ASSESSMENT OF SOURCE WATER MICROBIAL QUALITY USING BAYESIAN BELIEF NETWORKS AND DATA BALANCING ALGORITHMS

by

Atefeh Aliashrafi Zagi

B.Sc., University of Tabriz, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES

(Civil Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

May 2022

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis entitled:

ASSESSMENT OF SOURCE WATER MICROBIAL QUALITY USING BAYESIAN BELIEF NETWORKS AND DATA BALANCING ALGORITHMS

submitted by Atefeh Aliashrafi Zagi in partial fulfillment of the requirements of

the degree of Master of Applied Science.

Dr. Nicolas Peleato, School of Engineering

Supervisor

Dr. Cigdem Eskicioglu, School of Engineering

Supervisory Committee Member

Dr. Rehan Sadiq, School of Engineering

Supervisory Committee Member

Dr. Abbas Milani, School of Engineering

University Examiner

Abstract

Cryptosporidium and *E. coli* are recognized as critical pathogens in source water with mortality risk. In order to protect public health from waterborne risks, monitoring of *Cryptosporidium* and *E. coli* in drinking water sources is essential. However, direct measurement of these pathogens is expensive and labor-intensive, resulting in limited information and time-delays for risk-based management of water systems. While these challenges slow down the real-time monitoring of pathogens' levels in ambient waters, AI-based techniques offer a fast and effective alternative for direct measurements.

Bayesian Belief Networks (BBNs) is one of these data-driven methods gaining traction in modelling environmental systems and capturing their uncertainties. BBNs can assist the decision-makers by visualizing the interaction of variables in the complex systems. In this thesis, BBNs have been used to estimate *Cryptosporidium and E. coli* levels to provide a real-time assessment of the microbial quality of source water and fill the time gap required for direct measurement.

However, available *Cryptosporidium* data are rare and unbalanced, mainly indicating absence or non-detectable levels of *Cryptosporidium*. To overcome this challenge, two data balancing algorithms, Adaptive Synthesized Sampling (ADASYN) and Synthetic Over Sampling Technique (SMOTE) have been utilized. The objective was to eliminate unbalanced features of the dataset and train the model in a way that can predict both presence and absence of *Cryptosporidium* based on unbalanced and real measurements. In current work, the BBN model has been used for *Cryptosporidium* prediction and trained for the first time with a balanced dataset generated through ADASYN and SMOTE algorithms. The application of balancing algorithms increased the prediction accuracy to more than 60%, compared with models developed by unbalanced datasets.

Furthermore, the sensitivity of pathogen's level to different water quality and weather parameters was also investigated with the aim of improving the information regarding factors influencing source waters quality. Although precipitation and temperature indicated a significant impact on target parameters, the scale of the impact was very site-specific. The observation indicated that besides weather and water quality characteristics, different characteristics of each monitoring site seem to affect the level of *Cryptosporidium* and *E. coli* in studied water sources.

Lay Summary

This thesis focuses on predicting the concentration of *Cryptosporidium* and *E. coli* in drinking water sources using data-driven models. The development of data-driven models requires the historical data of understudied pathogens. However, the persistent challenge in the measurement of *Cryptosporidium* has made the recorded data of its presence in ambient waters limited. Therefore, different algorithms have been developed in this work to synthesize new data of *Cryptosporidium's* presence based on real observations and improve the performance of the predictive models. Furthermore, the integration of different water quality and weather parameters to predict the presence of *Cryptosporidium* and *E. coli* in source water was also evaluated and discussed in this thesis. Statistical analyses and the developed predictive models were used to better understand which parameters are required to be controlled or monitored to manage the risk of the pathogens.

Preface

This thesis is based on the research work completed in the School of Engineering at the University of British Columbia, Okanagan, under the supervision and guidance of Dr. Nicolas Peleato. The results and observation in Chapter 4 have been presented in the Canadian Society of Civil Engineering (CSCE): Monitoring Microbial Quality of Source Waters Using Bayesian Belief Networks, Aliashrafi, A., Peleato, N.M., (May 2021).

Table of Contents

List of Tables

List of Figures

Nomenclature

%	Percent
ADASYN	Adaptive Synthetic Algorithm
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANN	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average
BBNs	Bayesian Belief Networks
CFU	Colony forming Unit
CPT	Conditional Probability Tables
DAG	Directed Acyclic Graph
DO	Dissolved Oxygen
DBPs	Disinfection By-products
E. Coli	Escherichia
mg	Milligram
mL	Milliliter

MLP	Multi-layer Perceptron
SMOTE	Synthetic Over Sampling Technique
SVM	Support Vector Machine
TDS	Total Dissolved Solids
TN	Total Nitrogen
TP	Total Phosphorus
TSS	Total Suspended Solids

Acknowledgements

I would like to extend my gratitude towards Dr. Peleato who believed in me and endlessly supported me throughout my master's program. This work would not have been possible without his advice and encouragement. I would also like to give special thanks to my husband and my family for their continued support and encouragement.

Chapter 1: Introduction

1.1 Background

The most critical ingredient for human life is water. Source waters, including lakes, rivers, groundwaters and reservoirs, are crucial sources for life and industry (Zhan et al., 2021). The quality of water sources is affected by several factors related to human activities or natural and climatic events. Examples of anthropogenic pressure on source water can be the use of fertilizer in agricultural land and establishing nitrogen or phosphorus pollution in aquatic systems (Mainali & Chang, 2018) as well as industrial or municipal wastewater discharges. Furthermore, the population increase accentuates these negative impacts on source waters by raising the demands on land and food suppliers.

Climatic events and weather conditions noticeably can impact water sources (Mainali & Chang, 2018). The changes in water temperature, runoff and flooding or draughts in different regions have the potential for affecting all available water sources. For instance, the resulting variations of precipitation patterns due to climate change could influence the stress level on water supply and quality on a global scale (Qiu et al., 2019). Therefore, not only the scarcity of fresh water on the earth but also the high sensitivity of the available clean water supplies makes the monitoring of source waters important. In order to have a better understanding of source waters' condition as well as enhanced monitoring over the scale of the pressure on water supply, it is necessary to have an appropriate water quality assessment.

Water quality can be defined by a broad spectrum of indexes such as chemical, physical, or biological characteristics. An acceptable range of these parameters represents water quality, and exceedance from the permissible standard can result in water with low quality. Although the undesirable changes over each of these properties can impact water quality, public health risk is most influenced by pathogen levels in the water supply. Therefore, assessing the pathogen level in the water source is essential because the presence of some bacteria and parasites can cause illness through drinking water if not properly treated.

The microbial contamination of drinking water sources, as well as recreational waters, is the main route for transmitting waterborne diseases and outbreaks. Waterborne outbreaks are one of the main mortalities (more than 2.2 million deaths per year) and even more morbidity in both developed and developing countries worldwide (Gleick, 2002). Based on a recent UNICEF report, 41.7% of people globally are at least at the medium risk of using unimproved water in their households (Islam et al., 2020). Also, a study on waterborne diseases resulting from recreational

waters estimates almost 4 billion events in surface waters, resulting in 90 million illnesses and $2.2-$ 3.7 billion costs in the United States. This study reports that 8% and 65% of the economic burden was attributed to treating the resulting moderate and severe illnesses, respectively (DeFlorio-Barker et al., 2018). Therefore, it is well-known that the impact of waterborne outbreaks can be serious and costly for all communities (Ma et al., 2022). In order to prevent the spread of waterborne outbreaks and reduce this risk on public health, it is of great importance to have an appropriate evaluation of the microbial quality of drinking water sources and recreational waters.

An important pathogen that commonly drives public health risk associated with drinking water treatment and delivery is the protozoa *Cryptosporidium spp*. (Omarova et al., 2018). Outbreaks of *Cryptosporidium* can impact a large proportion of the population in a short time frame due to its persistence in aquatic environments (Swaffer et al., 2018) and high probability of infection at low doses (Lal et al., 2015). A review of waterborne protozoa outbreaks has reported 381 outbreaks between 2011 and 2016, with *Cryptosporidium* identified as the most common cause (63% of cases) (Efstratiou et al., 2017a; Ligda et al., 2020a). In Africa and South Asia, *Cryptosporidium* is recognized as the second leading cause of diarrhea and mortality in infants (Kotloff et al., 2013). Also, a well-known example of a *Cryptosporidium* outbreak occurred in Milwaukee, Wisconsin, in 1993, resulting in approximately 400,000 cases of illnesses (Rosell et al., 1994). In addition, there are numerous other reports of outbreaks caused by *Cryptosporidium* in both developed and developing countries (Mason et al., 2010; Wallis et al., 2003; Wheeler et al., 1999).

