Smart Water Monitoring for better water quality

An affordable and sustainable way to predict unexpected environmental changes of water bodies.

Chia-Wen Yang

Master Thesis Series in Environmental Studies and Sustainability Science,
No 2023:016

A thesis submitted in partial fulfillment of the requirements of Lund University International Master's Programme in Environmental Studies and Sustainability Science (30hp/credits)

LUCSUS
Lund University Centre for
Sustainability Studies

Smart Water Monitoring for better water quality

An affordable and sustainable way to predict unexpected environmental changes of water bodies.

Chia-Wen Yang

A thesis submitted in partial fulfillment of the requirements of Lund University International Master's Programme in Environmental Studies and Sustainability Science

Submitted May 9th

Supervisor: Murray Scown, LUCSUS, Lund University

Empty page

Abstract

This article discusses the challenges associated with maintaining water quality, particularly the issue of eutrophication, and the importance of technology advancements for monitoring and managing water quality. The potential of machine learning is highlighted, along with the importance of affordable and effective water quality monitoring techniques. The study aims to identify factors contributing to eutrophication in Sweden, using various regression models, including Random Forest and XGBOOST. The exploratory data analysis showed that environmental parameters may not have a strong linear relationship with chlorophyll concentration, but other variables such as nutrient availability and light may play a more important role. The random forest model produced the most accurate predictions. The study also discusses the importance of technology diffusion in promoting sustainable water management practices in the Global South and emphasizes the need for collaboration between developed and developing countries.

Keywords:

Machine learning, technology diffusion, Water quality management, Sustainable development, Global South, Affordable solutions

Word count: 9039

Acknowledgements

I would like to take this opportunity to express my heartfelt gratitude to all those who supported and guided me throughout this program. Coming from a different discipline, adapting to the new challenges was not easy, but with your invaluable assistance, I was able to overcome them.

To my dear mental support group "Girl Skwad" - Alena, Eda, Mahima, and Sahana - thank you for being there for me during tough times. Your motivation and encouragement helped me stay on right track. The best part of this program was meeting you all and studying together until the very last minute. Elisa, I am grateful for your constant academic support and inspiration. Nkechi, your incredible hard work has truly inspiring, and I appreciate the time we spent practicing Swedish and sharing the difficulties in life. Thank you, Moonhee and Maja, my Korean mates, for being my language exchange partners and sharing the same hobbies as me. Niki, meeting someone from Taiwan who shares my passion for sustainability and studying in Sweden was truly wonderful.

A special thanks to my supervisor, Murray, for being patient and well-organized throughout the program. I would like to express my gratitude to Nour and Vijaya from Vaquita for providing me with the opportunity to work with water and enhancing my Python skills.

Last but not least, I want to thank my family. Lidia, thank you for being my constant support system. Mom and Dad, thank you for your unwavering encouragement means the world to me. And to myself, I am incredibly proud of achieving my dream of studying abroad and completing this thesis.

Table of Contents

Table of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
CA	Cluster Analysis
CHL	Concentration of Chlorophyll
DO	Dissolved Oxygen
EC	Electrical Conductivity
EDA	Exploratory Data Analysis
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
N	Total concentration of total nitrogen
P	Total concentration of total phosphorus
PCA	Principal Component Analysis
PR	Polynomial Regression
R^2	Coefficient of determination
RF	Random Forest
r_s	Spearman rank correlation coefficient
SDGs	Sustainable Development Goals
SVM	Support vector machines
SZE	Sweden Zero Eutrophication
T	Water Temperature
XGBOOST	Extreme Gradient Boosting

1. Introduction

1.1. Water quality and Sustainable Development Goals

Water is a vital resource to the Earth, supporting not only aquatic life but the entire ecosystem. The quality and quantity of water are critical to the health and wellbeing of human and other creatures. However, the rapid pace of industrial development and population growth has put significant pressure on water resources (Jamil & Shehab, 2021; Simeonov et al., 2003). As a result, water resources become vulnerable to pollution, depletion, and degradation. This has led to a lack of access to clean water for people, especially to those who live in developing countries (Rozenberg & Fay, 2019; Thacker et al., 2019). It is critical to take action to protect and preserve water resources to ensure that they remain accessible, clean, and sustainable for future generations.

The importance of water resources is reflected in the fact that two of the Sustainable Development Goals (SDGs), namely SDG 6 and SDG 14, are dedicated to addressing water-related issues. SDG 6 focuses on universal access to clean water and sanitation, while SDG 14 emphasizes the conservation and sustainable use of marine and freshwater resources. However, achieving these goals is becoming increasingly challenging due to various factors, including unexpected environmental changes in water bodies. Such changes, which can include alterations in water temperature, pH, nutrient levels, and dissolved oxygen concentrations, etc, can have significant impacts on aquatic ecosystems and the human communities that depend on them (Carmichael & Boyer, 2016; Chorus & Welker, 2021; Sin & Lee, 2020; Smith, 2003). Therefore, regular water quality monitoring can help the scientists and governments to understand the situation and deal with the ill effects as soon as possible (Mavukkandy et al., 2014; Noori et al., 2010).

Eutrophication is one of the most notorious water problems in the world, characterized by an excessive buildup of nutrients, such as nitrogen and phosphorus, in water bodies. The process of eutrophication is often caused by the discharge of untreated wastewater and agricultural runoff into waterways (Conley et al., 2009; Newcomer Johnson et al., 2016). Eutrophication can lead to harmful algal blooms (HABs), decreased water quality, and the depletion of oxygen levels, which can be detrimental to aquatic life and human health (Andersen et al., 2017; Carmichael & Boyer, 2016; Dimitra Kitsiou & Michael Karydis, 2019; Gregersen et al., 2023; Pawlak et al., 2009; *Plateau Lake Water Quality and Eutrophication*, 2023; Schindler, 1977; Scholten et al., 2005; Zhang et al., 2020). According to the study in 2018, over 60 % of total water bodies were eutrophic (Wang et al., 2018).

Furthermore, eutrophication can have far-reaching economic consequences, affecting industries such as fishing, tourism, and water treatment (Boesch et al., 2001; Dodds et al., 2009; Garcia-Hernandez et al., 2022; Jimeno-Sáez et al., 2020; Smith, 2003). With climate change exacerbating the problem (Howarth et al., 2006), developing solutions to monitor and manage water quality and HABs has gained considerable global attention (Yajima & Derot, 2017), and advanced techniques to forecast can be instrumental in this effort.

1.2. Monitoring water quality for sustainable development

Although there are various operational methods available for inspection, monitoring, and data collection, many of them can be time-consuming and expensive in practice. The current monitoring methods used by authorities typically involve taking water samples directly from the site of interest and transporting them to laboratories where they are analyzed using spectrophotometry techniques. However, this process can be time-consuming and may take from one day to several weeks depending on the schedule and delivery times (Palmer et al., 2015). This delay in obtaining the results can hinder timely actions to be taken to prevent further degradation of water quality. Moreover, this method is often labor-intensive and can be costly (J. Chen et al., 2013; Papenfus et al., 2020). Meanwhile, long-term monitoring projects may experience data gaps due to budget constraints or a lack of funding (Pinto et al., 2013). To overcome these challenges, it is crucial to investigate and implement more efficient and cost-effective techniques for monitoring and data collection. Additionally, it is essential to ensure sustained data collection over the long term, which should be affordable and accessible to developing countries.

In addition to monitoring, forecasting eutrophication can be an efficient and valuable tool in improving efforts to address water-related problems. By analyzing historical data and current water quality information, a forecast model can forecast future changes in water quality, such as the potential for harmful algal blooms or changes in nutrient levels. This can enable water managers and policymakers to take proactive measures to prevent or mitigate potential negative impacts on aquatic ecosystems and human communities. For instance, a forecast model can provide early warnings of potential harmful algal blooms, allowing water treatment plants to adjust their treatment processes accordingly, or advising recreational users to avoid swimming in affected areas. In addition, by identifying areas at risk of eutrophication, forecast models can help guide land-use planning and decision-making to reduce nutrient inputs into water bodies, such as implementing best management practices in agriculture or wastewater treatment.

Monitoring programs often collect a vast amount of data that can be challenging to analyze and interpret due to its complexity and size (Gurjar & Tare, 2019). The cost of conducting these monitoring projects, and building forecast models, can be a significant barrier for vulnerable countries, making it difficult to maintain them for an extended period. While advanced technologies offer promising solutions for monitoring eutrophication, they can be expensive and limited in their applicability. As a result, there is a need for more efficient and cost-effective methods that can provide real-time or forecast models on water quality. Meanwhile, technological solutions can play a crucial role in managing water resources and preventing water pollution if they are developed and disseminated effectively.

1.3. Research Aim and objectives

This thesis aims to tackle the critical issue of forecasting and preventing unexpected environmental changes in water bodies, with a specific focus on forecasting algae blooms. The primary objective is to identify effective and affordable parameters for forecasting algae concentration and discuss how technological solutions for monitoring and forecast modeling can be applied to sustainability challenges related to water.

In order to achieve the objectives, this thesis research asks one overarching question and four sub questions:

What is the extent of predictive power that basic environmental factors have on chlorophyll concentrations (CHL) based on historical data and how complicated do the predictive models need to be?

1. Can univariate statistical correlations forecast chlorophyll concentrations?
2. Can univariate polynomial linear regression forecast chlorophyll concentrations?
3. Can multiple linear regression forecast chlorophyll concentrations?
4. Can machine learning models forecast chlorophyll concentrations?

Identifying the most influential factor in forecasting CHL in water bodies is crucial, as it can aid in pinpointing the critical factors that contribute to algae growth and developing more accurate predictive models. After presenting the findings of the study, I will delve into the implications of these results for the diffusion and implementation of technological solutions aimed at addressing water quality challenges around the world. This will involve exploring the potential transferability of the methods and techniques used in this study to other contexts, as well as examining the broader socio-economic, political, and environmental factors that may affect the adoption and impact of such solutions. By doing so, this study can contribute not only to the understanding of eutrophication in Skåne län but also to the broader global efforts to promote sustainable water management practices.

2. Background and approach

2.1. Technological solutions and diffusion

In response to the challenges of climate change, technologies have advanced in a more efficient, cost-effective, and sustainable way to achieve the SDGs. From renewable energy to food production, technology solutions have the potential to transform how we approach sustainable development. However, the diffusion of these technologies can be a challenge, particularly in developing countries, where technical and financial limitations may hinder their widespread adoption.

The theory of diffusion of innovations posits that a novel idea or product will gradually propagate throughout a particular group or community as time passes (Rogers, 2003). Consequently, individuals within that society will modify their behavior in order to conform to these new developments. The theory suggests that the rate and pattern of adoption of a new technology is influenced by various factors. There are five stages of the adoption process to new technology: knowledge/awareness, persuasion, decision, implementation, and confirmation/continuation. Additionally, there are several key elements that affect the adoption process, including innovation, adopters, communication channels, time, and the social system (Rogers, 2003).

Take Carbon Capture and Storage as a notable example of technology solutions to climate change. This approach can potentially address the issue of greenhouse gas emissions while also producing a useful resource. The development of such technologies is crucial for the mitigation of climate change and the creation of a more sustainable future. Initially, in the knowledge/awareness stage, scientists and stakeholders disseminated information to catch the attention of the public. People recognized the severity of climate change and sought to develop technologies to capture carbon from the air. In the persuasion stage, the benefits of adopting new techniques, such as effectively capturing carbon emissions and slowing down the rate of warming, were presented to other stakeholders. The decision stage involved policymakers deciding whether to invest in or implement the new technology in a specific area. In the implementation stage, the innovation was adapted to the study area, policies were developed to support the technology, and feedback was used to improve it. Finally, in the confirmation stage, stakeholders evaluated the effectiveness of the technology in reducing greenhouse gas emissions and mitigating climate change, as well as the costs and benefits of using the technology.

In the case of carbon capture and storage, the successful diffusion of this innovation is influenced by various factors. The key elements include the innovation itself, which refers to the techniques of carbon capture and storage, and the potential adopters, who could be various stakeholders such as researchers, government officials, and energy companies. Effective communication channels, such as research journals, academic conferences, and other media platforms, are also essential for disseminating information and promoting adoption. The adoption process for this technology is likely to be a long and complex one, requiring significant investment in time, resources, and infrastructure. Finally, the broader social, economic, and political systems in which innovation operates, collectively referred to as the social system, will also play a critical role in determining the success of carbon capture and storage as a means of mitigating greenhouse gas emissions.

Despite the importance of technological solutions in achieving sustainable development goals, the access to the latest technology is not always universal, and vulnerable countries often face significant delays in implementing new technology. Developing countries, particularly those in the Global South, face a range of socio-economic challenges, including limited resources and infrastructure, a shortage of technical expertise, and political and regulatory barriers, hindering their progress (Hanson et al., 2017; Sano et al., 2013).

Only when technological solutions have become increasingly accessible and user-friendly, resulting in easier adoption and diffusion. For example, the use of emerging technologies such as blockchain, artificial intelligence, and the Internet of Things (IoT) can further accelerate the diffusion of sustainable technology solutions in the developing countries. These technologies have been massively applied to various uses in daily life from fully automated factories and smart home systems (Kravchenko et al., 2017).

