and the x-axis the theoretical quantiles. The diagonal line in a Q-Q plot represents an ideal match between observed and expected quantiles. Four values stand out from the rest of the points and are labelled with their sample ID.

Table 6: The values of the samples that were identified as outliers. The table shows each sample's ID, date, location, average *E. coli* concentration, water temperature and rainfall.

Sample ID	Sampling Date	Location	Average *E. coli* concentration (cfu/100 mL)	Water temperature (°C)	Rainfall (mm)
17	2023-07-17	6	895.9	18.7	40.0
18	2023-07-17	5	895.9	19.1	40.0
19	2023-07-17	5.5	895.9	19.3	40.0
31	2023-07-24	8	895.9	18.8	47.4

Cook's distances for all observations for the second data in displayed in *Figure* 20. The threshold value was calculated to 0.0833. The two sample ID:s of the observations identified as overly influential were 15 and 19. Notice that 19 also is an outlier (*Figure* 19). The values for the identified samples can be seen in *Table* 7.

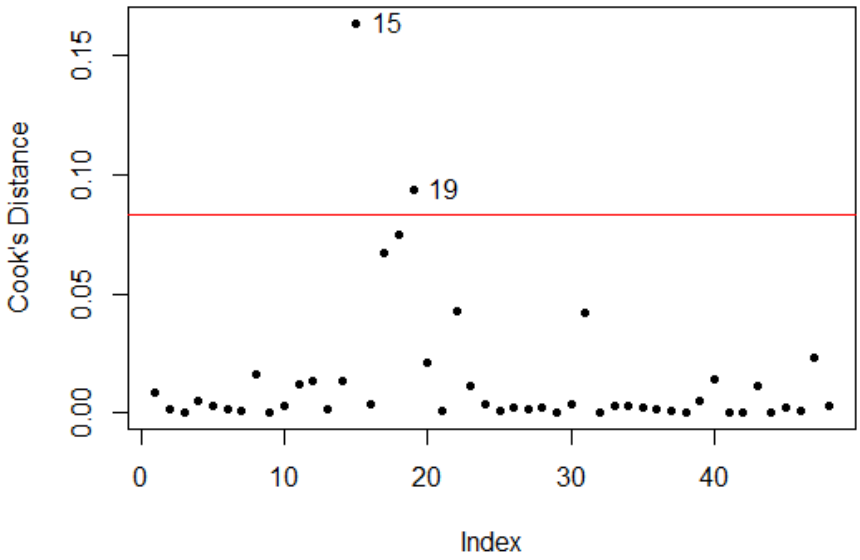

Figure 20: Cook's distance. The y-axis is the Cook's distance and on the x-axis are the sample ID:s. The black dots are the Cook's distance for respective sample. The red line is the threshold value, where the values above are considered overly influential. All samples that have a value higher than the threshold value are labelled with their respective sample ID.

Table 7: The values of the samples that were found to be overly influential. The table shows each sample's ID, date, location, average *E. coli* concentration, water temperature and rainfall.

Sample ID	Sampling Date	Location	Average *E. coli* concentration (cfu/100 mL)	Water temperature (°C)	Rainfall (mm)
15	2023-07-10	8	15.7	22.1	9.5
19	2023-07-17	5.5	895.9	19.3	40.0

6. Discussion

6.1 Result Summary

The spring model had estimated water temperature and rainfall to 12.62 respective 6.65. Both fixed effects were found to be statistically significant. The summer model estimated water temperature to 18.86 but it was not statistically significant.

6.2 Model Assumptions

6.2.1 Spring Model

The residuals were determined to be heteroskedastic (see *Figure* 16), meaning that the variance of the residuals was higher for larger fitted values. The result of this could be that the model is less accurate at making predictions at higher concentrations. Although heteroskedasticity was detected, no single observation could be assessed as an outlier. However, when studying influential values by calculating Cook's distance, four observations were identified as being overly influential. The observations had

sample ID:s of 5, 10, 26 and 52. When looking further into the influence of the observations it was found that the samples 5, 10 has a negative influence on the water temperature parameter while sample 52 had a positive influence. The parameter for rainfall was negatively influenced by sample 26.