Water is one of the main transmission routes for *Cryptosporidium,* and drinking water is a significant pathway for outbreaks (SMITH, 1992). These protozoa are particularly important to consider when setting water treatment objectives since they are small enough to pass through some physical treatment barriers and are particularly resistant to disinfection by chlorine (Lechevallier & Au, 2004). While *Cryptosporidium* is often a major source of public health risk associated with the drinking water (Baldursson & Karanis, 2011a) information on source water concentrations and facility-specific removal efficiencies is usually unknown. In part, the lack of information on source water concentrations is due to significant sampling and measurement challenges. Quantification requires several steps, including concentration and manual detection, which are pretty challenging considering the typically low amount of *Cryptosporidium* in the samples (Usepa, 2005). For example, cyst counts are often reported per 100 L volume to account for low numbers. However, due to measurement and treatment challenges, it is likely that elevated cyst levels are present in both source and drinking waters in many systems, as suggested by

studies showing improvements to water treatment (like adding to regulations and intervening with facilities such as the filtration process) can reduce reported sporadic cases of cryptosporidiosis in served populations (Goh et al., 2004; Lake et al., 2007; Petterson & Ashbolt, 2016).

Although *Cryptosporidium* is among the most common reasons for waterborne outbreaks, due to the mentioned challenges, other indicator bacteria are used for reporting the microbial quality of water bodies. For instance, compared to the other pathogens such as *Giardia* or *Cryptosporidium*, it is easier to measure *E. coli.* Therefore, *E. coli* as a Fecal Indicator Bacteria (FIB) is considered an essential indicator of fecal contamination in water and is commonly used for the microbial assessment of drinking and recreation water. Also, *E. coli* is a human pathogen (Riley et al., 1983) that can survive in the gastrointestinal system and easily transmit or impact the surface/ground, recreational and stream waters (Khan et al., 2021). Epidemiological investigation in gastrointestinal illness in 1980 in the United States showed a correlation between these diseases with the level of *E. coli* in recreational water. (Edberg et al., 2000). These investigations resulted in using *E. coli* as an index of fecal contamination in the freshwater of beaches (Jang et al., 2017; USEPA., 1986). Therefore, most guidelines and regulations evaluate the microbial quality of water bodies based on *E. coli* levels. However, the permissible level of *E. coli* depends on different factors, including the sampling frequency and the kind of water under assessment. For instance, in order to exempt the filtration of drinking water, the level of *E. coli* per 100 ml should be less than 20 CFU in 90% of the weekly samples of the source water during the last six months (Drinking Water Officers' Guide, BC, 2017). Also, according to the USEPA, the level of *E. coli* shall not exceed 235 CFU per 100 ml for a single sample in beaches' freshwater (Jang et al., 2017).

Sewage discharges are one of the primary sources of transmitting pathogens such as *E. coli* into water sources that contribute to lack of microbial quality of water (Cheng et al., 2013). Storm events can intensify the sewer overflows and wash out livestock feces from soil into the water bodies. Climatic incidents seem to be one of the main parameters that impacts the microbial quality of source waters because it can indirectly affect all chemical and physical parameters of the water. There are several studies showing the significant impact of weather events on pathogens such as *E. coli* (e.g., Allende et al., 2017; Edge et al., 2021; Tolouei et al., 2019). For instance, one common factor in many historical pathogen outbreaks is extreme rainfall (Mac Kenzie et al., 1994) or other hydrometeorological events (Markó, 2005; Sylvestre et al., 2021).

Although the changes in the level of pathogens in source waters are expected to be rapid and transient, there is limited ability to assess pathogen risk in short time frames because direct

measurement of the quality and quantity parameters require at least more than 24 hours and more. As such, there is a need for increased source-specific monitoring of pathogens in short time frames in order to inform treatment operations better and reduce the associated public health risk associated with drinking water.

The application of data-driven models and statistical methods seems a promising approach to fill the time gap required for direct and day-to-day measurement of different water quality parameters. Data-driven models can be a faster substitution for the traditional solution of water quality managements because 1) they can learn the complex relationship between physiochemical or climatic parameters that are non-linear and highly complex, 2) the result and output of these models can be helpful with other risk-based assessment methods, 3) the methods offer the ability of simultaneous monitoring of water quality without the demand of real-time information about the parameters (Vasquez et al., 2000), 4) can significantly reduce the required costs of onsite measurement facilities (Al-Adhaileh & Alsaade, 2021). Therefore, the application of different data-driven models has triggered the attention of many researchers to examine these models' capability in predicting water quality and quantity.

However, one of the main challenges regarding applying data-driven models is the lack of available data since they completely rely on historical data to evaluate target parameters. In contrast to water quantity parameters widely available, the data for some of the water quality parameters are scarce due to logistical challenges and analytical complexities associated with frequent measurement. Data generation algorithms could be used to address the lack of data by synthesizing additional artificial data based on real-time and measured data. These algorithms are reported to compensate the demand for high-quality data and improve the performance of data-driven based models even for parameters with a low number of samples.

1.2 Objectives

The objective of this work was to use data-driven models for improving the risk-based management of source water quality and addressing the associated challenges of direct measurement of pathogens. The specific objectives of this thesis are listed below:

1) *Predict the presence and absence of Cryptosporidium and E. coli as two pathogens of concern using Bayesian Belief Networks (BBNs).*

Presence of *Cryptosporidium and E. coli* can pose a severe public health and real time assessment of these pathogens can prevent or reduce this associate risk. While several challenges are associated with conventional monitoring of these pathogens, data-based models can offer a faster and continues assessment compared to direct measurements and conventional methods.

Several parameters can impact the level of *E. coli* and *Cryptosporidium* in source waters and the diversity of these parameters makes assessment of these pathogens a complex system. BBNs as one of the data-based models allows for graphical representation of parameters interaction in such complex systems and have been reported to present a good performance in modeling environmental systems and capturing their associated uncertainty. Therefore, the capability of this model in estimating the presence of *E. coli* and *Cryptosporidium* has been assessed in this thesis. The reason for considering these pathogens included the persistent challenges and time delay in their direct measurement, the resistance to common disinfection processes, and the associated risk of their presence to public health. Furthermore, the presence of *E. coli* can be an indicator of fecal contamination, and this pathogen is the basis of most water quality monitoring regulations. Also, besides the prediction of pathogens, the capability of the developed models in unfolding the interaction between weather characteristics and water quality parameters was examined. In order to explore how weather parameters such as precipitation and temperature can impact the microbial quality parameter or/and how microbial quality indicators will react to the different climatic scenarios,

2) *Investigate the capability of data balancing algorithms in compensating the deficiency of Cryptosporidium observation and the lack of data in developing data-based models.*

The direct measurement of *Cryptosporidium* is a more complex, expensive, and labor-intensive process compared to the measurement of other pathogens. These persistent challenges in direct measurement of *Cryptosporidium* have resulted in lower observation and record of this parasite in source waters. In this work, the impact of data synthesizing algorithms on generating artificial data based on real data was investigated because the availability of balance data is crucial in developing data-driven models. The objective was to see if the data-balancing algorithms could improve the performance of BBN models in predicting *Cryptosporidium* or not.

The body of this thesis is structured as follows. The second chapter has included a broad literature review of the topics that have been covered in this work. The detail of methods that were used in deriving the results, including the source of employed data, BBN model, SMOTE, ADASYN and efficiency criteria of model performances, have been described in the 3rd chapter. The result of predicting and modelling *Cryptosporidium* and *E. coli* have been presented in chapters 4 and 5 respectively, followed by the analysis of variables interaction that has been reported in chapter 6. Finally, chapter 7 has summarized the conclusion of the thesis and discussed the future works that can be done in this area.

Chapter 2: Literature Review

2.1 Assessment of Microbial Quality in Source Waters

Population increases in the modern world are followed by urbanization, industrialization and agricultural activities (Poonam et al., 2013). All of these anthropogenic activities put extra pressure on limited water sources and introduced more pollution to water supplies. Therefore, it is of great importance to assess the quality of available water sources to prevent future risks to public health, especially while climate change makes the risks even more tangible.

The assessment of surface water quality started in the early twentieth century because population and industrial booming prompted people to protect available water sources. Meanwhile, the wastewater effluents were contaminating the river water, which triggered more assessment on the biological quality of water sources because sewage flows were affecting the overall quality of water bodies by introducing different microorganisms and reducing the oxygen level. Before this time, the outbreak incidents were attributed to the lack of wastewater treatment because it was believed that running water could purify itself. However, the increment of waterborne diseases such as a typhoid outbreak in Butler in Pennsylvania forced public health regulations to develop more protocols on limiting the discharge of untreated wastewater effluent to water sources. Still, the only guidelines on source water quality evaluation were limited to quantity assessment such as flood control. However, the development of filtration and chlorination offered a cheaper and more effective approach than only limiting the sewage effluent in water sources and helped in initiating more regulations over the assessment of the drinking water (Tarr, 1996).