2.2. Effective and Affordable Water Quality Monitoring Techniques

In recent years, with advancements in technology and research, there has been a continuous improvement in the methods used for monitoring and managing eutrophication. These improvements have included the development of new sensors and devices for monitoring soil composition, nutrient levels, and other important environmental factors, as well as the use of machine learning algorithms and other advanced analytical tools to analyze and interpret the data collected.

To effectively and affordably monitor water quality, a thorough understanding of the environmental factors that promote algal growth is necessary. Algal growth in bodies of water can be prevented by monitoring the biomass of algae and related factors such as chlorophyll (CHL), phosphorus, and nitrogen concentrations (Chapman et al., 1996; Malek et al., 2011). Traditionally, optical sensor-based techniques have been used to measure these parameters, with CHL serving as the primary photosynthetic pigment and a reliable indicator of total algal biomass in surface water (Boyer et al., 2009; Carneiro et al., 2014; Li et al., 2018; O'Sullivan & Reynolds, 2004). However, there are some limitations to the use of traditional methods. For example, the experimental optical sensors for chlorophyll are often unportable or expensive, and there is a risk of sample pollution during delivery. It is important to be aware of these limitations to ensure accurate and reliable data collection.

One such technology is quantitative ocean color remote sensing, which is used to sense the concentration of chlorophyll (CHL) in water bodies using shortwave infrared bands. However, this method is limited in its applicability to coastal and inland water bodies due to atmospheric interference (J. Chen et al., 2013). Other advanced technologies, such as fluorescence spectroscopy and hyperspectral imaging, have been explored for their potential to monitor eutrophication, but their efficacy in real-world settings remains to be fully tested and evaluated (Alminagorta et al., 2021).

Portable real-time CHL sensors offer the advantage of providing more immediate and on-site measurements, but they do come with certain limitations. From the previous studies, environmental factors such as temperature, pH, dissolved oxygen (DO), weather, nutrient levels, and electrical conductivity (EC) have been shown to be strongly related to or promote algal growth (Beretta-Blanco & Carrasco-Letelier, 2021; Gardner-Dale et al., 2017; J. Kim et al., 2022; Liu et al., 2010; Ras et al., 2013; Shoener et al., 2019). Therefore, it is necessary to monitor these factors in order to assess and prevent eutrophication. Recent advancements in sensor technology have resulted in more affordable and portable sensors, which have made water quality monitoring more accessible and convenient for researchers. For instance, it is now possible to use simpler sensors to detect basic environmental parameters and input the data into a well-developed model to obtain the other more complicated factors. By adapting forecasting models to these new sensors, researchers can optimize water quality monitoring techniques, resulting in more accurate and cost-effective water quality monitoring. The development of such methods can significantly aid in the timely management of water resources and the prevention of water pollution.

To improve the efficiency of monitoring, statistical methods have been employed in combination with sampling. Statistical methods such as cluster analysis (CA) and principal component analysis (PCA) have also been employed with sampling to improve the efficiency of monitoring (Simeonov et al., 2003). Additionally, linear regression is a widely used statistical model that is used to describe the relationship between variables, assuming that there is a linear correlation between them (Mendenhall

& Sincich, 2012; Sen & Srivastava, 1997). Multiple Linear Regression (MLR) is a simple yet effective linear regression model used for predicting outcomes based on linear relationships between variables.These methods reduce the burden on local authorities by providing a more streamlined and cost-effective approach to collecting and analyzing data. By using historical data and statistical methods to identify trends and point sources, Pinto et al. (2013) was able to optimize the design of future monitoring programs and focus on the areas that are most critical for maintaining water quality. This approach can also help to prioritize resources and funding for monitoring programs and make them more cost-effective.

2.3. Machine Learning Methods

In recent years, the development of open-source machine learning software and cloud-based platforms has democratized access to these tools. This has enabled businesses, individuals, and organizations to leverage the power of machine learning without significant financial investments. Therefore, the widespread availability of machine learning technology has facilitated its diffusion, which is essential for achieving sustainable development goals.

The emergence of machine learning (ML) has revolutionized various fields since the mid-20th century with the early developments in artificial intelligence (AI) (Michalski & Anderson, 1982). However, it was only after the rise of big data analysis that the power of ML was taken seriously. During the 1980s and 1990s, researchers developed a range of ML algorithms, including decision trees, support vector machines, and Bayesian networks, among others. These algorithms were used in a variety of applications, such as speech recognition, computer vision, and natural language processing. With the advent of big data in the 21st century, machine learning has become even more crucial in several fields such as finance, healthcare, and marketing. Today, ML algorithms are used in a wide range of applications, from image recognition to fraud detection to personalized recommendation systems. The usefulness of machine learning is also reflected in sustainability, where it can help address complex challenges. For instance, ML algorithms can optimize crop yields, reduce waste, and improve soil health. By analyzing data on soil quality, weather patterns, and crop performance, machine learning can assist farmers in making more informed decisions about planting, fertilization, and irrigation (K et al., 2023). Overall, machine learning provides a promising tool to better understand sustainability challenges and develop effective solutions.

The ability of Machine Learning (ML) to analyze complex datasets and generate accurate predictions has made it an integral part of various fields, especially in the realm of water quality forecast (Nearing et al., 2021). There are numerous ML methods available, each with its own strengths and weaknesses (de Vita et al., 2022; Khullar & Singh, 2020).

Artificial Neural Networks (ANN) have proven to be a useful ML method in various fields due to their ability to mimic the functioning of the human brain (Cho et al., 2011; Fausett, 1994; Haykin, 1994). ANN consists of several layers and has been widely applied to address various water quality issues on a large scale (Campolo et al., 1999; E. Kim et al., 2019). However, the number of hidden layer neurons is crucial to ensure the prediction accuracy of the ANN. If the number of hidden layer neurons is too low, it will compromise the prediction accuracy, and if it is too high, the modeling process would become computationally expensive and time-consuming (Charulatha et al., 2017).

Support vector machines (SVM) is another well-established ML method that is effective in solving both classification and regression problems. This method can handle nonlinear data and achieve high accuracy in predictions (Yusri et al., 2022).

Random Forest (RF) is an ensemble learning method that utilizes a collection of decision trees to enhance predictive accuracy. RF has the advantage of handling incomplete data and still achieving satisfactory results (Breiman, 2001; Díaz-Uriarte & Alvarez de Andrés, 2006; Fang et al., 2021). The RF model, like many other models, is also prone to overfitting (Breiman, 2001). A study for water quality conducted in 2021 reported that the achieved accuracy of RF was 92.94% (Xu et al., 2021).

Extreme Gradient Boosting (XGBOOST) is an ensemble learning method that combines the strengths of multiple decision trees and gradient boosting. It can solve various problems related to regression, classification, ranking, and user-defined prediction tasks with high accuracy and efficiency (T. Chen & Guestrin, 2016; Yusri et al., 2022). A study conducted in 2022 revealed that XGBOOST achieved high levels of accuracy, precision, and recall, specifically 95%, 96%, and 96%, respectively (Garabaghi et al., 2022).

Machine learning (ML) methods have revolutionized various fields, including water quality forecasting, due to their ability to analyze complex datasets and generate accurate predictions. They offer numerous benefits over traditional methods and have proven to be effective in solving different types of problems. However, instead of relying on expensive and complex models for forecasting the concentration of chlorophyll (CHL), this research would utilize two efficient and readily available ML methods that are well-suited to the dataset obtained.

3. Methodology

3.1. Thesis Roadmap

The research will commence by investigating the data features and testing the simplest statistical model (i.e., correlation) to establish the presence of any significant univariable factor that strongly influences the growth of algae. If no such factor is identified, the univariable polynomial regression and multiple regression model will be explored to determine if either of these models is a better fit. Ultimately, the Random Forest and XGBOOST machine learning models will be introduced to the dataset to determine the best fit for the data.

3.2. Study area and data collection

Sweden is a country that has set up goals to make significant progress in the reduction of eutrophication in its water bodies. To achieve this, the Swedish government launched the Sweden Zero Eutrophication (SZE) project, which aimed to reduce nutrient inputs to the Baltic Sea and achieve good environmental status by 2021. The SZE project used a combination of monitoring methods to assess the effectiveness of measures taken to reduce nutrient inputs. Additionally, the project involved collaboration between government agencies, NGOs, and stakeholders to develop and implement measures to reduce nutrient inputs (Naturvardsverket, n.d.). The SZE project serves as an example of a comprehensive approach to addressing eutrophication through monitoring and collaborative efforts. Unfortunately, many countries around the world lack the financial resources necessary to address environmental threats through targeted projects or programs.

Specifically, this study focuses on Skåne län, the most southern province in Sweden, also known as Scania County. Skåne län is an agricultural region with a high population density, comprising around 13 percent of Sweden's total population and covering around 3 percent of the country's land area. It is home to around 500 lakes larger than 0.01 square kilometers, many of which have a rich plant and animal life, giving them significant natural value (Henestål et al., 2021). Most of the counties in Skåne län have a warm humid continental climate, while a small part of the counties located near the coast belong to the Oceanic climate zone.

Agricultural dominance in Skåne län has led to significant issues with eutrophication (IVL, 2021). Region Skåne reports that the region's land serves 30 percent of the entire food chain in Sweden, with food products being the largest export goods. Therefore, the wastewater from the agricultural land became a critical issue. It’s important to understand the factors contributing to eutrophication in this region and develop accurate forecasting models to mitigate its effects. The average yearly temperature in this region of interest is 9.65ºC, with a precipitation level of approximately 90.02 millimeters per year (Weather and Climate, n.d.).

The environmental data used in this study was obtained from Miljödata, a database managed by Sveriges lantbruksuniversitet (SLU). This database has been collecting land, water, and environmental data since 1943, with a particular focus on water bodies and agricultural lands in Sweden. The dataset used contained crucial environmental variables such as chlorophyll concentration (CHL, µg/L), pH, electrical conductivity (with the correction value at 25 degree, EC, mS/m), turbidity, dissolved oxygen (DO, mg/L), water temperature (T, °C), turbidity (FNU), concentration of total phosphorus (P, µg/L) and total nitrogen (N, µg/L), sampling date, and location (municipality), among others. The data covered a vast range of catchments and surface water in Skåne län, spanning almost 50 years, from June 3rd, 1973, to October 25th, 2022. The source of the collected data can be found in Figure 1, which shows that most of the data was collected from the middle part of Skåne län's lakes, Västra and Östra Ringsjön. However, due to limitations in the Python database, it is difficult to accurately depict the appearance of Skåne län. Therefore, the entire country of Sweden was used instead.

Weather data was collected using the OpenWeather One Call API 3.0, which provided information on ambient temperature, atmospheric pressure, latitude, longitude, cloud coverage, visibility, wind speed, wind direction, rain volume for the last hour, and snow volume for the last hour. The reason why I collected the weather data was because the CHL can also be affected by factors such as sunlight, temperature, precipitation, and wind speed (M. Chen et al., 2011; Wu et al., 2014). This information was used to complement the environmental variables in the analysis and provide a more comprehensive understanding of the factors affecting CHL in the water bodies. The database limitations include the lack of exact time of sample collection and the potential bias in using a fixed time of 12 pm for weather data collection, which may not accurately represent the weather conditions throughout the day in Skåne län.

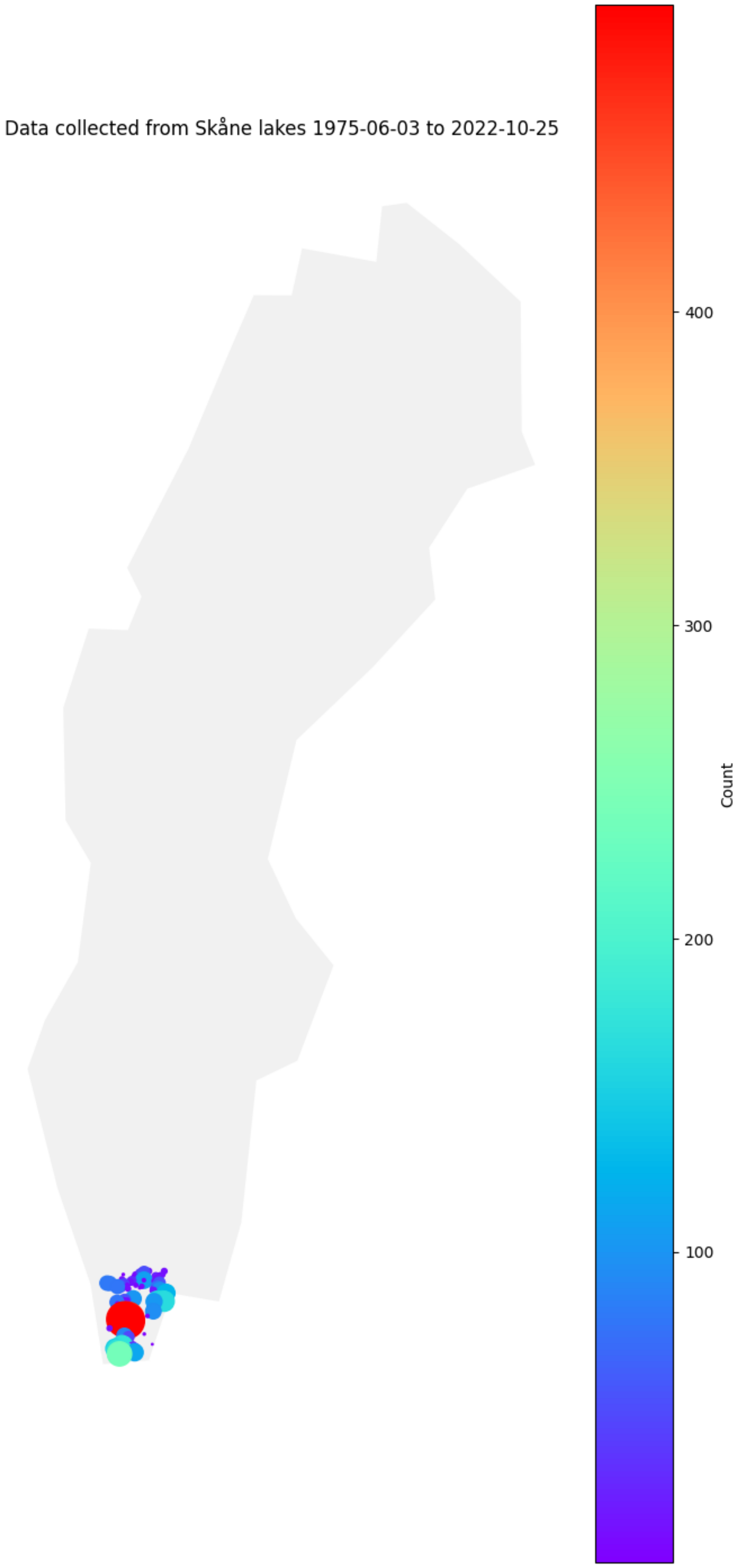

Figure 1. Map of Sweden and the sampling lakes from 1953 to 2022 generated from Python.