Sample 5 was one of the observations that lacked a measurement for water temperature. The values for the water temperature on the 2023-04-24 were imputed as explained in section "Handling missing data". Considering that the value for water temperature was imputed and that the observation is overly influential suggest that the imputed value may not be reliable. The other samples at the same location had a higher concentration at lower temperatures, suggesting that the imputed temperature was probably too high. Looking at sample 10, it had a higher concentration than other samples at the same location but with a lower temperature. Sample 10 had an original concentration above 2500 *E. coli*/100 ml but was decreased by Winzoring to better fit the model. The high concentration for sample 10 is most likely not a cause of weather effects and therefore it might not be an observation that is worth having when fitting the model. Sample 26 and 52 were higher than the other samples taken in the same locations, contributing to the influence of the parameters but nothing suggests that these measurements were abnormal like sample 10.

The model was reconstructed by removing sample 5 and 10 and the new model results can be seen in *Table* 8. The model without both observation 5 and 10 gave estimates of the parameters that are similar to the original model, where the parameters were 12.34 and 6.25, but with a higher significance. This suggests that the model is a better fit when the influential values were removed.

Table 8: Estimates of the fixed effects in the spring model overly influential values had been evaluated and the model after reestimation. The table shows the estimation of each fixed effect in the model. The t-value and p-value come from performing a t-test on the fixed effects significance. Both variables were found to be significant.

Parameter	Estimate	t-value	p-value
Water temperature	13.07	2.705	0.0096
Rainfall	7.16	2.801	0.0078

6.2.2 Summer Model

In contrary to the spring model, when looking at the residual plot in *Figure* 19 four samples could be indicated as outliers. These samples had ID:s 17, 18, 19 and 31 and the values of these samples can be seen in *Table* 6. All these samples have the same measured *E. coli* concentration at 895.9 cfu/100 mL. The reason for this was that samples 18, 19 and 31 had concentrations outside of the IQR and was changed by Winsorization to the highest concentration inside of the IQR which was the concentration of sample 17. Looking at the sampling dates, samples 17, 18 and 19 were all taken during the same day and sample 31 was taken the next week. Another interesting point is that the samples taken on the 2023-07-17 are located next to each other. Sample 31 was located at the mouth of the stream, and the high concentration, one week after the concentration was unusually high in the stream, could be a result of the stream flushing out *E. coli* into Barnviken which was theorized by Dwite (2023).

Looking at sample 15 which was indicated as being overly influential by Cook's distance, its values differed from the other samples taken during the same date. Sample 15 had a much higher water temperature than the other samples taken during the same day. Sample 15 had a temperature of 22.1 °C and the other samples taken during the same day had a temperature between 18.3-18.7 °C, except the

sample in location 7 which had a temperature of 19.5 °C. It's hard to say whether this measurement is correct or not. Location 8, at which sample 15 was taken is located in Barnviken and not actually in the stream which could be reason for it having a higher temperature. However, the sample hade an exceptionally low *E. coli* concentration of just 15.7 cfu/100 mL, which is one of the lowest of all samples. The high temperature and low concentration is what has caused the sample to be overly influential.

The summer model was reestimated after removing samples 15, 17, 18, 19 and 31. The new model estimated the water temperature parameter at 5.38. The t-value was 0.412 and the p-value 0.683. The new model is less significant than the original model where the p-value for the water temperature parameter was 0.369. Compare this to the spring model, whereby removing the outliers resulted in more significant estimates.

6.3 Interpreting the Results

The results provided two models that have some differences from each other. The prediction model for the spring model gave increments in *E. coli* concentration for increments of both water temperature and rainfall and both parameters were statistically significant (see *Table* 4). The assumption that the random effects follow a normal distribution was fulfilled. The residuals, however, were indicated as heteroscedastic.

The prediction model for the summer model gave a similar yet different prediction. First of, the model had a better fit without including rainfall as a fixed effect, comparing that to the first model where rainfall was both included and found to be significant. The estimate for the water temperature gave a similar value to that of the first model, 18.86 compared to 12.34. However, the parameter was not found to be statistically significant. The residuals had a more even distribution than the spring model, but there were four samples that were indicated as outliers.

In summary, the spring model showed a positive trend with the explanatory variables and by hypothesis tests of the parameters and studying model assumptions, it is reasonable to assume that the model could be able to make predictions in the future. The summer model also showed a positive trend but, but the parameter is not statistically significant. The model appeared inadequate for making additional predictions. So why does it appear that water temperature and rainfall can predict *E. coli* concentration during the spring but not in the summer? The reasons for the differences in the models will be discussed in the next sections. First, environmental causes will be discussed, meaning all reasons capable of explaining the differences in the models that are not related to statistics. In the section after that we will discuss limitations to the samples and the linear mixed model that could have affected the result.