Since the early twentieth century several quality indexes or parameters have been developed to assess water quality. One of the challenging aspects of quality assessment of water sources is the evaluation of biological or microbial quality indicators. Several natural and anthropogenic activities such as animal manure, weather events, agricultural and industrial activities can introduce new viruses, bacteria, or parasites to drinking water systems, and some of these microorganisms are resistant to disinfection of the treatment system. Therefore, it is essential to have a microbial quality assessment over the water supplies to better plan for developing treatment systems or controlling drinking water quality. The microbial quality assessment of water refers to the evaluation of pathogens that can potentially infect people and develop diseases or cause mortality. These pathogens can be categorized into three groups: bacteria, protozoans, and viruses. According to the US Environmental Protection Agency (EPA),

the important bacterial pathogens include *E. coli* Shigella and Salmonella. Also, *Cryptosporidium* and *Giardia* are the main parasites of concern. Each of the mentioned parameters has a specific permissible level in source water, and this level also can vary for intended water uses such as drinking, recreational or irrigation purposes (Jamieson et al., 2004).

2.1.1 Assessment of *E. coli* in Source Waters

Although each specific pathogen has a specific permissible level or infection rate, the difficulties in measuring each parameter have led to selecting one indicator organism. Most regulations rely on indicator organisms to estimate the concentration and persistence of other pathogens in the water body. One of these indicator organisms is *E. coli* which the USEPA recommends as the principal indicator organism in freshwaters (Jamieson et al., 2004). *E. coli* can pose a serious risk to public health, and the presence of *E. coli* can be an alert for having other fecal contaminants in the environment. Although the mere presence of this bacteria does not necessarily indicate the existence of other pathogens, it can imply the possible contamination of water with other fecal-related microorganisms such as Salmonella and hepatitis (Odonkor & Mahami, 2020).

Several studies or research with the objective of microbial assessment of water quality have focused on assessing *E. coli* levels in specific source waters: lake, river, revisors, etc. Also, some research, such as a study conducted by Odonkor and Mahami (2020), has considered several drinking water sources such as dams, rivers, streams, and underground water sources (in Dangme West District in Ghana) for assessment of *E. coli* concentration. Although the study has reported the contamination of all investigated locations to fecal bacteria, the concentration was observed to be different for each monitoring site. Samples recorded from the dams were reported to have the highest risk, and samples from the groundwater were found to have the lowest risk of contamination. The study has suggested anthropogenic activities as the main source of fecal contamination (Odonkor & Mahami, 2020).

Elevated *E. coli* or microbial contamination of source waters can impact treated drinking water quality and safety. Not all treatment systems are equipped with a filtration process, and some regulations allow for an exception for the filtration process. An investigation by Barragan et al. (2021) concluded that structural deficiencies in treatment, distribution, and storage as the reason for poor microbial quality in Villapanizon, Columbia. This study has evaluated the health risk associated with drinking water systems and used *E. coli* as a microbial quality indicator. The implemented QMRA method in the work estimated the risk of exposure to pathogenic *E. coli* for the population of the region. Different regions have been evaluated, and the resulting

observations show that the quality guidelines have not been fulfilled, and *E. coli* has been observed in samples from all testing locations (Barragán et al., 2021). Also, another study that investigates the *E. coli* and FIB contamination in all drinking water sources of Rohingya camps in Bangladesh has reported that 34.7% of households were contaminated with *E. coli* and 73.96% with fecal coliforms. The study has indicated that although the decontaminating drinking water treatment plant efficiently removed these pathogens, secondary contamination still occurs during storage or collection. The lack of personal knowledge and domestic hygienic practices were identified responsible for this contamination (Mahmud et al., 2019).

Due to the mentioned reasons, *E. coli* was the source of several serious outbreaks during history. For example, the Burdine Township, Missouri outbreak was recorded between 15 December 1989 and 20 January 1990. Four people died, and 3126,243 people were exposed to and developed diseases from *E. coli*, with 36 persons hospitalized (Swerdlow et al. 1992). The reason has been reported to be the overflow of sewage and inadequate sewage treatment. Also, a water break that occurred after the outbreak intensified the contamination and resulted in the main peak (Swerdlow et al. 1992). Similarly, the contaminated and unprotected water supply in the Highland region of Scotland (Licence et al. 2001) resulted in another outbreak in the region as the animals were permitted to graze in the region of water supply and have resulted in fecal contamination (mainly *E. coli*) to the water. There is even more report of *E. coli* contamination in both developed and developing countries, such as the Wyoming outbreak during 1998, the Swaziland outbreak during 1992 in South Africa and the *E. coli* outbreak in Grampian in Scotland during 1990, were all due to lack of microbial quality of water. However, the largest and most serious outbreak was the Walkerton outbreak occurred during May and June 2000 in Ontario (Hunter, 2003). More than thousands of people were infected, with six fatalities and 65 people hospitalized. This outbreak was also associated with the contamination of source water with cattle feces. When most of these outbreaks could have been prevented if the source water quality had been adequately monitored and the treatment or distribution systems were improved.

Followed by this outbreak, new provisions were developed for preventing similar events in the future. For instance, Walkerton Clean Water Centre (WCWC) is developed in this way to improve safe water management and more research and preparedness over water supplies (Lisnyj & Dickson-Anderson, 2018). Although regulations have been improved due to the past incidents, there is still some weakness that can pose a risk to public health. For instance, most of the provisions rely only on *E. coli* levels to evaluate the microbial quality of water. In this matter, an investigation was conducted in Brazil on the Alto Pranayama River basin to see if *E. coli* solely

can be a potential microbial quality indicator of water sources. The sampling location of this study was from a region of the basin where the populations of neotropical otters were high, and the possibility of having fecal contamination was high consequently. However, among the 26 samples, only 30% of samples were *E. coli* positive. This observation indicated that *E. coli* cannot represent other pathogens that, with a high probability, are present in samples (Oliveira et al., 2017). Also, in the study conducted by Reinoso et al. (2008), the relationship between indicator organisms such as *E. coli* and pathogenic parasites such as *Cryptosporidium* and *Giardia* was investigated. The testing samples were collected from different points/regions of wastewater treatment facilities, such as effluent during the treatment process and effluent. The observation of the study indicated that although there was a correlation between fecal bacteria (*E. coli*) and other pathogens of concern (such as *Giardia* and *Cryptosporidium*), this correlation between indicator bacteria and other pathogens was varied along with the system due to the changes in the survival rate of pathogens (Reinoso et al., 2008). Therefore, the study recommends a more specific assessment of the pathogens in order to improve the protection of public health adequately (Harwood Valerie J. et al., 2005). Also, the study by Kartim et al. (2004) found different "die-off rates" for indicator bacteria and other pathogens in constructed surface-flow wetland. This observation indicated that each microorganism can have a specific trend in aquatic systems, and surrogating other pathogens with indicator bacteria could be an optimistic approach (Karim et al., 2004). Waterborne diseases can be caused by bacteria and because of viruses and protozoa and relying only on fecal coliforms or *E. coli* as indicator organisms result in misleading information (Gordon, 2001). In part, several experiments have indicated a very poor relationship between *E. coli* or indicator organisms with *Cryptosporidium* (Payment et al., 2000; Fu et al., 2010; Isaac-Renton et al., 2005).

Therefore, microbial quality assessment depending on indicator organisms is likely to miss pathogens such as *Giardia*, *Cryptosporidium*, *Salmonella*, etc. (Pandey et al., 2014). Therefore, other pathogens are required to be monitored besides indicator bacteria to address the limitations in identifying other water contaminants (Price & Wildeboer, 2017).

2.1.2 Assessment of *Cryptosporidium* in Source Waters

Cryptosporidium is one of the important intestinal pathogens that can cause long-lasting illnesses and nutritional disorders such as diarrhea in the case of digesting contaminated food or drinking/ recreational waters. The resulting disease by *Cryptosporidium* is one of the leading causes of morbidity and mortality in the world, specifically in developing countries that have higher

populations and lower quality sanitation systems (Fletcher et al., 2012; Khalil et al., 2018). One of the main reasons for frequent waterborne outbreaks was reported to be *Cryptosporidium* (Baldursson & Karanis, 2011b; Gallas-Lindemann et al., 2013; Breternitz et al., 2020). Similarly, a review of historical protozoan outbreaks conducted by Efstratiou et al. (2017) has reported *Cryptosporidium* was responsible for 63% of worldwide waterborne parasite protozoan outbreaks. Some of these outbreaks that have occurred in British Columbia have been presented in Table 2.1. Thus, the presence of this pathogen in the water supply should be considered a severe risk to the public health (Efstratiou et al., 2017).

Table 2.1 Example of waterborne outbreaks occurred in British Columbia.

Location	**Causing Pathogen**	**Source**	**Year of Outbreak**
Cranbrook	Cryptosporidium	Watershed contaminated by Cattle	1996
Kelowna	Cryptosporidium	Wastewater Discharge	1996
Chilliwack	Cryptosporidium	Watershed contaminated by Wildlife	1998

In addition to the increasing number of waters transmitted outbreaks caused by *Cryptosporidium*, the magnitude of the resulted illnesses is underreported due to lower number of reports from underdeveloped communities. This poor reporting and deficiency of data in these regions imply an even higher number of illnesses caused by *Cryptosporidium* worldwide.