3.3. Data analysis

The data was imported from Microsoft Excel into Python and analyzed using Pandas. Exploratory data analysis (EDA) is an essential preliminary step that helps us understand the data, identify patterns and relationships, detect outliers, and generate hypotheses about the underlying process generating the data. It also involves visualizing and summarizing the data using various statistical and techniques, which informs the machine learning process in several ways. First, EDA can help to identify the most relevant variables for the machine learning task, which can reduce the dimensionality of the data and improve the performance of the model. Second, EDA can help to identify the relationships between the variables, which can inform the choice of the modeling approach and the selection of the appropriate algorithms. Third, EDA can help to identify the distributional properties of the data, which can inform the selection of the appropriate data preprocessing and transformation techniques. In this study, EDA was employed to analyze and investigate the data sets collected from Miljödata, allowing us to identify critical indicators and summarize the findings.

The first stage of the study models involved using univariate linear regression to investigate the relationship between two variables, where one variable is considered the dependent variable and the other variable is considered the independent variable. I utilized univariate linear regression as the basic stage of prediction to determine the relationship between CHL, which served as the dependent variable, and the various environmental parameters, including pH, T, EC, DO, P, and N, which were considered the independent variables.

To ensure comparability across the remaining stages, I applied data scaling to the dataset. By scaling the dataset, I standardized the range of values for each variable to fall within a similar numerical range. This process enabled us to reduce the potential impact of variables with larger values on the prediction models, which could have led to biased results. Then, the extreme values were removed from the scaled training dataset. To apply with other models, the dataset was split into a training set and a test set using the 80/20 rule to prevent overfitting of the models, with 80% of the data used for training and 20% for testing.

For the second stage, I use multiple linear regression and Polynomial regression (PR). Multiple linear regression (MLR) is a statistical method that can be used to examine the association between a dependent variable and several independent variables simultaneously. The method has been applied to classify and model environmental data to prevent misinterpretation of environmental monitoring data (Reisenhofer et al., 1996). In this study, I performed multiple linear regression to forecast the CHL based on several environmental parameters. The multiple linear regression model allows us to assess the contribution of each independent variable to the variation in CHL, while controlling for the effects of the other variables. Like MLR, PR is a technique used in machine learning to model the relationship between a dependent variable and one or more independent variables as well. However, while MLR models a linear relationship between the variables, polynomial regression allows for modeling of nonlinear relationships (Ahmed et al., 2019). It is important to note that increasing the degree of the polynomial may also lead to overfitting, where the model becomes too closely tailored to the training data and does not generalize well to new data. It has been utilized for monitoring water quality due to its ability to capture data influenced by both natural and artificial factors that are often distributed non-linearly (Huang et al., 2017).

Machine learning is a branch of AI that involves developing models that can learn from data without being explicitly programmed. Unlike traditional regression models, machine learning models can capture complex or nonlinear relationships that might exist in the data, which can lead to more accurate predictions. Additionally, machine learning models can handle large amounts of data and discover hidden patterns that might be overlooked by traditional statistical methods. To develop accurate predictive models, two commonly used machine learning models were applied in this study: Random Forest and XGBOOST. Random Forest is a decision tree-based ensemble learning algorithm that is capable of handling nonlinear relationships between input and output variables. It constructs multiple such decision trees and combines them to achieve a more precise and consistent prediction. XGBOOST, on the other hand, is an optimized gradient boosting algorithm that combines weak prediction models to make more robust and accurate forecasts. The performance of these models was evaluated using various metrics such as mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). The model with the lowest MAE and RMSE values and the highest R^2 value was considered the best performing model.

To find the best fitting model, I need to set up the hyperparameters. Hyperparameters are parameters that are set prior to the training process and cannot be learned during training. These parameters affect the behavior of the model during training and can have a significant impact on its performance. Choosing the appropriate hyperparameters is crucial for achieving the best possible performance of the model on the given task. The random forest models used two hyperparameters: "n_estimators," which set the number of trees to build before taking the maximum voting or averages of predictions. While a higher number of trees can perform better, it also generates results slower. Another hyperparameter was "random_state," which ensures that the same results are always produced when given the same parameters and training data. The other hyperparameter was “max_depth”, which limited the depth of the tree. In contrast, the hyperparameters for XGBOOST models in this study were more complicated and included max_depth, min_child_weight, gamma, subsample, colsample_bytree, and learning_rate. "max_depth" specifies the maximum depth of a tree, "min_child_weight" specifies the minimum sum of instance weight (hessian) needed in a child, "gamma" specifies the minimum loss reduction required to make a further partition on a leaf node of the tree, "subsample" specifies the subsample ratio of the training instances, "colsample_bytree" specifies the subsample ratio of columns when constructing each tree, and "learning_rate" specifies the step size shrinkage used in updates to prevent overfitting.

Overall, the machine learning techniques employed in this study allowed us to identify which environmental parameters have the strongest impact on CHL and to develop accurate predictive models that can be used to inform management decisions related to water quality. The results of the data analysis and modeling are presented in the next section.

This research is conducted in collaboration with Vaquita Technologies, a company specializing in software and sensor solutions for water quality assessment. The collaboration has provided invaluable expertise and resources, playing a vital role in the development and implementation of the research methodology. The findings presented in this study are conducted with the commitment to impartiality and objectivity. The aim of this study is to advance knowledge in the field and make meaningful contributions to the broader sustainability science community.

4. Results

4.1. Traditional statistical models

The descriptive statistics of the environmental parameters and their scaled counterparts were presented in Table 1. The average CHL was found to be 13.19 µg per liter, which is within the range of concentrations typically observed in freshwater systems. The sampling day of the year, as indicated by the variable "Day of a year," occurred on average at 188.13, approximately at the beginning of July. This suggests that the data was collected during the summer months, which is a period when freshwater systems are often susceptible to eutrophication. The latitude of Skåne län is 55.9903° N, 13.5958° E, which is consistent with the coordinates in the data, namely 56.04 ° N, 13.35 ° E. The pH value of the water samples was observed to be neutral, with a mean value of 7.66. The water temperature (T) was found to be 12.18 °C, which is also typical of freshwater systems during the summer months. The concentration of total phosphorus (P) was measured to be 38.07 µg per liter, while the concentration of total nitrogen (N) was 1351.72 µg per liter. The dissolved oxygen content was 10.81 mg per liter, which is within the acceptable range for freshwater systems. The electrical conductivity (EC) was observed to be 23.12 milli Siemens per meter, while the turbidity was found to be 4.26 FNU.

The first analysis (univariate pairwise correlations) showed weak relationships between the environmental parameters and the CHL, as indicated by the correlation coefficients between CHL and the different parameters (see Figure 2). The Spearman rank correlation test revealed that turbidity had the most significant relationship with CHL, followed by the concentration of total phosphorus and electrical conductivity. However, it should be noted that the high correlation coefficient between CHL and turbidity might be due to the presence of algae, which can block sunlight and lead to increased turbidity, while also contributing to higher CHL (Keller et al., 2018).

On the other hand, factors such as the date of the year, water temperature, latitude, longitude, dissolved oxygen, and pH value had weak or no relationships with CHL. This suggests that these environmental factors are not good predictors of CHL and that other factors, such as nutrient availability and light availability, may play a more important role. Overall, the results showed that there was little to no direct relationship between the CHL and weather factors, indicating that other environmental variables may need to be considered in order to accurately predict CHL in freshwater systems.

In the first stage of analysis, I utilized scatter plots to visualize the relationship between CHL and various environmental parameters. It shows the scatter plots for T(a), pH(b), (c) EC, (d) Turbidity, (e) P, (f) N, (g) Longitude, (h) Latitude, (i) day of a year, and (j) DO. Among these variables, temperature did not show a clear relationship with CHL, as the scatter plot exhibited a wide range of values. In contrast, scatter plots for pH, electrical conductivity, turbidity, total P concentration, and dissolved oxygen appeared to cluster in certain areas, suggesting a potential relationship with CHL. However, it is important to note that scatter plots may not always reveal the full picture of the relationships between variables. In this study, I found weak correlations between CHL and several environmental factors in the scatter plots, which is consistent with the results in Figure 2. Based on these findings and previous studies, I speculate that CHL may be influenced by a variety of environmental factors. Thus, it is crucial to consider all relevant variables in constructing multiple linear regression models to better understand the complex relationship between CHL and environmental parameters.

To ensure comparability for the rest of the analysis, I scaled the dataset. While scaling does not change the relationship between variables, it standardizes the range of values across variables, allowing for easier comparison of the magnitudes of coefficients in the MLR models.

For the second stage, I performed multiple linear regression (MLR) using the training dataset. The model was then applied to the testing data, and the results were visualized in Figure 3. The R^2 value of the multiple linear regression was 0.42, which dropped to 0.04 in the test set, indicating poor predictive performance. These findings suggest that accurately predicting CHL based solely on environmental parameters might be challenging using a linear model. However, it is still possible that there are non-linear relationships between the variables. To explore this possibility, I employed polynomial regression as an alternative approach.

Although the polynomial regression model appeared to fit well when applied to the training set, as shown in Figure 4, there was a significant deviation from the model when applied to the test set. This result suggests that polynomial regression may not be a suitable approach for predicting CHL in this case.

Table 1. Descriptive statistics of environmental parameters.

	CHL	Day of a year	Latitude	Longitude	pH	Temperature	P	DO	EC	Turbidity	N
Mean	13.19	188.13	56.04	13.35	7.66	12.18	38.07	10.81	23.12	4.26	1351.72
STD	7.12	88.78	0.21	0.18	0.58	6.56	73.47	1.74	9.90	2.58	604.53
Min	4.10	14.00	55.48	13.04	6.38	0.60	5.00	6.70	7.50	0.90	410.00
25%	6.70	133.00	55.89	13.28	7.20	5.50	18.00	9.60	10.90	2.60	900.00
50%	12.00	194.00	56.09	13.31	7.66	12.80	27.00	10.60	25.50	3.60	1300.00
75%	19.00	230.00	56.28	13.55	8.10	18.30	38.00	12.00	29.30	5.30	1700.00
Max	27.00	349.00	56.35	13.98	9.04	22.90	880.00	15.80	51.90	16.00	4100.00