6.4 Rainfall as a Predictor

It is unknown why rainfall was able to b a significant parameter for the spring model but was not even included in the summer model. However, some theories can be discussed. One reason could be that it simply rained too much. To assume that the *E. coli* concentration would continue to grow linearly with huge amounts of rainfall would be unreasonable. Looking at the models, rainfall seem to have an impact on the concentration for lower concentrations, but only up to a certain point. After that the linear effect of rainfall is no longer evident. But that is just an assumption. It could also be that the linear effect of rainfall continuous even for higher values but that it rained so much that concentration became diluted, increasing the amount of *E. coli* but not the overall concentration. This could be an explanation for why the rainfall does not seem to have any effect on the concentration at all in *Figure* 13. It could also be that the large amount of rainfall caused the creek to flush out into the sea, taking the *E. coli* with it.

Another possible explanation is that the quantity of rain may not be significant; rather, the crucial factor might be whether it rains at all. The first sampling period had two weeks where it did not rain, but the

second sampling period did not any dry periods, and rainfall was only significant in the first model. Rain can increase the concentration through sedimentation, bringing nutrients to the *E. coli*. However, it might be that the amount of rain does not result in a significant increase in sedimentation, leading to the parameter only showing significance when there are samples from both rainy and non-rainy periods.

6.5 Environmental Reasons

Referring to the article *Investigation of Fecal Contamination in the Hammar´s stream in Malmö, Barnviken* (2023), Dwita mentions that the high *E. coli* concentration might be a consequence of a leaking sewage pipe from a nearby toilet. This theory can also be backed up by the result from the models. Assuming that more people would use the toilet during the summer, which could be likely since it is located close to the beach, the leakage would have a bigger impact during the summer than the spring. Faecal contamination could be an explanation factor that would outweigh water temperature and rainfall, making the summer model futile. It is worth noting that, so far, no evidence has shown that there might be a leakage. It is simply something that could be further investigated.

One reason could be the nearby camping called Malmö Camping & Feriecenter. It can also be assumed that the camping site is more populated during the summer season than during the spring. If that is true than pollution in the form of waste could be a reason for higher concentrations during the summer. This could also be an explanation for the difference between the models since waste could outweigh water temperature and rainfall as parameters. One reason could also be that the number of seagulls in the area increases, as before, contaminating the water and outweighing the other factors.

6.6 Statistical Reasons

The difference between the models could also be a consequence of the type of model used and the sampled data. A potential explanation might be that there are fewer data points for some locations during the first sampling period compared to the second one. Locations 3.5 and 8 had three measurements, and location 5.5 had five, while all other locations had six measurements during both sampling periods. Having fewer data points makes it so that the model can more easily find a pattern that fits all the points. On the contrary, the model cannot find a pattern when the number of data points increases. The first model might be able to provide significant parameters because the regression can easily fit all points. This might lead to overfitting, meaning that the model gives accurate predictions for the existing data but not when trying to use it for predictions on new data.

7. References

7.1 Literature Sources

Altman, N. & Krzywinski, M. (2016) . Analyzing outliers: influential or nuisance?. *Nature Methods,* Volume 13, pp. 281-2. https://www.nature.com/articles/nmeth.3812

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* https://people.math.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf (Accessed 2023-11-11).

beaches.ie. (2023). *Frequently Asked Questions.* https://www.beaches.ie/frequently-asked-questions/ (Accessed 2023-11-11).

Busemeyer, J. R. & Diederich, A. (2014). Chapter 4 - Estimation and Testing of Computational Psychological Models. In: Glimcher P.W & Fehr E. (eds). *Neuroeconomics.* 2:nd ed. Academic Press, pp. 49-61. DOI: https://doi.org/10.1016/B978-0-12-416008-8.00004-8

Center for Disease Control and Prevention. (2014). *E.coli (Escherichia coli): Questions & Answers.* https://www.cdc.gov/ecoli/general/index.html (Accessed 2023-09-25).

Dwite, A. (2023). *Internship Project: Investigation of Fecal Contamination in the Hammar´s stream in Malmö, Barnviken.* Lund: Lund University.

Ford, C, University of Virginia Library. (2015). *Understanding QQ Plots.* https://library.virginia.edu/data/articles/understanding-q-q-plots#:~:text=A%20QQ%20plot%20is%20a,a%20line%20that's%20roughly%20straight (Accessed 2023-10-23).