The primary reason for surface water contamination with *Cryptosporidium* are

1) Wastewater discharge and sewage overflow (Santos & Daniel, 2017; Sato et al., 2013),
2) Climatic incidents, including temperature variation or heavy precipitation (Lal et al., 2013),
3) Resistance of *Cryptosporidium* to regular disinfection processes (Carmena, 2010)

However, the difficulties in measuring *Cryptosporidium,* make it even more challenging to be monitored or studied. Therefore, the available literature on how *Cryptosporidium* can be studied as the representative of microbial quality assessment of water sources is limited. One of these studies is conducted by Kui et al. (2021) that has evaluated drinking water quality for regions with high populations in China, based on the *Cryptosporidium* and quantitative microbial risk assessment (QMRA). The study has investigated 45 samples from source water and 45 samples

from the effluent of the water treatment plant. The observation indicated that the probability of infection exceeds the permitted limit (10^{-4}) recommended by US EPA (CAO et al., 2021). Also, Breternitz et al. (2020) have evaluated the risk of *Cryptosporidium* occurrence in water sources based on the land use and coverages around watersheds. The study has examined 11 distinct municipalities of Sao Paulo in Brazil and reported more *Cryptosporidium* occurrence in watersheds covered with higher populations and the lands with livestock application. In contrast, the watersheds close to the urban areas without any livestock activities have shown the lowest frequency and concentration of *Cryptosporidium* presence (Breternitz et al., 2020). In another study, Xiao et al. (2021) have investigated the relationship between the occurrence of the parasite and discharge of wastewater treatment plant effluent into surface waters. A QMRA to analyze was conducted to evaluate the associated risk of *Cryptosporidium* presence in the drinking water supply. The study's observation indicated that due to the high concentration of *Cryptosporidium* in the effluent, the associated risk of discharging sewage effluent is higher than the threshold of 10^{-6} that is released by WHO (Hu et al., 2018).

Despite the reported wide range of contamination sources, a common element to previous studies focusing on the occurrence of *Cryptosporidium* in the environment is the high level of uncertainty and consistent influence of various factors between watersheds. Also, the observation of the literature indicates that the possibility of *Cryptosporidium* in water suppliers is quite high, and this pathogen can cause serious risk for public health in the case of intaking to water treatment systems. Therefore, it is of critical importance to monitor and evaluate the concentration and presence of *Cryptosporidium* in source waters.

2.2 Impact of Weather events on Microbial Quality of Source Waters

One of the main challenges regarding the assessment of source water quality is their sensitivity to various factors, including climatic changes. The studies on how weather or climate events can impact water quality are quite a few because this impact depends on different factors such as the nature of water sources (if it is considered groundwater, river or reservoirs, etc.), the geographical condition and the “hydro-climatic” condition of the water (Goderniaux et al., 2011; Macdonald et al., 2009).

Understanding of how natural water quality will respond to the climatic changes is in very early stages (Howard et al., 2010; Levy et al., 2009a; Sadik et al., 2017; Taylor et al., 2009; Wu et al., 2009). Although there is not a good understanding of the relationship between weather conditions and water quality, historical outbreaks caused by heavy rainfalls or temperature

flocculation indicate the existence of a dependency. For instance, one of the early *E. coli* outbreaks during 1992 was due to extreme rainfall after a drought period. The heavy precipitation washed out fertilizers and animal feces into surface waters in South Africa and Swaziland (Isaacson et al. 1993). Similarly, the heavy rainfall in Walkerton during 2000 in Canada resulted in a serious outbreak in which more than 2000 people were infected with *E. coli* (Hrudey et al., 2002). The outbreak happened following heavy spring precipitation (143 mm in 5 days). Rainfall with this magnitude was estimated to occur 1 in 60 years (O'Connor, 2002). The source of *E. coli* contamination was the cattle manure or agricultural lands that were washed out with the rainfall flow and resulted in the failure of the treatment system (Hrudey et al., 2002).

The occurrence of such unexpected precipitation and rainfalls of this kind can become increasing common due to climate change. These non-regular droughts or precipitation events are potential of contaminating the waters sources. *E. coli* outbreaks in Grampian, Scotland, are an example of how low precipitation and high temperature during summer can eliminate a drinking water supply of the community (Dev et al. 1991). For instance, a study conducted by Jacklin in 2015 investigated the factors that are involved in the microbial quality of groundwaters. The statistical analysis has indicated that *E. coli* levels were highest during the wet seasons, likely due to concurrent high temperatures that increased the growth rate and survival of *E. coli* in the water (Jacklin, 2016). A similar observation was reported by Abia et al. (2015) in Australia that has reported a higher level of *E. coli* during the season with higher precipitation because of the direct relationship of *E. coli* with the air temperature (Abia et al., 2015). However, Balleste et al. (2019) discussed that the influence of environmental parameters such as rainfall depends on the source of natural water and pollution. For instance, the investigation in their study has indicated that higher precipitation in "human polluted" areas can dilute the fecal contamination. At the same time, in the case of the agricultural regions, runoff can decrease the quality in terms of fecal pollution (Ballesté et al., 2019).

A similar observation has also been observed in a study conducted by Levy et al. (2009), which examined the impact of seasonal changes on the variation of microbial quality of water and *E. coli* levels. Like other studies, it was also observed that the impact of rainfall on *E. coli* is stronger during the wet season than the dry season. As such, the study suggested that a 1 cm increase of rainfall was associated with a 3% decrease in *E. coli* counts in source waters during the dry season and a 6% reduction during the wet season. According to the finding of the study, soil moisture can be a reason for this variation as it can impact the microbial processes during rainfall (Levy et al., 2009b). Also, this study has explored the different time scales (hourly, daily,

and weekly scales) to measure *E. coli* and reported more variability and quality flocculation based on an hourly scale that is again explainable due to the flushing feature of rainfalls. Therefore, observation of this investigation indicates that the impact of weather events on microbial quality of water can vary based on runoff effect, dilution effect and the time scales consideration (Levy et al., 2009b).

Not only are *E. coli levels* or fecal contaminations affected by climatic features, but other contaminants and pathogens such as *Cryptosporidium* levels are sensitive to weather variations. For instance, a study conducted by Atherholt et al. (1998) has examined the impact of seasonal changes and rainfall on the concentration of Giardia and *Cryptosporidium*. Delaware river has been selected as the sampling location, and the recorded data was representative of four seasons. The highest concentration of *Cryptosporidium* was observed in winter and fall within the samples that have the highest flow rate and turbidity. This observation shows a direct relationship between precipitation and parasite level in the tested area. Therefore, the study concluded the elevated runoff and sediment resuspension (as the reflection of seasonal changes) as the reason for *Cryptosporidium* increase *(Atherholt et al., 1998)*. In addition, a study by Craun et al. (1998) reviews the historical *Cryptosporidium* outbreaks in the United States, United Kingdom and Canada. Investigating the several outbreaks in studied countries, the study has reported sewage discharge and runoff from agricultural land as the primary contamination source of *Cryptosporidium*. Also, this review has indicated that heavy rainfalls by worsening these contamination sources can lead to lifted parasite concentration in source waters and impaired treatment systems. Therefore, most of the studied outbreaks, such as Carrollton in 1987; Sheffeld in 1986; North Humberside in 1990; Isle of Thanet in 1991; Bradford in 1992, have occurred during heavy rainfalls (Craun et al., 1998).

Other climatic changes such as frequency and duration of droughts and high temperature can impact parasite levels in the source water. For instance, Leland et al. (1993) investigated one of the *Cryptosporidium* outbreaks that has happened in Talent in 1992. The study has reported the drought condition and associated low flow as one of the possible reasons that can result in less diluted sewage discharge in the Bear Creek water supply and higher parasite concentration (Leland et al., 1993). Also, Masina et al. (2019) have studied the level of *Cryptosporidium* and *Giardia* with environmental conditions such as precipitation and air/water temperature in Iqaluit. The study reported an increase in number of protozoa because of the increased air and water temperature in the Arctic (sampling region). However, despite other studies, this investigation has not observed any association between *Cryptosporidium* and precipitation. The main reason could

be the difference in wildlife and the lower human population in Northern Canada. Also, because there are very few sources of *Cryptosporidium* in Iqaluit and rainfall cannot introduce carry over or increase its level consequently.

As the result of studies indicates, different case studies with varying sampling sites can change the impact of weather conditions on microbial contamination of source water. These observations from literature studies imply sophistication of the relationship between climatic events and *E. coli*/*Cryptosporidium* level. Therefore, there is more demand on studying this correlation because now the outcomes of global change such as heavy rainfalls, non-regular flooding, precipitation, and high temperatures are getting more tangible.