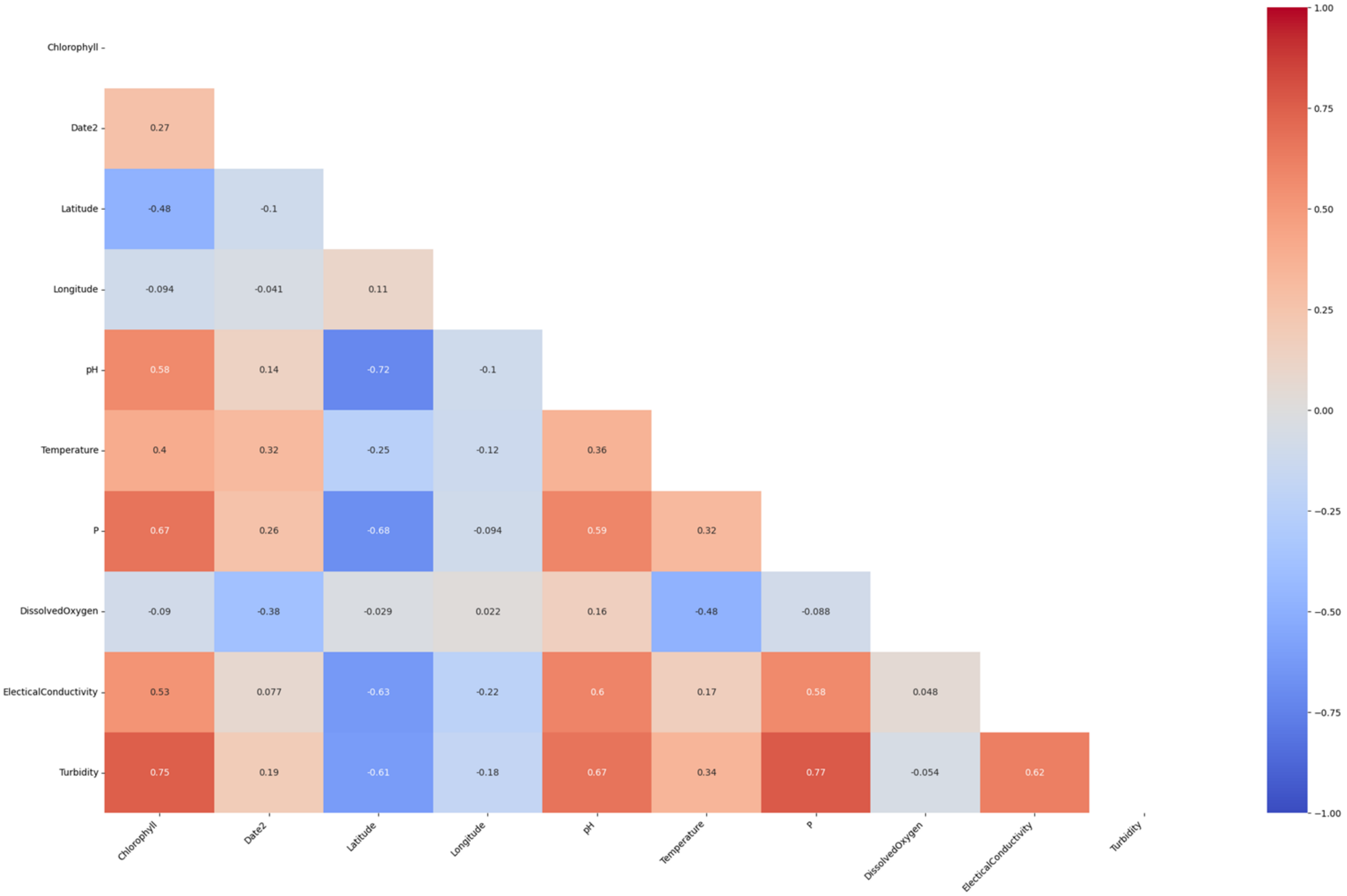

Figure 2. The heatmap of Spearman's rank correlation coefficient (r_s).

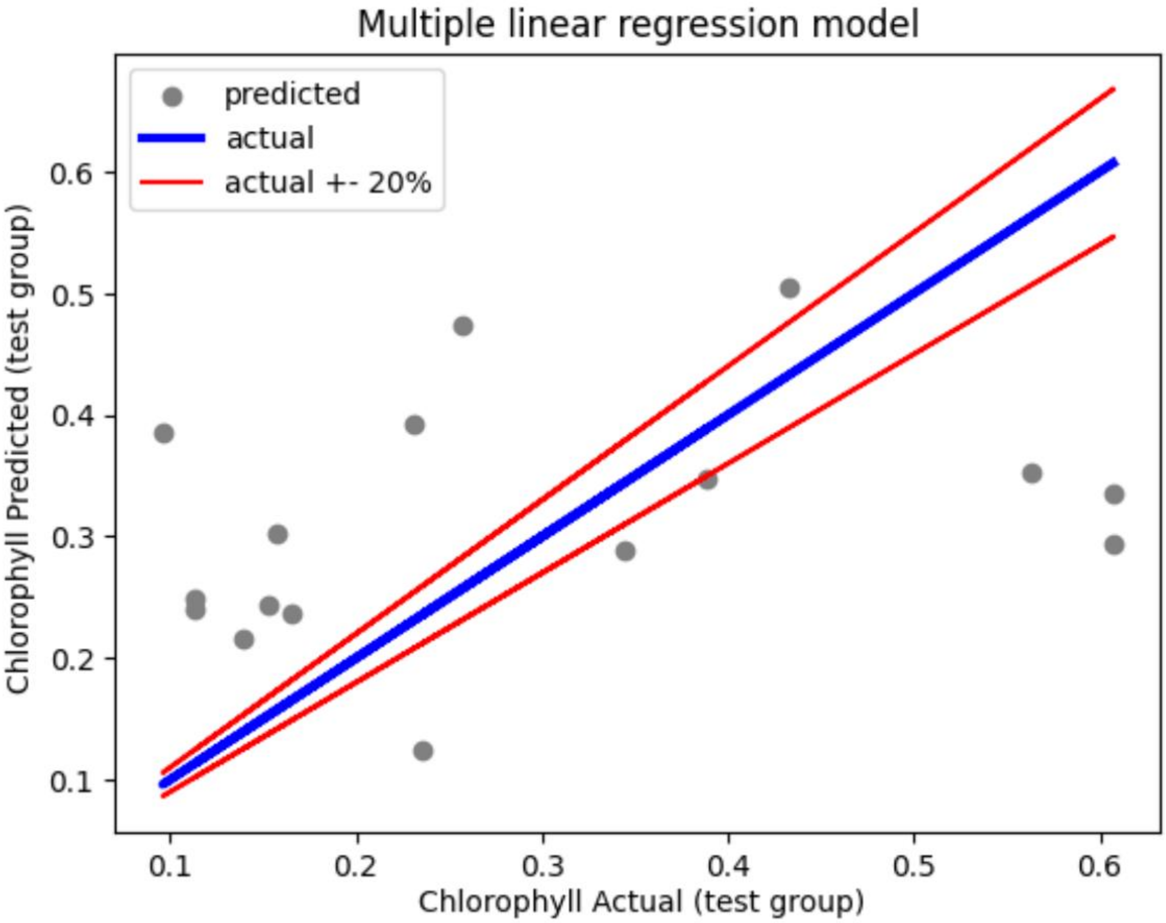

Figure 3. Multiple linear regression model with testing dataset generated by Python.

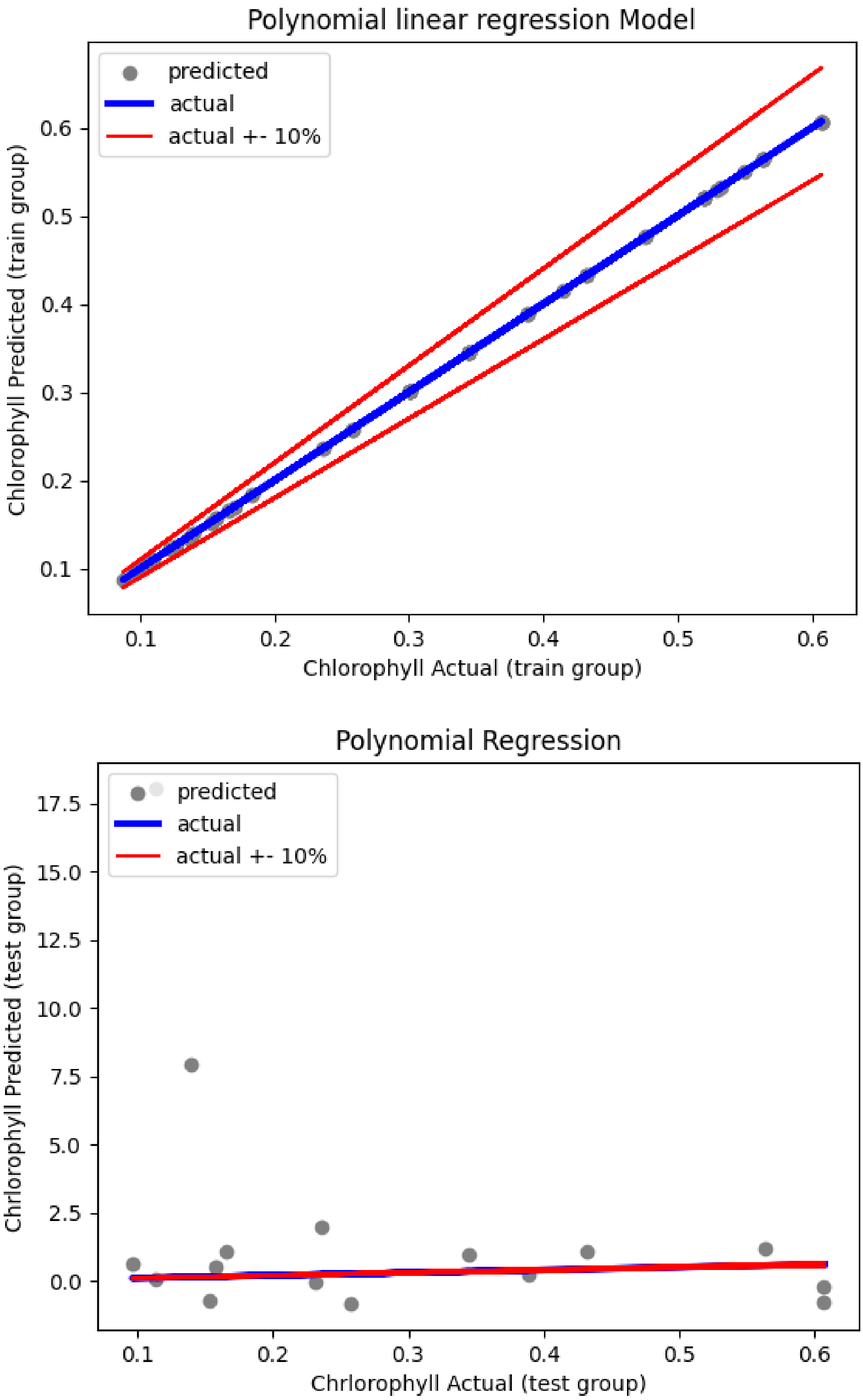

Figure 4. Polynomial Regression generated by Python.

4.2. Machine Learning Models

To identify the most effective models for predicting chlorophyll concentration, I applied two existing machine learning (ML) algorithms, Random Forest (RF) and XGBOOST model, to the dataset and evaluated their performance.

As illustrated in Figure 5(a) and 5(b), the Random Forest model produced a much better fit on the training data compared to the traditional statistical models, with an R^2 value of 0.88. The model also performed well in predicting the CHL in the test set, with an R^2 value of 0.69, which was higher than the multiple linear regression and polynomial models. These findings suggest that the Random Forest model can effectively predict the CHL based on environmental parameters.

Next, the XGBOOST model was used to predict the CHL. Although the XGBOOST model's performance on the test set was not as strong as the Random Forest model, with a mean squared error (MSE) of 72.94, the training set was well-predicted, as shown in Figure 5(c) and (d). MSE suggests that the XGBOOST model's predictions on the test set were not as accurate as the RF model, which had a higher R-squared value. The resulting hyperparameters for the XGBOOST model were colsample_bytree = 0.6, gamma = 0.2, learning rate = 0.1, max depth = 7, min_child_weight = 1, and subsample = 1.0. The results of the XGBOOST model were similar to those of the Random Forest model, indicating the potential of machine learning models in predicting CHL.

Overall, my findings suggest that machine learning models, such as Random Forest, are highly effective in predicting CHL based on environmental parameters. This model is particularly promising for future water quality prediction and monitoring efforts in Skåne län. By providing insight into the complex relationships between environmental parameters and CHL, machine learning models can aid in better understanding and managing the water quality in the region.

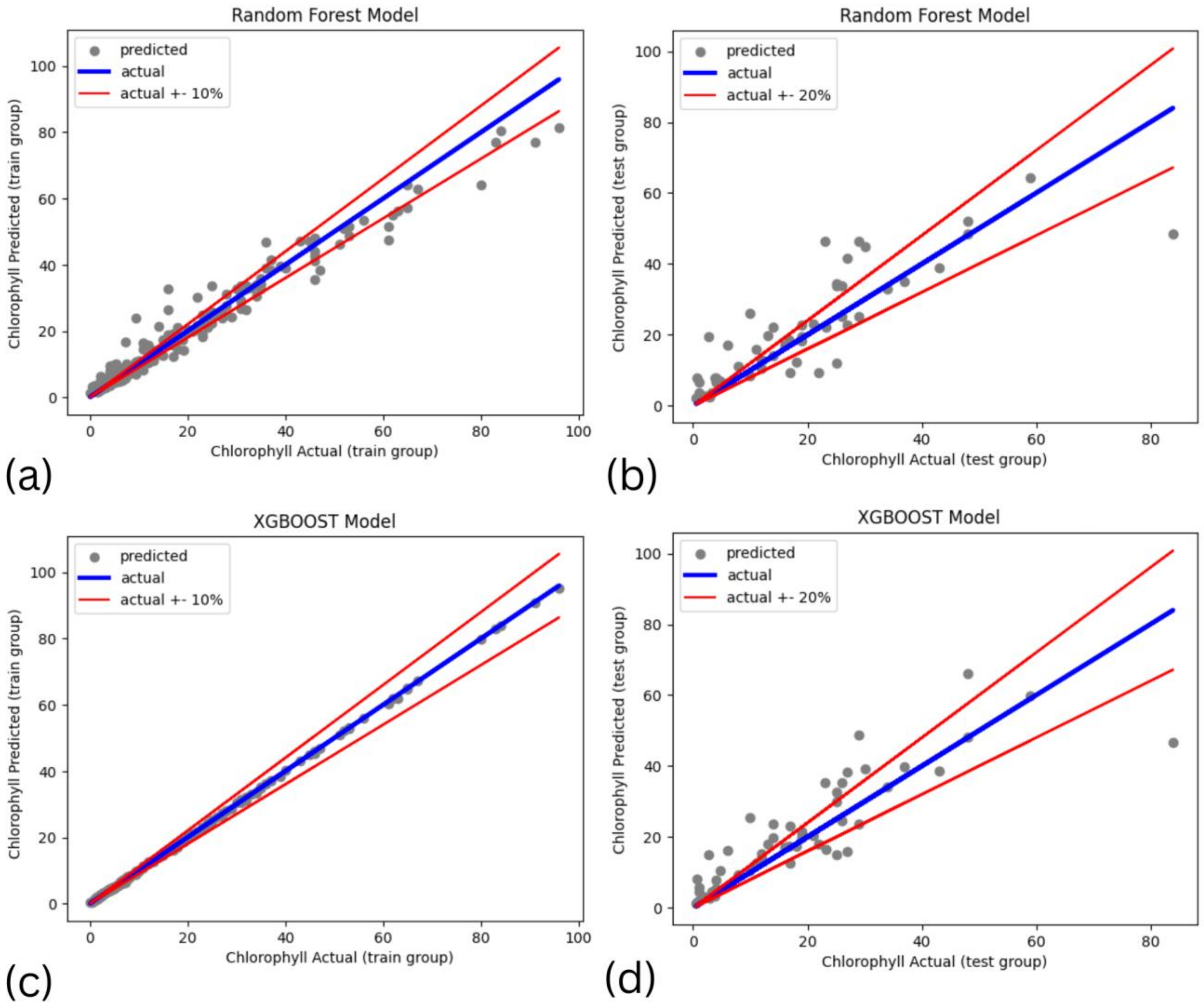

Figure 5. Random Forest models and XGBOOST models generated by Python. (a) Random Forest model based on training dataset. (b) Random Forest model based on testing dataset. (c) XGBOOST model based on training dataset. (D) XGBOOST model based on testing dataset.