Glen, S, Statistics How To. (2023a). *Residual Plot: Definition and Examples.* https://www.statisticshowto.com/residual-plot/ (Accessed 2023-10-23).

Glen, S, Statistics How To. (2023b). *Winsorize: Definition, Examples in Easy Steps*. https://www.statisticshowto.com/winsorize/ (Accessed 2023-10-23).

Jiang, J. & Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications.* 2nd ed. New York: Springer. https://link.springer.com/book/10.1007/978-1-0716-1282-8

Körner, S. & Wahlgren, L. (2018). *Statistisk dataanalys.* 5th ed. Lund: Studentlitteratur AB.

Moberg, L.J. (1985). Fluorogenic assay for rapid detection of Escherichia coli in food.. *Applied and Environmental Microbiology,* 50(6): pp. 1383-7. DOI: 10.1128/aem.50.6.1383-1387.1985

Nieuwenhuis, R., Grotenhuis, M. & Pelzer, B. (2012). influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *The R Journal,* 4(2), pp. 38-47. https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Nieuwenhuis~et~al.pdf (Accessed 2023-11-04).

Public health agency of sweden (2015). *Sjukdomsinformation om escherichia coli-infektioner i tarmen.* https://www.folkhalsomyndigheten.se/smittskydd-beredskap/smittsamma-sjukdomar/escherichia-coli-infektioner-i-tarmen/ (Accessed 2023-09-25).

Ramachandran, K. & Tsokos, C. (2021). Chapter 11 - Categorical data analysis and goodness-of-fit tests and applications. *Mathematical Statistics with Applications in R.* 3:rd ed. Amsterdam: Elsevier pp. 461-90. DOI: 10.1016/B978-0-12-817815-7.00011-7

RDocumentation. (2023). *mice: mice: Multivariate Imputation by Chained Equations*. https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice (Accessed 2023-12-12).

rdrr.io. (2023). *isSingular: Test Fitted Model for (Near) Singularity.* https://rdrr.io/cran/lme4/man/isSingular.html (Accessed 2023-12-17).

Swedish Agency for Marine and Water Management. (2013). *Vägledning för badvatten enligt direktiv 2006/7/EG (EU-badvatten).* Göteborg: Swedish Agency for Marine and Water Management. https://www.blaflagg.org/wp-content/uploads/2016/04/Badvattenvagledning_HaV_v9.pdf (Accessed 2023-12-12).

Swedish Agency for Marine and Water Management. (2022). *Vägledning och administration.* https://www.havochvatten.se/badplatser-och-badvatten/vagledning-och-administration.html (Accessed 2023-12-30).

Swedish Agency Marine and Water Management. (2018) *EU Bathing Sites.* https://www.havochvatten.se/en/facts-and-leisure/bathing-water-quality/eu-bathing-sites.html (Accessed 2023-11-27).

UCLA: Statistical Consulting Group. (2021). *INTRODUCTION TO LINEAR MIXED MODELS.* https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/ (Accessed 2023-10-23)

UCLA: Statistical Consulting Group. (2021). *HOW DO I PERFORM MULTIPLE IMPUTATION USING PREDICTIVE MEAN MATCHING IN R? | R FAQ.* https://stats.oarc.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/ (Accessed 2023-12-12).

West, B. T., Welch, K. B., Galecki, A. T. & Gillespie, B.W. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. New York: Chapman and Hall/CRC. https://lubcat.lub.lu.se/cgi-bin/koha/opac-detail.pl?biblionumber=1806618

Westerberg, O. (2018). Höga halter tarmbakterier i Barnviken. *Sydsvenskan*, 24 Jul. https://www.sydsvenskan.se/2018-07-24/hoga-halter-tarmbakterier-i-barnviken (Accessed 2023-12-12)

Yang, Z.R. & Yang, Z. (2014). 6.01 Artificial Neural Networks. *Comprehensive Biomedical Physics 6:* pp. 1-17. DOI: https://doi.org/10.1016/B978-0-444-53632-7.01101-1

7.2 Data Sources

Dwite, A. (2023). *Internship Project: Investigation of Fecal Contamination in the Hammar´s stream in Malmö, Barnviken* [Dataset]. Lund University.

Swedish Meteorological and Hydrological Institute (SMHI), (2023). *Nederbördsmängd (dygn)* [Dataset]. https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer/#param=precipitation24HourSum,stations=core,stationid=52350 (Accessed 2023-09-22)