2.3 Modeling Microbial Quality of Water Sources

Water quality modelling can be a helpful approach to estimate the associated impact of climate change or the risk of new pollutants in the source waters (Feng et al., 2013). Water quality modelling simulates the interaction between different water quality characteristics and can provide a quantitative basis to better understand and interpret their relationship (Reichert et al., 2001). However, generating accurate knowledge about the water environment can be quite challenging because there are several complex processes involved, such as eutrophication, chemical and biological pollutant transport and diffusion, algal blooms, weather events, etc. While these scenarios can be simulated and predicted by the water quality modelling (Kim et al., 2021). Two approaches can be employed in modelling water quality. One classical modelling method uses physical models that can simulate the mechanisms of underlying processes in the monitoring environment. Physically based models are based on detailed knowledge and information of the system, which should be achieved by exact measurement and analysis. Therefore, these methods require extensive cost and time to be developed. However, data-based models offer a faster and more cost-efficient approach for modelling environmental systems such as source water quality. Furthermore, the fast growth of computer and monitoring techniques such as the application of wireless sensors are improving the availability of the data (Aliashrafi et al., 2021). Besides, the successful report of applying data-driven models in other fields of science and engineering (Gilpin et al., 2018; Montáns et al., 2019), has resulted in the development of more data-based models for assessment of water quality. These types of models can generate timely and accurate information of the system, which can be of great help for decision-makers and regulation developments (Telci et al., 2009).

2.3.1 Application of Machine Learning Methods in Assessment of Source Water Quality

The rapid improvements to computational power and advances in developing new machine learning algorithms have resulted in numerous examples of data-driven or machine learning models (Bishop, 2006). Due to the ability of these models in tackling the time and cost challenges, several data-driven models have been implemented in predicting water quality.

There are a number of studies with the application of data-driven models in predicting source and drinking water quality. Among them, a review study conducted by Tyralis et al. (2019) claims artificial neural networks (ANNs) as the most applied method in forecasting water quality (Tyralis et al., 2019). Although predicting quality indicators is the same objective in most of these studies, the case study, the selected target parameter and the input variables are different for each investigation (Al-Adhaileh & Alsaade, 2021; Gazzaz et al., 2012; Kim et al., 2021; Sarkar et al., 2021). For instance, Palani et al. (2008) have used the ANN algorithm to predict the seawater quality parameters (DO), salinity and temperature in Singapore coastal waters (Palani et al., 2008). Kalin et al. have applied the same method for predicting the total dissolved solids (TDS) and total suspended solids (TSS) as the water quality indicators in Western Georgia (Kalin et al., 2010). Similarly, Seo et al. (2016) have successfully applied the ANN algorithm on Cheongpyeong dam in Korea and were able to predict a wide range of quality parameters (temperature, DO, pH, Electric Conductivity, TN, TP, Turbidity and Chlorophyll-a) (Seo et al., 2016). Besides quality indicators, ANNs have been widely applied to predict disinfection by-products (DBPs) in the drinking water systems (Hong et al., 2020; Peleato et al., 2018; Xu & Liu, 2013). In addition to ANNs, linear or non-linear regression methods (Chen & Liu, 2015; Harvey et al., 2009; Yang et al., 2017), Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Deng et al., 2015; Solgi et al., 2017) and Support Vector Machines (SVM) (King et al., 2000; S. Liu et al., 2013; Xiang & Jiang, 2009) are examples of classical approaches that are commonly applied to predict water quality indicators.

In terms of microbial quality, the real-time estimation capability of data-driven models can compensate for the time delay of measuring pathogen concentration in source waters because the measurement of present microorganisms in water can take even more than 24 hours, and during this time gap, there is a possibility for pathogen levels to exceed the permitted standard. Also, rapid changes in microbial activities demand a real-time assessment of their concentration. Therefore, to address the challenges of time delay in monitoring the quality indicators, mathematical and data data-driven models based on historical data have been used to predict

pathogen concentrations on a day-to-day basis (Benham, 2006; Coffey et al., 2010; Nevers & Whitman, 2011; Nicholas et al., 2016).

Due to the large variety of data-based techniques and availability of wide datasets, the selection of the modelling techniques is important and challenging. Therefore, besides utilizing the predictive models, the focus of research in literature was on comparing the predictive powers and performance of different algorithms. For example, Brooks et al. (2016) have compared 14 distinct regression techniques for predicting *E. coli* concentration in recreational water of Wisconsin. The study denoted the superiority of multi-linear regression algorithms (MLR) over other methods (Brooks et al., 2016). Similarly, Malzer et al. (2016) have compared ANN, logistic and process-based modelling performance in predicting *E. coli* of Ruhr River in Germany and reported the higher performance of the ANN method for all sampling sites over the river (Mälzer et al., 2016). Tousi et al. (2021) have developed SVM, logistic regression and ridge classifier methods to predict *E. coli* level in agricultural water and reports the outperform of SVM algorithms for this purpose. Another recent study conducted by Sokolova et al. (2022) has compared different data-based predictive models' performance in predicting *E. coli* and evaluating the impact of each predictor variable on the pathogen level. The study has utilized the Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA), Regression and RF models and denoted the superiority of the RF model compared to the other techniques (Sokolova et al., 2022a). Some example of availble litreture that have employed machine learning algorithms for predicting water quality has been summorized in Table 2.2.

In addition, another successful application of the RF model has been reported by Mohammed et al. (2017), which has utilized these algorithms for predicting FIB, *E. coli* and Intestinal enterococci in the Svartdiket water treatment plant in Norway. Although the study has used a wide range of physio-chemical characteristics of water (pH, colour, turbidity, conductivity), the colour of water and the season of the year were still the most significant variables in predicting these pathogens (Mohammed et al., 2017). Evaluating the impact of predictor variables on target parameters has also received attention. One of the main advantages of data-driven methods is their capability to capture the complex relationship between parameters of the dynamic systems that are hard to learn by conventional approaches. Therefore, the application of data-driven methods not only provides a real-time prediction of microbial contaminants but also can be helpful in investigating the underlying relationship of water quality parameters or the impact of different environmental systems such as weather events on pathogen levels. For instance, in order to assess the effect of precipitation on *E. coli* Clostridium and coliforms' levels in source water,

Tornevi et al. (2014) have utilized the time series, regression model. The study has denoted that the heavy rainfalls (>15mm/day) are associated with a higher concentration of *E. coli* and FIB and can capture a strong correlation between precipitation over two days with microbial contamination of the water (Tornevi et al., 2014).

Table 2.2 Summary of literature examples that have used the machine learning algorithms for predicting water quality.

Water System	**Applied Model**	**Predicted Water Quality Indicator**	**Reference**
Marine Water	ANN	DO	Palani et al. (2008)
River Water	ANN	TDS	Kalin et al. (2010)
Tap Water	RBFS-ANN	DBPs	Xu et al. (2013)
River Water	ANN-Logistic	*E. coli*	Malzer et al. (2016)
Recreational water	None-Linear Regression	*E. coli*	Brooks et al. (2016)
Reservoir Water	ANN	Do, pH, TN, TP	Seo et al. (2016)
River Water	ANFIS - SVR	BOD	Solgi et al. (2017)
Drinking Water Treatment Effluent	RF	*E. coli* - Intestinal enterococci	Mohammed et al. (2017)
Drinking Water Treatment Effluent	ANN	DBPs	Hon et al. (2020)
Agricultural ponds	SVM-logistic	*E. coli*	Tousi et al. (2021)
Groundwater	ANN	*E. coli*	Khan et el. (2021)
Groundwater	ANN	*E. coli*	Khan et el. (2021)
Marine water	ANN-SVR	Turbidity	Kumar et al. (2022)
River Water	RF- ARIMA	*E. coli*	Sokolova et al. (2022)

Furthermore, by application of RF methods, Sokolova et al. (2022) were able to show that the water temperature, microbial concentration upstream of testing point and precipitation have the highest importance in predicting *E. coli* (Sokolova et al., 2022b). Khan et el.,(2021) have developed a superposition-based learning algorithm (SLA) and ANN model for the same purpose. The study's observation reports the high importance of pH and the low impact of DO in predicting *E. coli* concentration of groundwater. The result of the study has indicated that the ANN model developed for predicting *E. coli* achieved the best performance by having pH, Turbidity, TDS and Electrical conductivity (Khan et al., 2021). Similarly, Rossi et al. (2020) reported successful use of multivariate regression to predict *E. coli* levels above and below a threshold level and identified the influence of relevant factors such as pH, temperature, and turbidity variations in *E. coli* levels (Rossi et al., 2020).