5. Discussion

5.1. Using Statistics to Predict Water Quality

To investigate the relationship between basic environmental factors and chlorophyll concentrations (CHL), this study aimed to determine the extent of predictive power of various models based on historical data. The main research question was supported by four sub-questions that explored the efficacy of univariate linear, multiple linear regression, polynomial linear regression, and machine learning models in forecasting CHL. Analysis of the results revealed that basic environmental factors such as EC, DO, and pH may not have a strong linear relationship with CHL, at least based on the historical dataset in Skåne län. The simple linear regression model and polynomial model were found to be inadequate in achieving accurate predictions. However, the random forest (RF) model produced significantly better forecasts.

To illustrate the inner workings of the RF model, Figure 6 presents a decision tree that represents its underlying structure. Random forests are composed of multiple decision trees, each of which is generated through a random selection of features and observations. With the aid of this decision tree, the visualization of the results of the Random Forest (RF) model is made easier. This is because the decision tree breaks down the complex process of the RF model into a series of simple, logical steps that can be easily understood. By following the path of the decision tree, it is possible to see the specific criteria that the model uses to classify the data and make predictions.

In Figure 6, the decision tree starts by identifying that all the turbidity levels of the data are below 11.5 FNU. This leads to the data set being divided into two paths: the green path to the right for data with turbidity below 11.5 FNU, and the red path to the left for data with turbidity levels above 11.5 FNU. Figures 6(a) and 6(b) provide a closer look at the original figure 6. Following the green path, the data is subjected to another specific criterion, where P (phosphorus) levels are checked against a threshold value of 107.0 µg/L. If P is higher than this value, the data takes the red path to the left. If it is lower, the data continues along the green path. The next box in the green path checks if T (temperature) is below 20 °C. If T meets this criterion, the decision tree leads to a prediction of CHL (chlorophyll) at 25.0 µg/L. The initial model was configured to have a maximum depth of 4 layers. However, if the dataset size is larger, the depth can be increased up to 100 layers to improve the precision of the results. Overall, this decision tree helps to provide a clear and structured framework for understanding how the RF model operates. It also helps to highlight the specific variables and criteria that the model uses to make predictions. This information is essential for interpreting the results of the model and for assessing its accuracy and reliability. (The decision tree figure was recreated using an online decision tree maker called Miro, using the original model generated from Python.)

In our study, we found that the RF model outperformed traditional statistical models and XGBOOST in terms of forecasting accuracy. This finding is consistent with other studies such as Mozo et al., who also reported better results with the RF model (Mozo et al., 2022). For example, a study conducted by Xu showed that the RF model could achieve an accuracy rate of up to 92.94% for nearshore waters (Xu et al., 2021). However, it should be noted that some studies have pointed out that while the RF model can predict in the right direction, its accuracy still needs improvement, as reported by Shin et al. (2020) and Yajima & Derot (2017). Furthermore, there are some studies that suggest XGBOOST performs better than the RF model in certain cases. For example, a study reported a higher accuracy rate of 96.9696% using XGBOOST (Garabaghi et al., 2022). However, it is important to note that the choice of the model ultimately depends on the specific characteristics of the data and the research objectives. In our case, the RF model was the most suitable for our data set and research question.

In line with our findings, a recent study conducted by Huang in 2022 also concluded that P is the primary factor influencing CHL, while turbidity does not have a significant impact (Huang et al., 2022). This reinforces the validity of our research and suggests that our model is effective in identifying key factors affecting CHL levels. Additionally, it is worth noting that during the development of our model, we excluded outliers to prevent bias. However, a recent study has revealed that RF models have good noise immunity and are not highly sensitive to outliers (Xu et al., 2021). This finding suggests that the exclusion of outliers may not be necessary when using RF models for CHL forecasting. Further research is needed to explore the extent to which outliers can be included in the data set without negatively impacting the accuracy of the model. Overall, the combination of our findings and those from other studies suggests that RF models are a promising tool for forecasting CHL levels and have potential applications in a variety of contexts.

One limitation of the study was that the remaining data collected over almost 50 years was fewer than expected, as I wanted the full data with all parameter values to be applied in the model. Additionally, the experimental methods used to measure the total concentration of phosphorus and nitrogen were not always the same, as the methods improved over time. However, the data is comparable. Despite these limitations, the RF model was able to process predictions even with limited amounts of data. As studies about diffusion and water quality monitoring with machine learning are still relatively scarce, the RF model's ability to work with limited data is a crucial advantage.

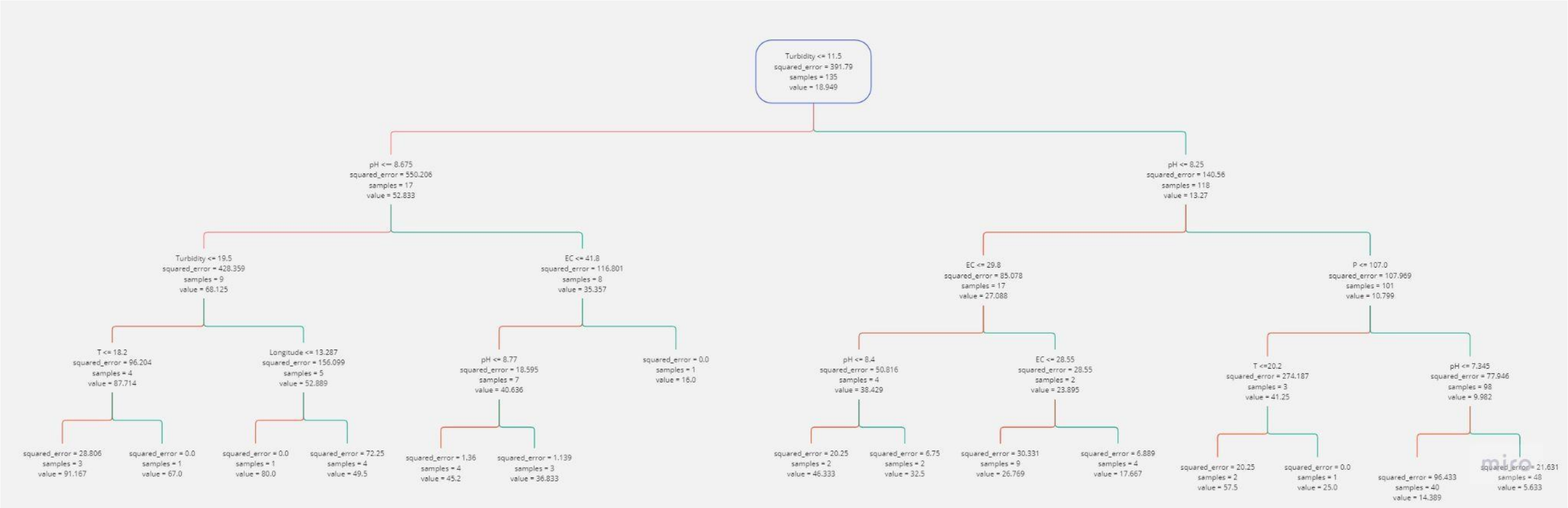

Figure 6. Decision Trees based on the Random Forest model. (Full picture)

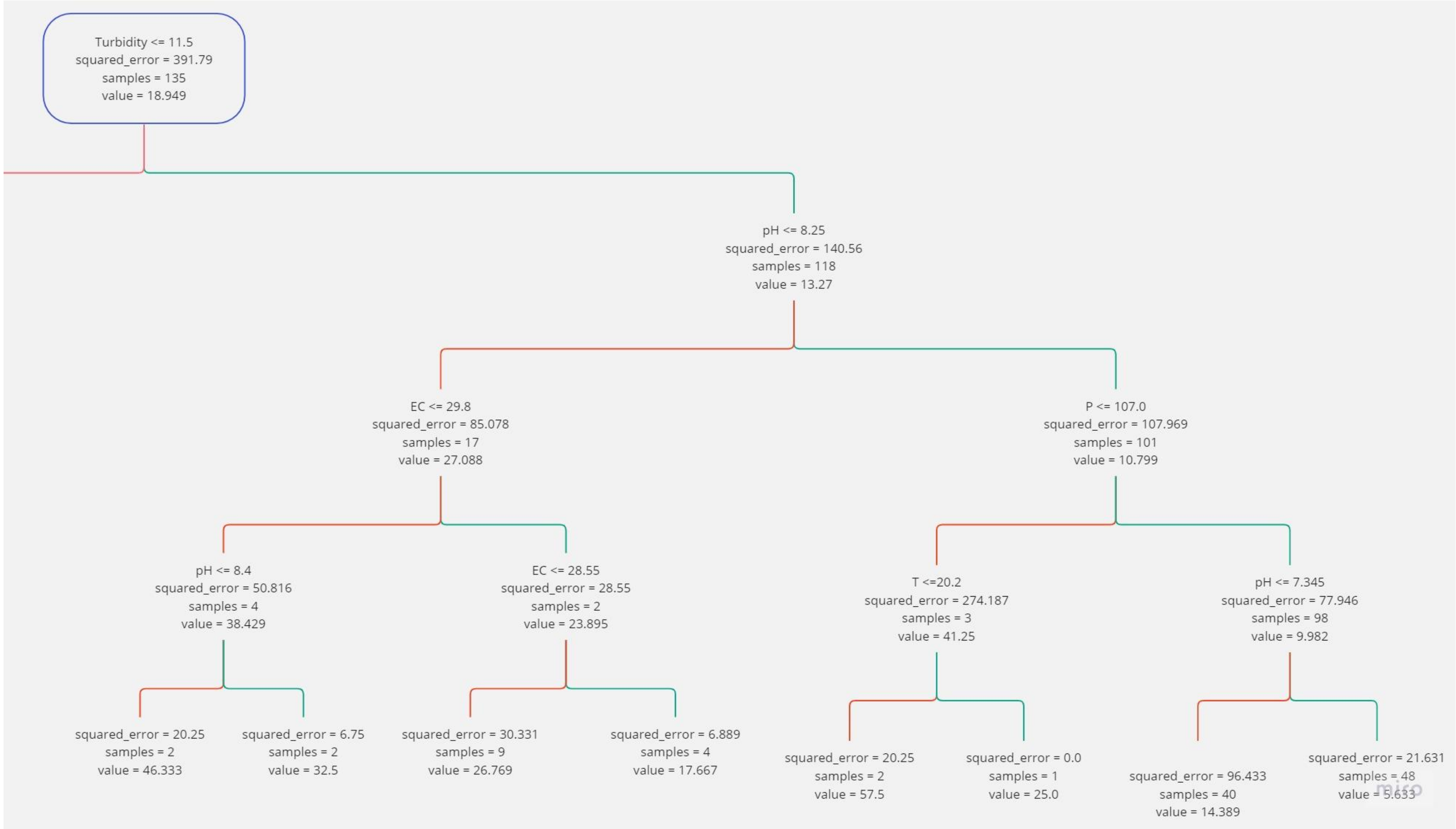

Figure 6. (a) Decision Trees based on the Random Forest model.