The available literature review indicates that the impact of the predictor can vary based on the monitoring source waters or the availability of parameters. For instance, Tousi et al. (2021) have evaluated the impact of incorporating sediment characteristics in improving the prediction power of SVM, logistic regression and ridge classifier methods. The study has revealed that the inclusion of sediment information can result in a more accurate prediction of *E. coli* in agricultural waters (Tousi et al., 2021). However, monitoring the estuarine waters, Gonzalez et al. (2012) have found the weather and water quality data, including precipitation over five days, DO and salinity, the most impactful parameters for predicting *E. coli* and enterococci (Gonzalez et al., 2012). Also, Fancy et al. (2013) reported the impact of streamflow and air temperature on *E. coli* levels in freshwater beaches (Francy et al., 2019). Therefore, it can be observed that the nature of the monitoring source water and environmental condition can play a major role in the determination of variables that can explain the concentration of pathogens (Francy et al., 2020).

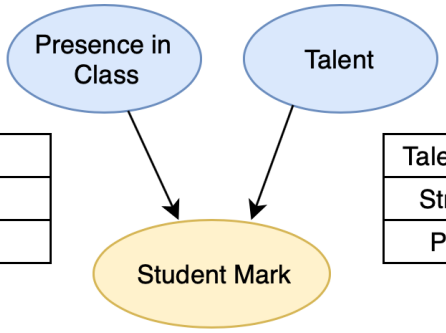
Water quality regulations mainly consider *E. coli* and Fecal Indicator Bacteria or Organisms (FIB/FIO) as indicators for overall microbial quality due to their prevalence and ease of enumeration. As such, the majority of previous studies have focused on predicting these indicator organisms (Dorner et al., 2004; Francy et al., 2013; Nevers & Whitman, 2011). However, a study by Lalancetter et al. (2014), which has studied the ratio of *E. coli* to Cryptorpsodium for different water sources, has reported that *E. coli* was a poor microbial quality indicator for drinking water intakes (Lalancette et al., 2014b). Not only for drinking water but also an investigation on source waters has indicated the poor correlation of *E. coli* as an indicator bacteria with *Cryptosporidium* in surface waters (Payment & Locas, 2011; Wohlsen et al., 2006a). However, there are limited studies that focus on the prediction of *Cryptosporidium*. These could be due to

the lack of available data representing Cryptosprodium presence and difficulty in measuring the parasite in research labs. Most of the attempts in the field were to elucidate factors driving the occurrence of *Cryptosporidium* in water bodies or modelling this pathogen with the Soil and Water Assessment Tool (SWAT) (Bergion et al., 2017; Coffey et al., 2010; Tang et al., 2011) or Quantitative Microbial Risk Assessment (QMRA) Models. For instance, Liu et al. (2018) have studied the fate and transport dynamics of *Cryptosporidium* in the Daning River using SWAT and reported *Cryptosporidium* variation among different seasons, without specific relationships to co-occurrence with heavy rainfalls (Liu et al., 2018). The study emphasized a combined impact of rainfall and regional fertilization on the level of the *Cryptosporidium*. Coffey et al. (2010) utilized a similar methodology and found that fertilization usage significantly impacts the *Cryptosporidium* level in a watershed in Ireland. Xiao et al. (2018) developed a QMRA model of *Cryptosporidium* and reported its strong relationship with the flooding frequency (Xiao et al., 2018). Hunter et al. (2011) incorporated the QMRA method with Bayesian Belief Network (BBN) to improve the risk assessment of *Cryptosporidium* and Giardia in private water supplies in England and France. The employed BBN model in the study was reported to be able to prioritize the involved factors in health risks associated with *Cryptosporidium* and Giardia (Hunter et al., 2010). While the efficiency of BBNs' integration with QMRA models in the assessment of relationships among microbial and physiochemical variables have been investigated in literature (e.g., Beaudequin et al., 2016a; Goulding et al., 2012; Gronewold et al., 2011; P. Hunter et al., 2010; Staley et al., 2012), the studies with direct application of this method for predicting the level of pathogens in source waters is quite a few.

2.3.1.1 Application of BBN Methods in Assessment of Source Water Quality

A promising approach that is well suited to modelling systems with high levels of uncertainty is the Bayesian Belief Networks (BBNs) (Bertone et al., 2016a; Herrig et al., 2019; Uusitalo, 2007a). BBNs make probabilistic graphical models that can explicitly define a dependency between variables and represent the probability of a given observation based on connected variable states (Aguilera et al., 2011; Fenton & Neil, 2012). BBNs construct these probabilities that are called conditional probabilities based on Bayes Theorem. Initiated by Reverend Thomas Bayes, this theorem is a rule for calculating the conditional probability of an outcome without knowing the joint probabilities (Grover, 2013). Therefore, Bayes Theorem allows for inferring the probability of causes based on the information of dependent effects and updating the probabilities based on new observations (Pressini, 2018). For instance, Figure 2.1 indicates an example of a simplified

BBNs and shows how the mark of a random student can be defined by other random variables: presence in the class and the talent of the student in that subject.



Presence in Class (C)	Probability
Yes	P(C=Y)
No	P(C=N)

Talent (T)	Probability
Strong	P(T=S)
Poor	P(T=P)

Talent	Strong		Poor	
Presence in Class	Yes	No	Yes	No
Pass	P(Pass | T= S, C =Y)	P(Pass | T= S, C =N)	P(Pass | T= P, C =Y)	P(Pass | T= P, C =N)
Fail	P(Fail | T= S, C =Y)	P(Fail | T= S, C =N)	P(Fail | T= P, C =Y)	P(Fail | T= P, C =N)

Figure 2.1 Example of BBNs and the Simplified Probability Tables. (Performance of a Student in an Exam).

BBNs are also called directed acyclic graphs due to the direct arcs pointing out from the parent node toward the child nodes without any loops (Hassall et al., 2019). Set of probabilities for each child nodes based on the occurrence of different states of parent nodes formes the CPTs and by the increase of node numbers and their states the resulted CPTs can be larger consequently. Considering the mentioned example, the probability of passing or failing an exam can be different based on a distinct state's combination of the parent nodes (Hassall et al., 2019). Thus, BBNs allow for considering different states for predicting an outcome of interests and provide an insight over the sensitivity of variables' relationship in a complex system.

Due to the advantages of BBNs, this method has gained traction in different fields with diverse objectives, such as scoping and intuitive presentation of relationships in the natural systems (McCann et al., 2006; Ban et al., 2014; Andriyas & McKee, 2015; Avilés et al., 2016; Forio et al., 2021; Frizzle et al., 2022;) and risk assessments in environmental structures (Avila et al., 2018a; Baldock et al., 2019; Laurila-Pant et al., 2019). Also, BBNs have efficiently been employed for monitoring and assessment of drinking and source waters quality and quantity. Such as classifying the level of lead in the drinking water systems (Fasaee et al., 2021b), quality assessment of groundwaters (Aguilera et al., 2013), evaluating the impacting factors on chlorophyll dynamics in the river water (Alameddine et al., 2011), determination of required

sanitation for water suppliers (Dondeynaz et al., 2013), evaluating the formation of DBPs in drinking water systems (Li et al., 2021) and predicting water quality indicators such as total nitrogen and phosphorus (Yu & Zhang, 2021).

Despite being popular and a widely applied method in different water quality aspects, BBNs have seen a limited application of this model has been reported for predicting pathogens in source waters. Donald et al. (2009), utilized this approach as a supplementary analysis to determine the most critical variables in increasing the health risk of waterborne pathogens (Donald et al., 2009). Since then, BBNs have been used to integrate the QMRA method for evaluating the microbial risk of water sources (Beaudequin et al., 2016b; P. Hunter et al., 2010). A few studies have employed this method to predict *E. coli* and FIB. Avila et al. (2018) have compared different data-based models for predicting the *E. coli* counts for recreation water across bathing sites of Southland, New Zealand. This investigation reported the superiority of BBN models compared to none-linear regression, logistic regression, regression tree and RF models. Besides high accuracy in predicting *E. coli*, this study indicated the efficiency of BBN models in handling the missing data (Avila et al., 2018b). Similarly, Herrig et al. (2019) used Bayesian linear regression models to predict *E. coli* and FIO, highlighting the importance of considering relationships of weather parameters such as rainfall, temperature and solar radiance on *E. coli* concentrations (Herrig et al., 2019). Improved performance of BBNs in predicting *E. coli*/FIB by integrating weather characteristics has also been reported in a study by Pandihapu et al. (2020). This study has compared the BBNs' performance with conventional models such as RF, logistic regression and Naïve Bayes. The results denoted that BBNs not only successfully predicted *E. coli*/FIB for all seven monitoring sites but demonstrateed a meaningful prediction based on an incomplete data (Panidhapu et al., 2020).

Despite the established efficiency of the BBN approach in modelling complex systems, the implementation of this model for studying *Cryptorpsdirum* in source waters is limited to the study by Bertone et al. (216). This study focused on a risk assessment based on stakeholders' inputs and developed a BBN for evaluating the impact of different weather-related parameters and factors on levels of *Cryptosporidium*, colour, and turbidity (Bertone et al., 2016).

2.3.1.2 Structure Learning of BBNs Method

The connection between variables and CPT of BBNs is constructed based on the initiated structure of the models. Therefore, defining the optimal topology of the network prior to training the model is important but challenging. Several algorithms have been used and proposed for this

purpose from the early stage of BBNs development during the 1980s. Earlier, the algorithms fell into two groups of score-based and constraint-based algorithms (Beretta et al., 2018). More recently hybrid algorithms that combine score-based, and constraint-based algorithms have been created (Guo & Li, 2022). Consequently, there are several methods for learning BBN structure, and the number of these methods is increasing with the augmented number of BBN applications.

The high number of possible structures based on potential connections has made learning the optimal structure an NP-hard problem. Furthermore, the increasing number of variables can make these possible connection even more complicated. Therefore, in order to find the optimal answers, most of the studies in the literature have proposed optimization approaches such as genetic algorithms to find the best structure. For instance, Hesar (2013) introduced a simulated annealing algorithm that is based on an evolutionary-based method (Hesar, 2013), and Yan and Cercone (2010) have proposed E-Algorithm for developing BBN models (Yan & Cercone, 2010). In addition, some researchers have coupled statistical methods such as PCA to improve the efficiency of genetic algorithms in learning the best structure (Rezaei Tabar et al., 2016).

Despite the high number of proposed approaches, some studies in the literature have indicated that none of the available algorithms can be universally the best approach for all datasets. For instance, Mittal and Maskara (2011) have compared six different structures learning methods commonly used in literature: Greedy Search; Greedy Equivalent Search; Bayesian Network Power Constructor; PC Algorithm; Minimum Weighted Spanning Tree and K2 algorithms. The study has applied these algorithms on two different datasets and has observed that none of the approaches is necessarily efficient for all datasets (Mittal & Maskara, 2011a). This finding was also aligned with the conflicting observations in other studies. As such, the study conducted by Scutari et al. (2019) has compared several approaches, including PC algorithm, Greedy Search, K2 and Max-Min hill Climbing algorithms, and reported PC algorithm as one of the fastest learning algorithms.

In contrast, a similar study conducted by Tsamardinose et al.(2006) has denoted the lower speed of the PC algorithm compared to other methods such a tabu search Max-Min hill Climbing algorithms (Tsamardinos et al., 2006). These observations imply that no unique generalized approach for developing the topology of the BBNs can be the best method, and the effectiveness of the structure learning algorithm can be specific to each case study. Parameters such as the number of considered variables or the structure of data (if the data continues or discredited) can impact the efficiency of the learning algorithms (Beretta et al., 2018). Sensitivity of the learning structures to different elements can be why most of the developed BBN structures with the

objective of water quality modelling in literature mainly rely on expert Knowledge instead of using structure learning algorithms (e.g., Aguilera et al., 2013; Phan et al., 2019; Forio et al., 2021). Since the number of involved factors affecting the quality of water sources is very diverse, this complexity between variables' interconnection is likely to make learning the optimal structure more complicated.

2.3.1.3 Application of Data Balancing Methods in Assessment of Source Water Quality

A persistent challenge in modelling and monitoring the water quality with data-driven models is the low number of samples. This challenge is particularly salient for *Cryptosporidium* monitoring since the measurement is expensive and labour-intensive. Furthermore, the proportion of non-detects or zero values in *Cryptosporidium* monitoring datasets is high. Unbalanced datasets will significantly reduce the ability of data-driven models to learn variable relationships and limit the ability to predict positive cases when protozoa will be present.

In order to address this issue, balancing algorithms have been used in different fields to overcome the drawback of data-based algorithms in learning the variety of samples with minority classes. Although a wide range of balancing algorithms has developed (Batista et al., 2004; Han et al., 2005a), the SMOTE algorithm is the most commonly used approach for this purpose. SMOTE was developed by Chawla et al. (2002) and since then has been utilized as the base algorithm for generating other data balancing methods (Luengo et al., 2011a). One of these algorithms which are gaining more attention in the practical research field is the ADASYN algorithm. The generated data by this algorithm can reduce the introduced bias by class impabalan and can shift the capability of the models to better learn the samples that are hard to be learned (He et al., 2008).

Although several studies have reported improved prediction accuracy by balancing datasets with various balancing algorithms (Gosain & Sardana, 2017; Yang Bai, 2008; Han et al., 2005b; Luengo et al., 2011b), there are rare examples of utilizing these methods in improving the modelling of environmental or natural systems. As such, Kim et al. (2021) have used Adaptive Synthetic Sampling Algorithm (ADASYN) for improving the performance of two ANN and SVM models in predicting alga Bloom. The study reported that employing a data balancing algorithm could lift the accuracy of the prediction model by more than 33.7% (Kim et al., 2021). Also, another successful application of the ADASYN method has been reported by Xu et.,(2020). The study has applied the generated data by ADASYN in different AI_based models, including k-mean nearest neighbour, boosting decision tree, SVM and multi-layer perceptron (MLP) for five various bathing

sites in Auckland of Newzealand. Aside from the applied model, the employment of ADASYN is denoted to increase the accuracy to more than 90% (Xu et al., 2020). Due to the reported efficiency of data balancing algorithms and the aforementioned challenges in handling the *Cryptosporidium* data, the SMOTE and ADASYN method has been employed in this study to generate new synthetics samples of *Cryptosporidium* presence. Therefore, the current work aims to predict the presence or absence of *Cryptosporidium* for Kensico Reservoir, which supplies drinking water in New York City. A novel application of two widely applied balancing algorithms, SMOTE and ADASYN, are investigated to address issues with collecting *Cryptosporidium* samples and resulting unbalanced datasets. Model have been trained using synthetic samples are examined with BBNs. The performance of this method is assessed on a test set of data comprising only real samples.

2.4 Research Gaps

2.4.1 Previous Work

BBNs have established a successful performance in capturing the associated uncertainties in the environmental systems since their initiation during the 1980s (Newton, 2009). BBNs' capability in visualizing the systems with a high degree of uncertainty is considered an asset for decision-makers. As their probabilistic graphical features improve the transparency of the model and assist all involved parties to have an enhanced understanding of the system and consider possible affecting factors (Laurila-Pant et al., 2019). Therefore, the method has been widely applied in modelling water quality and quantity with various objectives such as predicting the quality indicators (Aguilera et al., 2013) and identifying the relationship between the water characteristics (Daniel et al., 2020). The method was previously used in some research as a predictive model and forecasting the pathogens in source and drinking waters. However, *Cryptosporidium*, one of the critical pathogens that are associated with several serious outbreaks worldwide, has not been modelled by this method. Predicting *Cryptosporidium* by rigorous and probabilisitic methods such as BBNs could be an effective solution to the difficulties in its direct measurement of this pathogen.

2.4.2 Research Challenges

The challenges in the measurement of pathogens have led the water quality regulations to rely only on a limited number of indicator bacteria for assessing the microbial quality of water. While some research has reported the poor connection between the presence of pathogens such as *Giardia* and *Cryptosporidium* to the existence of indicator bacteria (Lalancette et al., 2014).

Therefore, several studies have aimed to explore this relationship with different methods and case studies. However, the diversity of answers indicates the dependency of this relationship to different features of the case studies.

The application of data-driven models such as BBNs is gaining attention in this matter due to their capability in rapidly identifying the connection between water quality variables. However, the lack of available data and the associated challenges of measuring some parameters make data availability an obstacle in developing data-based models. In order to overcome this issue, data balancing algorithms have been newly introduced to enhance the predictive and data-based models by generating new samples based on real observations (Xu et al., 2020). The generated data can reduce the pressure on demands of data and the related costs of measurements.

Chapter 3: Prediction of *Cryptosporidium* using Bayesian Networks and Balancing Algorithms

3.1 Introduction

Cryptosporidium is a widespread pathogen in source waters imposing risk to the public and environmental health (Hamilton et al., 2018). This pathogen was responsible for several worldwide outbreaks as it can cause infection even in low doses (Brion et al., 2001). The source of *Cryptosporidium* causing these outbreaks was reported to be mainly from agricultural runoff, fecal inputs and human wastes (Brion et al., 2001; Craun et al., 1998). One of the main features of this protozoa is its resistance to the natural decay in the environment and a very high survival rate in cold and dark conditions. Compared to the other waterborne pathogens, *Cryptosporidium* has one of the highest resistance to chemical treatments and other common disinfection processes in treatment systems (Kim et al., 2004). Investigations in the fate of *Cryptosporidium* in the environment, have reported that it can survive for months and even years in these favorable conditions (*Cryptosporidium: Drinking Water Health Advisory*, 2001).

However, the main challenge of assessing *Cryptosporidium* in water bodies is the measurement difficulties. Since the required advanced molecular techniques and sufficient expertise can cause a high cost and time gap. Hence, early warning systems that can rapidly detect the presence of *Cryptosporidium* can be an effective surrogate for direct measurements. Prediction of *Cryptosporidium* in source water based on available data can detect and mitigate the risk before plant intake. Therefore, monitoring this pathogen before intaking it into water utilities is of critical importance to prevent potential health risks.