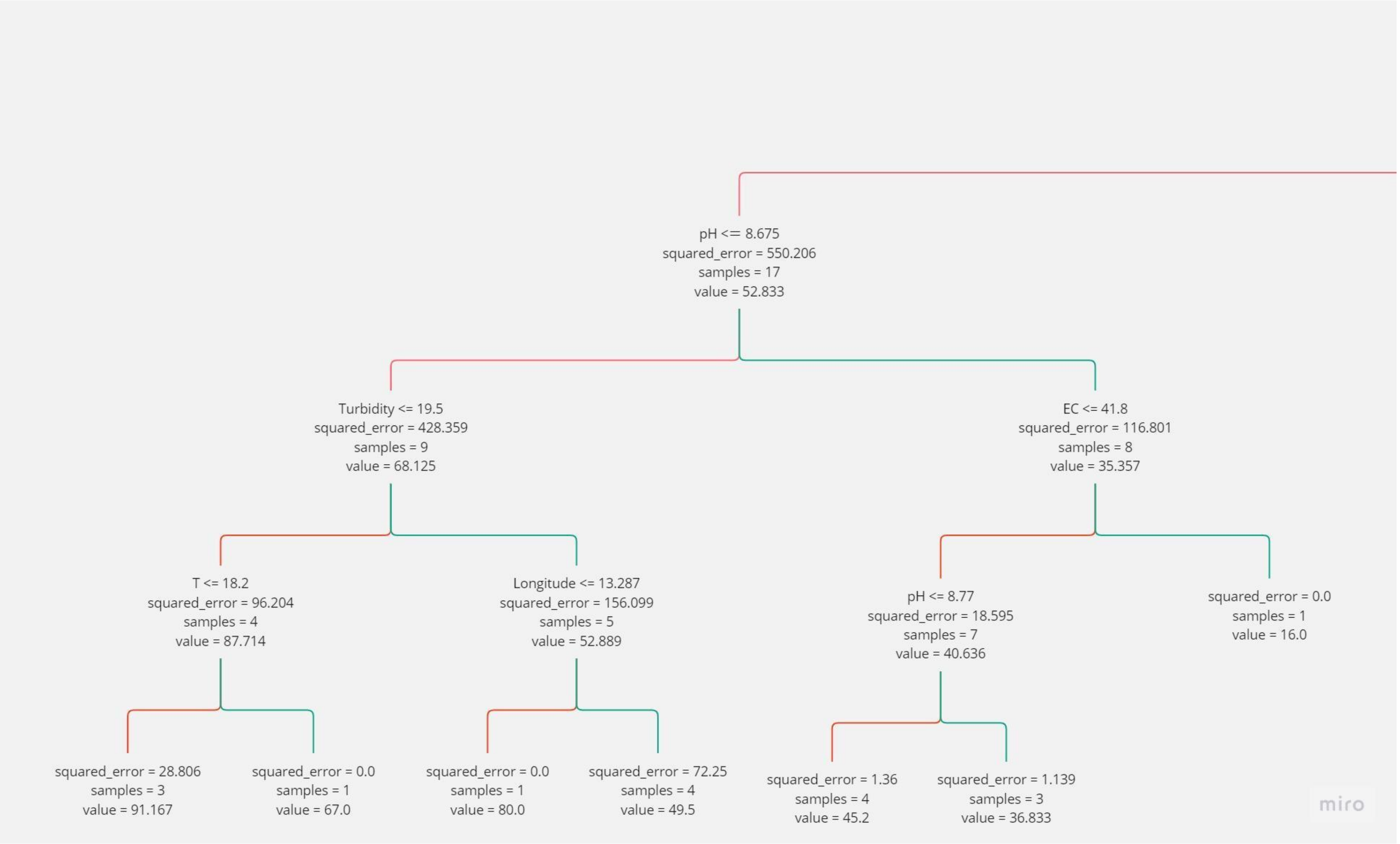

Figure 6. (b) Decision Trees based on the Random Forest model.

5.2. Technological Solutions for Sustainable Development

Technological innovation has the potential to make significant contributions to sustainable development, but it requires the diffusion of technology. However, advanced technology is often expensive, making it more accessible to wealthy countries in the Global North than to poor countries in the Global South. Technology diffusion is the process by which a new technology gains acceptance, ownership, and use by the members of a social group (Odekon, 2015). Most industrial nations' economic growth has been significantly influenced by the international diffusion of technology (Soete, 1985). The United Nations' promotion of the SDGs has brought attention to the importance of sustainable development. For example, monitoring and managing the water quality is a common issue in the Global South, where countries often face geographical limitations. Advanced technology, such as monitoring water quality using precise sensors or equipment, is not as easily accessible to the Global South as it is to the Global North. While the UN raised the concerns of water quality and the drinking water security remains a prominent issue that affects socioeconomic and human development, the countries in the Global South often face technical and financial limitations. Such limitations may affect the applicability of easy and affordable methods, raising the question of whether these methods can provide a sustainable solution for underprivileged communities. ML methods have been examples of technology innovation that apply efficiently in the water quality in simulating different types of problems. In my vision, I expect that I can use cheaper sensors to approach basic parameters and get the CHL from the ML model. Nevertheless, the findings of the study indicate that the prediction of CHL based solely on the selected environmental parameters was not entirely successful.

Water quality management is a common issue in the Global South, where countries often face technical and financial limitations. This raises the question of whether affordable methods can provide a sustainable solution for underprivileged communities. In this study, I explore the use of machine learning models to achieve accessible and approachable access to clean and safe water resources.

In the context of technology diffusion, water quality has been recognized as a significant concern. While this study has made significant progress in developing a model for forecasting chlorophyll and harmful algal blooms, further improvements are necessary to create a more comprehensive and reliable model. This may include the integration of real-time sensor data to improve the accuracy and reliability of the model. Additionally, the results of this study may be used to encourage further research and development in the field of water quality monitoring and management. One of the key benefits of this model is that it may be implemented in countries or areas with limited laboratory resources, providing decision-makers with a valuable tool for managing water quality. However, this will require the availability of suitable sensors, and sufficient data to build reliable models in new locations. In order to collect accurate and comprehensive data, a range of basic environmental parameter sensors such as thermometer, pH meter, EC, and DO sensors are necessary. Additionally, precision equipment such as chlorophyll sensors and sensors for P or N will also be required to obtain more detailed information. The use of such sensors will ensure that the collected data is both reliable and comprehensive, allowing for the development of accurate models that can be used for forecasting chlorophyll and harmful algal blooms. In the long term, financial assistance may be required to facilitate the diffusion of this technology and its eventual confirmation as a valuable tool for sustainable water management practices. Therefore, this study represents an essential step towards the goal of promoting sustainable water management practices globally, and it provides a foundation for further research in the field of water quality monitoring and management.

The diffusion of the model developed in this study would require careful consideration of the key elements of innovation, adopters, communication channels, time, and the social system (Rogers, 2003). The innovation is the use of machine learning to monitor and manage water quality, which may require significant investment in technology and personnel. The adopters of the technology could potentially encompass a wide range of stakeholders, including water management agencies, researchers, policymakers, and other relevant actors in sectors such as agriculture, and tourism. Communication channels could include scientific journals, conferences, and workshops to disseminate information about the model and its potential benefits. The time needed for adoption and adaptation of the model may be long, and ongoing monitoring and evaluation would be necessary to ensure its effectiveness. Finally, the social system, which includes the broader social, economic, and political systems in which the model would operate, would need to be considered. This would include factors such as governance structures, funding mechanisms, and public perception of the technology. Overall, the successful diffusion of the model would require a collaborative and interdisciplinary approach involving various stakeholders and addressing the challenges of adoption and implementation in different social contexts.

The exploration and development of alternative approaches to promote the widespread adoption of advanced technology in the Global South is crucial. Collaboration between developed and developing countries can facilitate the diffusion of technology, build capacity, share knowledge, and transfer technology, all of which can contribute to sustainable development. In addition, engaging local communities in the development and implementation of technology solutions can increase their sense of ownership, promoting the sustainability of the solutions. By working collaboratively with innovators, investors, and policymakers, sustainable technologies can be more effectively promoted and adopted, facilitating their diffusion. It is important to note that economic development alone does not necessarily solve the problem of poor water quality but rather transforms it and lays the foundation for future solutions. By considering these approaches and exploring the challenges and opportunities associated with technology diffusion, a better understanding of how technology can be effectively diffused to promote sustainable water management practices in the Global South can be achieved.

5.3. Limitations and Future Research

This study has several limitations that should be acknowledged to provide a clear understanding of the research findings. Firstly, the dataset employed in this study only covers a limited time range due to the experimental data's lack of integrity. Although this study still achieved significant findings, some models require a massive amount of data to build a comprehensive and precise model, and a more extended time range could yield more robust results. Secondly, the study does not include weather data due to the weak relationships observed. The original data did not record the sampling time, making it impossible to obtain accurate weather data. Therefore, the exclusion of weather data could have potentially influenced the results. Thirdly, this study only considers basic environmental factors and uses data solely from Sweden. Although these factors provide insight into the relationship between chlorophyll concentration and the environment, the addition of more parameters, such as biochemical oxygen demand (BOD), chemical oxygen demand (COD), and dissolved organic matter, could enhance the model's accuracy and provide further insight. Furthermore, the impact of environmental factors on chlorophyll concentration in water bodies may differ across countries due to differences in geography, climate, and land use patterns. Therefore, it is crucial to test the model's performance on data from different countries to ensure its validity and reliability for the diffusion of technology. However, given the scope and limitations of the study, the current selection of variables was deemed appropriate.

In future studies, real-time monitoring should be considered to combine with the forecast models to improve or verify the accuracy of the models. This approach has the potential to significantly enhance water resource management, protect aquatic ecosystems and human health, and contribute to achieving the Sustainable Development Goals related to water. As discussed in the limitations section, more complete water quality data is needed to develop a more robust model from different countries, and other factors, such as latent factors that may require expensive sensors to study, should also be included. Furthermore, field studies in countries in the Global South can be conducted to test the model's performance and ensure that these countries can also benefit from the diffusion of technology. In addition, investigating the effect of different land use patterns and agricultural practices on chlorophyll concentrations in water bodies can provide valuable insight into the factors that influence water quality.

To further promote the diffusion of technology for sustainable water management practices in the Global South, it is important to consider the role of government policies and regulations. From training local technicians and engineers to developing local factories and manufacturing industries, it will help the locals step away from the threat of poverty gradually and decrease the unemployment rate. Creating a supportive regulatory system and environment is also crucial to protect the environment and monitor the pollutants and hazards in the wastewater. With the support, the technology could adapt to the local context and meanwhile evaluate sustainability with economic growth.

6. Conclusion

This study investigated the use of statistical and machine learning models to predict chlorophyll concentrations based on environmental parameters in Skåne län, Sweden. The results indicated that the Random Forest (RF) model was the most effective in producing accurate predictions. Despite the study's limitations, such as a small dataset and inconsistent measurement methods for total phosphorus and nitrogen, it examined the diffusion of technology in water quality management for sustainable development, particularly in the Global South. The challenges and opportunities of technology diffusion were discussed, and the importance of collaboration with local communities to promote sustainable technologies was emphasized. Future research could investigate the use of machine learning models for water quality monitoring and explore alternative approaches to promote technology diffusion. By addressing these issues, the study suggests that sustainable water management practices can be promoted and that access to clean and safe water resources can be improved for all communities, regardless of their economic status.

7. Reference

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, *11*(11), Article 11. https://doi.org/10.3390/w11112210

Alminagorta, O., Loewen, C. J. G., de Kerckhove, D. T., Jackson, D. A., & Chu, C. (2021). Exploratory analysis of multivariate data: Applications of parallel coordinates in ecology. *Ecological Informatics*, *64*, 101361. https://doi.org/10.1016/j.ecoinf.2021.101361

Andersen, J. H., Carstensen, J., Conley, D. J., Dromph, K., Fleming-Lehtinen, V., Gustafsson, B. G., Josefson, A. B., Norkko, A., Villnäs, A., & Murray, C. (2017). Long-term temporal and spatial trends in eutrophication status of the Baltic Sea. *Biological Reviews*, *92*(1), 135–149. https://doi.org/10.1111/brv.12221

Beretta-Blanco, A., & Carrasco-Letelier, L. (2021). Relevant factors in the eutrophication of the Uruguay River and the Río Negro. *Science of The Total Environment*, *761*, 143299. https://doi.org/10.1016/j.scitotenv.2020.143299

Boesch, D. F., Brinsfield, R. B., & Magnien, R. E. (2001). Chesapeake Bay eutrophication: Scientific understanding, ecosystem restoration, and challenges for agriculture. *Journal of Environmental Quality*, *30*(2), 303–320. https://doi.org/10.2134/jeq2001.302303x

Boyer, J. N., Kelble, C. R., Ortner, P. B., & Rudnick, D. T. (2009). Phytoplankton bloom status: Chlorophyll a biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. *Ecological Indicators*, *9*(6), S56–S67. https://doi.org/10.1016/j.ecolind.2008.11.013

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campolo, M., Andreussi, P., & Soldati, A. (1999). River flood forecasting with a neural network model. *Water Resources Research*, *35*(4), 1191–1197. https://doi.org/10.1029/1998WR900086

Carmichael, W. W., & Boyer, G. L. (2016). Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes. *Harmful Algae*, *54*, 194–212. https://doi.org/10.1016/j.hal.2016.02.002

Carneiro, F. M., Nabout, J. C., Vieira, L. C. G., Roland, F., & Bini, L. M. (2014). Determinants of chlorophyll-a concentration in tropical reservoirs. *Hydrobiologia*, *740*(1), 89–99. https://doi.org/10.1007/s10750-014-1940-3

Chapman, D. V., Organization, W. H., UNESCO, & Programme, U. N. E. (1996). *Water quality assessments: A guide to the use of biota, sediments and water in environmental monitoring*. E & FN Spon. https://apps.who.int/iris/handle/10665/41850

Charulatha, G., Srinivasalu, S., Uma Maheswari, O., Venugopal, T., & Giridharan, L. (2017). Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arabian Journal of Geosciences*, *10*(6), 128. https://doi.org/10.1007/s12517-017-2867-6

Chen, J., Zhang, M., Cui, T., & Wen, Z. (2013). A Review of Some Important Technical Problems in Respect of Satellite Remote Sensing of Chlorophyll-a Concentration in Coastal Waters. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal Of*, *6*, 2275–2289. https://doi.org/10.1109/JSTARS.2013.2242845

Chen, M., Li, J., Dai, X., Sun, Y., & Chen, F. (2011). Effect of phosphorus and temperature on chlorophyll a contents and cell sizes of Scenedesmus obliquus and Microcystis aeruginosa. *Limnology*, *12*(2), 187–192. https://doi.org/10.1007/s10201-010-0336-y

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cho, K. H., Sthiannopkao, S., Pachepsky, Y. A., Kim, K.-W., & Kim, J. H. (2011). Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. *Water Research*, *45*(17), 5535–5544.

https://doi.org/10.1016/j.watres.2011.08.010

Chorus, I., & Welker, M. (Eds.). (2021). *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*. Taylor & Francis. https://doi.org/10.1201/9781003081449

Conley, D. J., Paerl, H. W., Howarth, R. W., Boesch, D. F., Seitzinger, S. P., Havens, K. E., Lancelot, C., & Likens, G. E. (2009). Controlling Eutrophication: Nitrogen and Phosphorus. *Science*, *323*(5917), 1014–1015. https://doi.org/10.1126/science.1167755

de Vita, C. G., Mellone, G., Barchiesi, F., Di Luccio, D., Ciaramella, A., & Montella, R. (2022). Artificial Intelligence for mussels farm quality assessment and prediction. *2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, 33–38. https://doi.org/10.1109/MetroSea55331.2022.9950875

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Dimitra Kitsiou & Michael Karydis. (2019). *Marine Eutrophication: A global perspective* (Electronic resources). CRC Press.

Dodds, W. K., Bouska, W. W., Eitzmann, J. L., Pilger, T. J., Pitts, K. L., Riley, A. J., Schloesser, J. T., & Thornbrugh, D. J. (2009). Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages. *Environmental Science & Technology*, *43*(1), 12–19. https://doi.org/10.1021/es801217q

Fang, X., Li, X., Zhang, Y., Zhao, Y., Qian, J., Hao, C., Zhou, J., & Wu, Y. (2021). Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon. *Environmental Research Letters*, *16*(5), 055003. https://doi.org/10.1088/1748-9326/abf395

Fausett, L. V. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall.

Garabaghi, F. H., Benzer, S., & Benzer, R. (2022). *Performance Evaluation of Machine Learning*

Models with Ensemble Learning Approach in Classification of Water Quality Indices Based on Different Subset of Feature. https://doi.org/10.21203/rs.3.rs-876980/v2

Garcia-Hernandez, J. A., Brouwer, R., & Pinto, R. (2022). Estimating the Total Economic Costs of Nutrient Emission Reduction Policies to Halt Eutrophication in the Great Lakes. *Water Resources Research*, *58*(4), e2021WR030772. https://doi.org/10.1029/2021WR030772

Gardner-Dale, D. A., Bradley, I. M., & Guest, J. S. (2017). Influence of solids residence time and carbon storage on nitrogen and phosphorus recovery by microalgae across diel cycles. *Water Research*, *121*, 231–239. https://doi.org/10.1016/j.watres.2017.05.033

Gregersen, R., Howarth, J. D., Atalah, J., Pearman, J. K., Waters, S., Li, X., Vandergoes, M. J., & Wood, S. A. (2023). Paleo-diatom records reveal ecological change not detected using traditional measures of lake eutrophication. *Science of The Total Environment*, *867*, 161414. https://doi.org/10.1016/j.scitotenv.2023.161414

Gurjar, S. K., & Tare, V. (2019). Spatial-temporal assessment of water quality and assimilative capacity of river Ramganga, a tributary of Ganga using multivariate analysis and QUEL2K. *Journal of Cleaner Production*, *222*, 550–564. https://doi.org/10.1016/j.jclepro.2019.03.064

Hanson, K., Puplampu, P., & Shaw, T. (2017). From Millennium Development Goals to Sustainable Development Goals: Rethinking African Development. In *From Millennium Development Goals to Sustainable Development Goals: Rethinking African Development*. https://doi.org/10.4324/9781315228068

Haykin, S. S. (1994). *Neural networks: A comprehensive foundation* (E-husets bibliotek LTH 1.5 Haykin). Macmillan.

Henestål, J., Ranung, J., Gyllander, A., Johnsen, Å., Olsson, H., Pettersson, O., Westman, Y., & Wingqvist, E.-M. (2021). *Arbete med SVAR version 2012_1 och 2012_2, Svenskt Vattenarkiv, en databas vid SMHI | SMHI*. SMHI. https://www.smhi.se/publikationer/publikationer/arbete-med-svar-version-2012-1-och-2012-2-svenskt-vattenarkiv-en-databas-vid-smhi-1.171026

Howarth, R. W., Swaney, D. P., Boyer, E. W., Marino, R., Jaworski, N., & Goodale, C. (2006). The influence of climate on average nitrogen export from large watersheds in the Northeastern United States. In L. A. Martinelli & R. W. Howarth (Eds.), *Nitrogen Cycling in the Americas: Natural and Anthropogenic Influences and Controls* (pp. 163–186). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5517-1_8

Huang, H., Wang, W., Lv, J., Liu, Q., Liu, X., Xie, S., Feng, J., & Wang, F. (2022). Relationship between Chlorophyll a and Environmental Factors in Lakes Based on the Random Forest Algorithm. *Water (Switzerland)*, *14*(19). https://doi.org/10.3390/w14193128

Huang, H., Wang, Z., Xia, F., Shang, X., Liu, Y., Zhang, M., Dahlgren, R. A., & Mei, K. (2017). Water quality trend and change-point analyses using integration of locally weighted polynomial regression and segmented regression. *Environmental Science and Pollution Research International*, *24*(18), 15827–15837. https://doi.org/10.1007/s11356-017-9188-x

IVL. (2021, July 6). *New report shows continued problems with acidification and eutrophication in southern Sweden*. https://www.ivl.se/english/ivl/press/press-releases/2021-07-06-new-report-shows-continued-problems-with-acidification-and-eutrophication-in-southern-sweden.html

Jamil, N. R., & Shehab, Z. N. (2021). Landscape Perspective to River Pollution: A Case Study of Bentong River, Malaysia. In A. Singh, M. Agrawal, & S. B. Agrawal (Eds.), *Water Pollution and Management Practices* (pp. 19–39). Springer. https://doi.org/10.1007/978-981-15-8358-2_2

Jimeno-Sáez, P., Senent-Aparicio, J., Cecilia, J. M., & Pérez-Sánchez, J. (2020). Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). *International Journal of Environmental Research and Public Health*, *17*(4), Article 4. https://doi.org/10.3390/ijerph17041189

K, S., M, P., B, N., & S, S. (2023). *An Efficient Machine Learning Approaches for Crop Recommendation based on Soil Characteristics*. 71–76. https://doi.org/10.1109/ICEARS56392.2023.10085361

Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake, A., & Hinz, S. (2018). Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll a, Diatoms, Green Algae and Turbidity. *International Journal of Environmental Research and Public Health*, *15*(9), Article 9. https://doi.org/10.3390/ijerph15091881

Khullar, S., & Singh, N. (2020). Machine learning techniques in river water quality modelling: A research travelogue. *Water Supply*, *21*(1), 1–13. https://doi.org/10.2166/ws.2020.277

Kim, E., Park, H., & Jang, J. (2019). Development of a Class Model for Improving Creative Collaboration Based on the Online Learning System (Moodle) in Korea. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(3), Article 3. https://doi.org/10.3390/joitmc5030067

Kim, J., Yu, J., Kang, C., Ryang, G., Wei, Y., & Wang, X. (2022). A novel hybrid water quality forecast model based on real-time data decomposition and error correction. *Process Safety and Environmental Protection*, *162*, 553–565. https://doi.org/10.1016/j.psep.2022.04.020

Kravchenko, Y., Starkova, O., Herasymenko, K., & Kharchenko, A. (2017). Technology analysis for smart home implementation. *2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, 579–584. https://doi.org/10.1109/INFOCOMMST.2017.8246467

Li, T., Li, S., Liang, C., Bush, R. T., Xiong, L., & Jiang, Y. (2018). A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. *Journal of Cleaner Production*, *190*, 1–11. https://doi.org/10.1016/j.jclepro.2018.04.121

Liu, Y., Guo, H., & Yang, P. (2010). Exploring the influence of lake water chemistry on chlorophyll a: A multivariate statistical model analysis. *Ecological Modelling*, *221*(4), 681–688. https://doi.org/10.1016/j.ecolmodel.2009.03.010

Malek, S., Syed Ahmad, S. M., Singh, S. K. K., Milow, P., & Salleh, A. (2011). Assessment of predictive models for chlorophyll-a concentration of a tropical lake. *BMC Bioinformatics*, *12 Suppl*

13(Suppl 13), S12. https://doi.org/10.1186/1471-2105-12-S13-S12

Mavukkandy, M. O., Karmakar, S., & Harikumar, P. S. (2014). Assessment and rationalization of water quality monitoring network: A multivariate statistical approach to the Kabbini River (India). *Environmental Science and Pollution Research*, *21*(17), 10045–10066. https://doi.org/10.1007/s11356-014-3000-y

Mendenhall, W., & Sincich, T. (2012). *A second course in statistics: Regression analysis* (7th, International ed ed.). Prentice Hall.

Michalski, R., & Anderson, J. R. (1982, October 1). *Machine learning—An artificial intelligence approach*. Symbolic computation. https://www.semanticscholar.org/paper/Machine-learning-an-artificial-intelligence-Michalski-Anderson/b6df5c2ac2f91d71b1d08d76135e2a470ac1ad1e

Mozo, A., Morón-López, J., Vakaruk, S., Pompa-Pernía, Á. G., González-Prieto, Á., Aguilar, J. A. P., Gómez-Canaval, S., & Ortiz, J. M. (2022). Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. *Scientific Reports*, *12*, 13529. https://doi.org/10.1038/s41598-022-17299-5

Naturvardsverket. (n.d.). *Zero Eutrophication* [Government News]. Zero Eutrophication. Retrieved April 16, 2023, from https://www.naturvardsverket.se/en/environmental-work/swedish-environmental-objectives/zero-eutrophication/

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Newcomer Johnson, T. A., Kaushal, S. S., Mayer, P. M., Smith, R. M., & Sivirichi, G. M. (2016). Nutrient Retention in Restored Streams and Rivers: A Global Review and Synthesis. *Water*, *8*(4), Article 4. https://doi.org/10.3390/w8040116

Noori, R., Sabahi, M. S., Karbassi, A. R., Baghvand, A., & Taati Zadeh, H. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. *Desalination*, *260*(1), 129–136. https://doi.org/10.1016/j.desal.2010.04.053

Odekon, M. (2015). *The SAGE Encyclopedia of World Poverty* (Second Edition, Vol. 1–5). https://doi.org/10.4135/9781483345727

O'Sullivan, P. E., & Reynolds, C. S. (2004). *The lakes handbook*. Blackwell Science.

Palmer, S. C. J., Kutser, T., & Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment*, *157*, 1–8. https://doi.org/10.1016/j.rse.2014.09.021

Papenfus, M., Schaeffer, B., Pollard, A. I., & Loftin, K. (2020). Exploring the potential value of satellite remote sensing to monitor chlorophyll-a for US lakes and reservoirs. *Environmental Monitoring and Assessment*, *192*(12), 808. https://doi.org/10.1007/s10661-020-08631-5

Pawlak, J., Laamanen, M., & Andersen, J. (2009). *Eutrophication in the Baltic Sea: An Integrated Thematic Assessment of the Effects of Nutrient Enrichment in the Baltic Sea Region: Executive Summary*. https://doi.org/10.13140/RG.2.1.2758.0564

Pinto, U., Maheshwari, B. L., & Ollerton, R. L. (2013). Analysis of long-term water quality for effective river health monitoring in peri-urban landscapes—A case study of the Hawkesbury–Nepean river system in NSW, Australia. *Environmental Monitoring and Assessment*, *185*(6), 4551–4569. https://doi.org/10.1007/s10661-012-2888-2

Plateau Lake Water Quality and Eutrophication: Status and Challenges. (2023). MDPI - Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/books978-3-0365-6464-7

Ras, M., Steyer, J.-P., & Bernard, O. (2013). Temperature effect on microalgae: A crucial factor for outdoor production. *Reviews in Environmental Science and Bio/Technology*, *12*(2), 153–164. https://doi.org/10.1007/s11157-013-9310-6

Reisenhofer, E., Adami, G., & Favretto, A. (1996). Heavy metals and nutrients in coastal, surface seawaters (Gulf of Trieste, Northern Adriatic Sea): An environmental study by factor analysis. *Fresenius' Journal of Analytical Chemistry*, *354*(5), 729–734. https://doi.org/10.1007/s0021663540729

Rogers, E. M. (2003). *Diffusion of innovations* (Ekonomihögskolans bibliotek Office; 5. ed.). Free press.

Rozenberg, J., & Fay, M. (2019). *Beyond the Gap: How Countries Can Afford the Infrastructure They Need while Protecting the Planet*. World Bank Publications.

Sano, F., Akimoto, K., Wada, K., & Nagashima, M. (2013). Analysis of CCS Diffusion for CO2 Emission Reduction Considering Technology Diffusion Barriers in the Real World. *Energy Procedia*, *37*, 7582–7589. https://doi.org/10.1016/j.egypro.2013.06.702

Schindler, D. W. (1977). Evolution of Phosphorus Limitation in Lakes. *Science*, *195*(4275), 260–262.

Scholten, M. C. Th., Foekema, E. M., Van Dokkum, H. P., Kaag, N. H. B. M., & Jak, R. G. (Eds.). (2005). Eutrophication and the Ecosystem. In *Eutrophication Management and Ecotoxicology* (pp. 1–20). Springer. https://doi.org/10.1007/3-540-26671-2_1

Sen, A., & Srivastava, M. (1997). *Regression Analysis: Theory, Methods, and Applications*. Springer Science & Business Media.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., Lee, C., Kim, T., Park, M., Park, J., & Heo, T.-Y. (2020). Prediction of Chlorophyll-a Concentrations in the Nakdong River Using Machine Learning Methods. *Water*, *12*, 1822. https://doi.org/10.3390/w12061822

Shoener, B. D., Schramm, S. M., Béline, F., Bernard, O., Martínez, C., Plósz, B. G., Snowling, S., Steyer, J.-P., Valverde-Pérez, B., Wágner, D., & Guest, J. S. (2019). Microalgae and cyanobacteria modeling in water resource recovery facilities: A critical review. *Water Research X*, *2*, 100024. https://doi.org/10.1016/j.wroa.2018.100024

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., & Kouimtzis, Th. (2003). Assessment of the surface water quality in Northern Greece. *Water Research*, *37*(17), 4119–4124. https://doi.org/10.1016/S0043-1354(03)00398-1

Sin, Y., & Lee, H. (2020). Changes in hydrology, water quality, and algal blooms in a freshwater system impounded with engineered structures in a temperate monsoon river estuary. *Journal of Hydrology: Regional Studies*, *32*, 100744. https://doi.org/10.1016/j.ejrh.2020.100744

Smith, V. H. (2003). Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environmental Science and Pollution Research*, *10*(2), 126–139. https://doi.org/10.1065/espr2002.12.142

Soete, L. (1985). International diffusion of technology, industrial development and technological leapfrogging. *World Development*, *13*(3), 409–422. https://doi.org/10.1016/0305-750X(85)90138-X

Thacker, S., Adshead, D., Fay, M., Hallegatte, S., Harvey, M., Meller, H., O'Regan, N., Rozenberg, J., Watkins, G., & Hall, J. W. (2019). Infrastructure for sustainable development. *Nature Sustainability*, *2*(4), Article 4. https://doi.org/10.1038/s41893-019-0256-8

Wang, S., Li, J., Zhang, B., Spyrakos, E., Tyler, A. N., Shen, Q., Zhang, F., Kuster, T., Lehmann, M. K., Wu, Y., & Peng, D. (2018). Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index. *Remote Sensing of Environment*, *217*, 444–460. https://doi.org/10.1016/j.rse.2018.08.026

Weather and Climate. (n.d.). *Skane, SE Climate Zone, Monthly Weather Averages and Historical Data*. Skane, Sweden Climate. Retrieved April 23, 2023, from https://tcktcktck.org/sweden/skane

Wu, Q., Xia, X., Li, X., & Mou, X. (2014). Impacts of meteorological variations on urban lake water quality: A sensitivity analysis for 12 urban lakes with different trophic states. *Aquatic Sciences*, *76*(3), 339–351. https://doi.org/10.1007/s00027-014-0339-6

Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X., & Zhang, Y. (2021). An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water*, *13*(22), Article 22. https://doi.org/10.3390/w13223262

Yajima, H., & Derot, J. (2017). Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics*, *20*(1), 206–220. https://doi.org/10.2166/hydro.2017.010

Yusri, H. I. H., Ab Rahim, A. A., Hassan, S. L. M., Halim, I. S. A., & Abdullah, N. E. (2022). Water Quality Classification Using SVM And XGBoost Method. *2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC)*, 231–236. https://doi.org/10.1109/ICSGRC55096.2022.9845143

Zhang, Y., Wu, L., Ren, H., Deng, L., & Zhang, P. (2020). Retrieval of Water Quality Parameters from Hyperspectral Images Using Hybrid Bayesian Probabilistic Neural Network. *Remote Sensing*, *12*(10), Article 10. https://doi.org/10.3390/rs12101567

8. Annex

8.1. Annex 1

Table 2. Descriptive statistics of scaled environmental parameters.

	CHL	Day of a year	Latitude	Longitude	pH	Temperature	P	DO	EC	Turbidity	N
Mean	0.32	0.53	0.66	0.32	0.47	0.57	0.03	0.42	0.35	0.20	0.23
STD	0.17	0.25	0.23	0.19	0.20	0.28	0.02	0.20	0.22	0.14	0.15
Min	0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
25%	0.16	0.36	0.47	0.25	0.32	0.37	0.02	0.29	0.08	0.11	0.13
50%	0.30	0.62	0.70	0.29	0.45	0.58	0.02	0.37	0.41	0.16	0.24
75%	0.48	0.63	0.91	0.54	0.63	0.80	0.04	0.55	0.49	0.25	0.32
Max	0.32	0.53	0.66	0.32	0.47	0.57	0.03	0.42	0.35	0.20	0.23

8.2. Annex 2

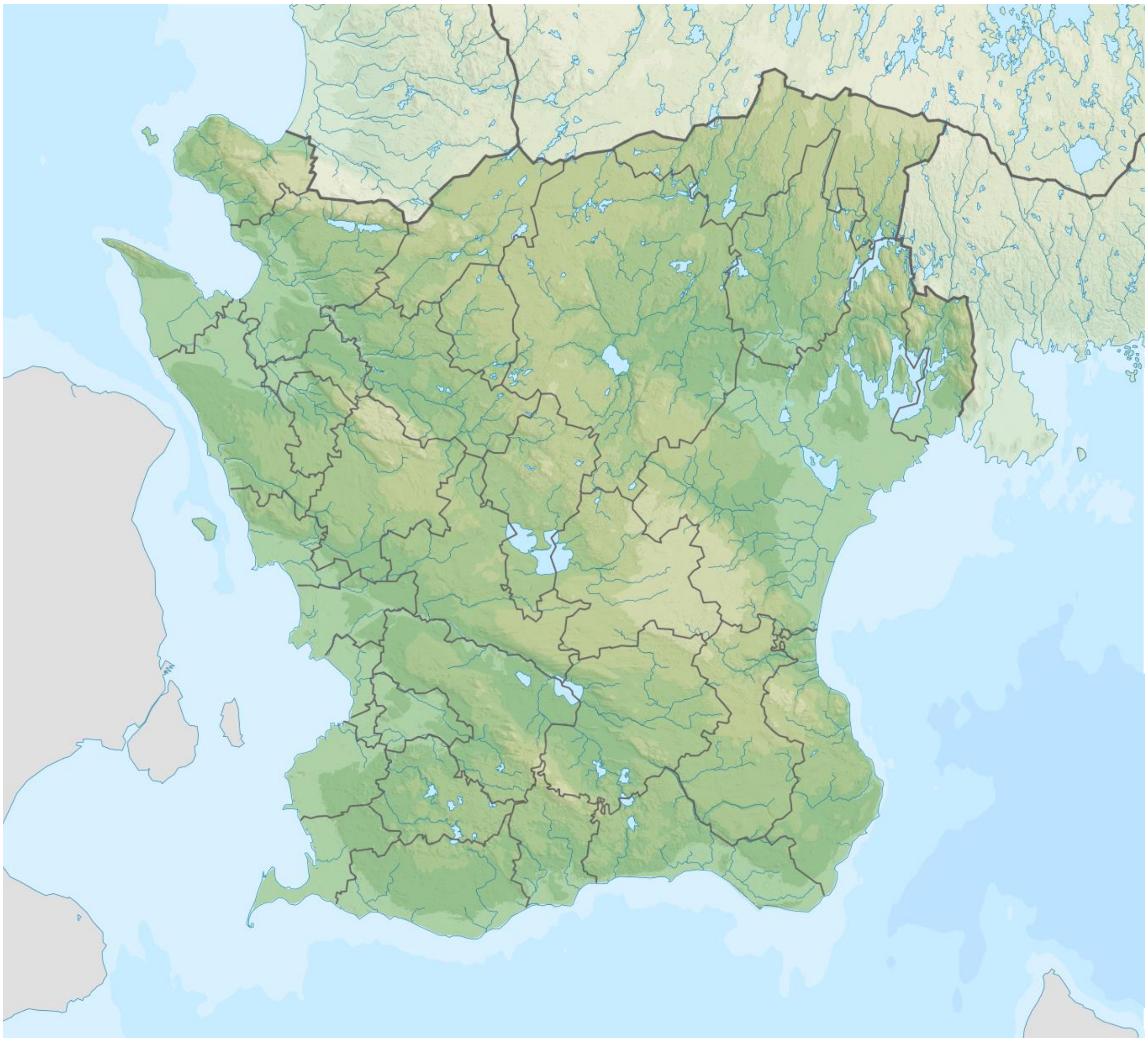

Figure 7. Maps of Skåne län.

8.3. Annex 3

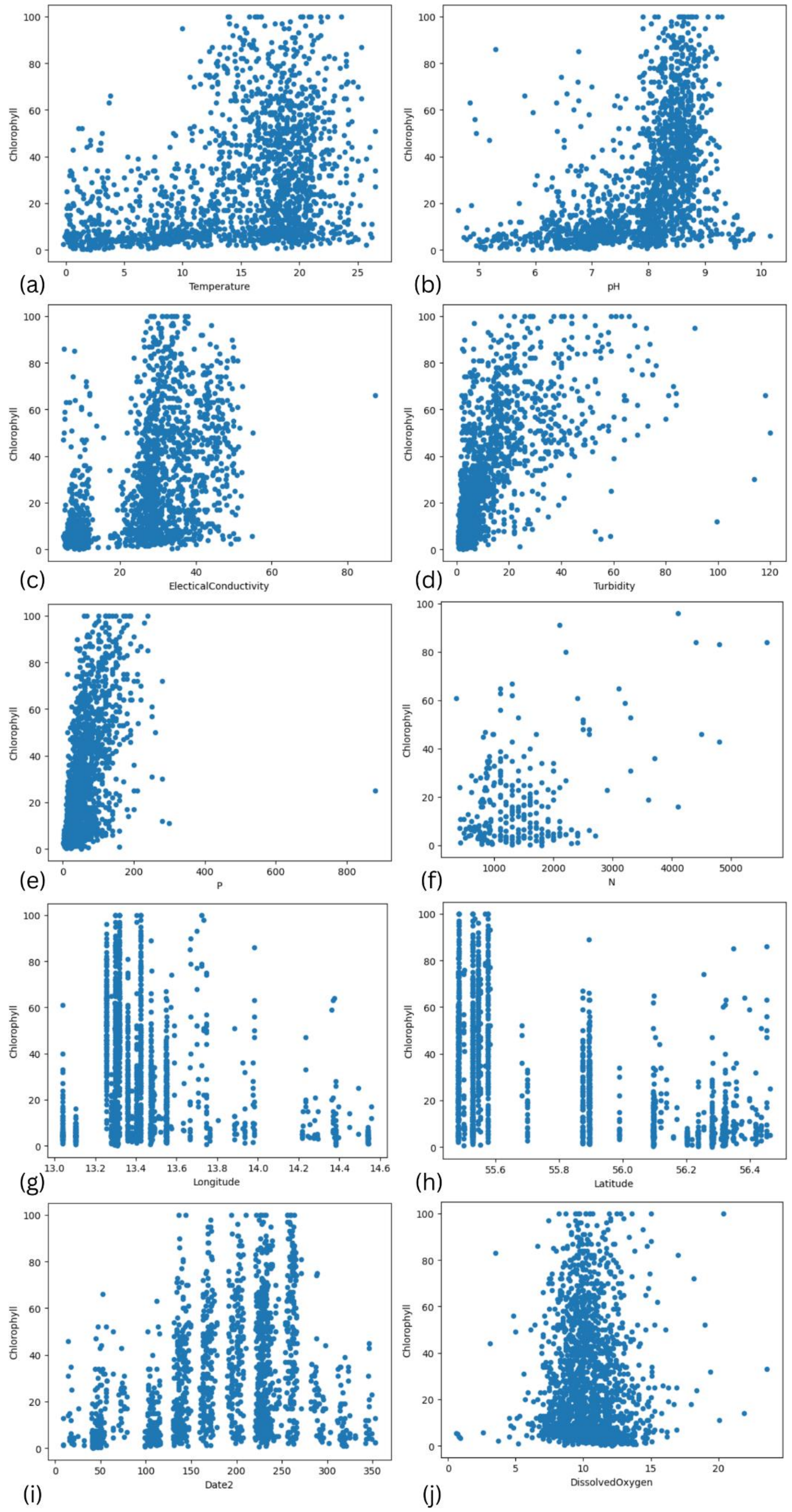

Figure 8. The scatter plot for mono-linear regression. Different parameters to CHL.
(a) T, (b) pH, (c) EC, (d) Turbidity, (e) P, (f) N, (g) Longitude, (h) Latitude, (i) day of a year, (j) DO.