Bayesian Belief Networks (BBNs) is a promising method that can be utilized to model the complex interconnection between different characteristics involved in source water quality and predict the *Cryptosporidium*'s absence or presence (Christophersen et al., 2018). BBNs are probabilistic graphical models well suited to represent the probability of a given observation based on the connected variable states (de Vries et al., 2021; McCann et al., 2006). BBNs are increasingly used for simulating environmental systems because they can capture the uncertainty associated with the natural systems by predicting the outcome of interest as likelihoods (Death et al., 2015; Uusitalo, 2007b). One of the advantages of using BBNs is being able to simulate different scenarios by defining possible cause-effect relations based on available knowledge (Landuyt et al., 2013; McCann et al., 2006). Furthermore, this method provides a visualization of these cause-effect relationships, making them an efficient communication tool for decision-

makers to better understand the systems under study. However, although BBNs are capable of learning the relationship of different variables from incomplete data, they cannot compensate for the lack of data from natural systems. Especially considering the evaluation of pathogens such as *Cryptosporidium*, the difficulty and cost of measurement techniques have resulted in databases without balanced information of both absence and presence of pathogens. Therefore, although this technique can be a powerful tool for assessing *Cryptosporidium* levels in source water, the issue of lacking data can significantly impact the prediction performance.

Data balancing algorithms can lift the limitation of data-driven methods in modelling systems with unbalanced data. These algorithms generate new synthetic data (based on the initial data distribution) of samples in low numbers to balance both minority and majority classes. Therefore, the use of synthesized datasets has been reported to improve the prediction capability and performance of the models (Luengo et al., 2011a; Xu & Liu, 2013). In the context of this work, the persistent challenge regarding modelling pathogens level and predicting *Cryptosporidium* is the lack of samples indicating the presence of *Cryptosporidium* and the high proportion of non-detects or zero values in *Cryptosporidium* monitoring datasets. These unbalanced datasets will significantly reduce the ability of data-driven models to learn variable relationships and limit the ability to predict positive cases when protozoa will be present.

Therefore, two commonly applied data balancing algorithms, Adaptive Synthetic Sampling Algorithms (ADASYN) and Synthetic Minority Over-sampling Technique (SMOTE) have been utilized to generate the samples indicating the presence of *Cryptosporidium*. The generated data using these algorithms were then utilized for developing BBNs model to predict the absence/presence of *Cryptosporidium*. The following sections present the applied steps and the observed results.

3.2 Material and Methods

3.2.1 Data Preparation and Discretization

This study utilized datasets for developing prediction models belong to four different monitoring sites. One of these datasets included *Cryptosprodium* concentrations, and the other three have contained the records of *E. coli* concentration. The dataset including *Cryptosporidium* was obtained from New York City's open database reported from the Department of Environmental Protection (NYC, Department of Environmental Protection, OpenData, 2020). The parameters and *Cryptosporidium* observation in this dataset were recorded from the effluent of the Kensico reservoir as the drinking water provider of New York City. This reservoir was constructed in 1915

with the capacity of holding 30.6 billion gallons. The primary water source of Kensico reservoir is from Catskill and Delaware Aqueducts, but it also receives 2% of the water from its own watershed. This reservoir is the last station of the other reservoirs making the drinking waters available for daily consumption of New York City. The location of the reservoir is 15 miles north of New York and 3 miles of white plains with a latitude of 41.0737078 and -73.7659656 longitude (NYC Environmental Protection). The samples were taken from the effluent of the reservoir before intaking to drinking water and before the disinfection process (NYC Environmental Protection). Besides *Cryptospridoum*, fecal coliform and turbidity as another representative of water quality indicators were extracted for the same case study. In addition to water quality parameters, weather variables were also obtained from the Westchester County Airport station (National Centers For Environmental Information,2020) as the closest weather station to Kensinco Reservior. The extracted weather parameters of this dataset have included precipitation on the sampling day, over three days prior, maximum air temperature, and minimum air temperature (average of minimum and maximum temperature was considered in this study as the indicator of sampling temperature). Therefore, the final considered parameters for the Kensinco Reservior were: turbidity, fecal coliforms (measured based on CFU/100 mL), and *Cryptosporidium,* with a total of 238 samples from the summer of 2015 up to the summer of 2020.

The BBN models were developed using GeNIe (Bayesfusion LLC, Pittsburgh,PA) software. However, before developing the BBN models the following steps were employed to prepare the models; Step 1) Excel software were used to make sure data and prameters are orgnized (i.e., all data has their own label such as turbidity, temperature etc.); Step 2) Data set imported to the GeNIe software and discretized based on uniform counts in each bin. In this work, the objective was a binary prediction of the pathogens. Therefore, the reported values of *Cryptosporidium* concentration were divided into two classes. The presence has included the dates that had more than zero (oo) cysts/100 liter, and the absence of *Cryptosporidium* was indicating a zero report of *Cryptosporidium* on that specific date. Turbidity, fecal coliform, temperature were disctretized to 3 bins and precipitation on sampling day and over three days were discretized to 2 bins. All parameters were discretized to have an equal number of data in each bin. These two parameters were discretized to 2 bins because recorded data were zero for most of the samples, and it was aimed to have an equal number of samples in each bin. These discretized data for each monitoring site were used for training the structures. The reason for discretizing data before developing BBN models is that discretization of variables helps with the complexity of the models by removing the need for assuming continuous probability distribution.

Also, speeding up the model performance reduces the computation costs for large-scale applications. Furthermore, most microbial parameters are usually assessed based on specific thresholds and classified ranges. However, discretizing causes missing some information that could be obtained from continues variables and data and it can be of future interest to initiate the models without discretizing the data; Step 3) Based on the explained details in section 3.2.4 structure of BBNs were created each parameters' data in training set were assigned to the appropriate nodes and model was trained based on discretized training data set. The performance of the model was assessed using test dataset and comparing the number of accurate predictions to the measured values.

3.2.2 Over Sampling Data with Balancing Algorithms

One of the main barriers to applying machine learning algorithms to environmental systems is class imbalance. In the dataset used for predicting Cryptosporidium, most reported were 'zero' concentrations of protozoa or below the detection limit (92% of all data) and there was a significant difference between the samples indicating the presence and absence of Cryptosporidium (n=18 and n=220). The small subset of data is called the minority class ($\mathrm{n}_{\mathrm{minority}} = 18$) and, 92% of samples make the majority class of the data ($n_{majority} = 220$). Unbalanced datasets with small minority classes (i.e., presence of the organism) are challenging to use and often result in poor-performing models. Data-driven models aim to reduce error and increase the accuracy of predicting both positive and negative classes, the algorithms will work better when the number of classes is equal. Also, an imbalanced data set can increase the chance of the model being overfitted with one class because the model only summarizes and repeats one specific class regardless of what other information is coming from input parameters. To address this issue, the capability of two data balancing algorithms to generate a number of positive samples and balance the dataset was investigated.

3.2.2.1 Synthetic Minority Oversampling Technique (SMOTE)

The core objective of data balancing algorithms is to generate artificial samples to make the number of all classes equal. SMOTE is one of the basic and initial algorithms proposed for balancing imbalanced data. The algorithm randomly selects some examples from the minority class (x_i) and then determines k-nearest neighbours of the selected samples. The distance between the nearest neighbour and the sample from the minority class gets multiplied to a random

number between 0 and 1 (w). The following equation simply shows how a new sample (z) can be generated based on the selected sample:

$$z = x_i + w(\, x_{iz} - x_i) \qquad (eq.3.1)$$

Here w is a random number in the range of 0 and 1, x_i is the chosen sample and x_{iz} is one of its, k neighbours.

SMOTE selects the samples and generates data based on them until the number of samples in both minority and majority classes gets equal. However, the algorithm selects the samples without paying attention to the overlaps between minority and majority classes and can choose solo samples of minority classes among the majority ones. This approach for some datasets can be a source of concern because the resulting high number of minor samples among the majority class makes the classification models confused and can decrease the model's performance. To overcome this concern, another balancing algorithm has been proposed based on SMOTE algorithm. These modified methods force the algorithm to be selective of the particular examples that are used from minority classes to generate the data. One of these algorithms is described in the following section**.**

3.2.2.2 Adaptive Synthetic Sampling Algorithms (ADASYN)

One of the most reliable data balancing algorithms is ADASYN, a modified version of the SMOTE method. The only difference of this method with its origin algorithm (SMOTE) is in the way of selecting samples in the minority class. Opposite to the SMOTE, which selects the sample arbitrarily, ADASYN selects data proportional to the density of samples in the minority class. For instance, in a region with a lower number of samples, ADASYN generates more samples as the density is low and fewer samples would be generated in regions of high density. This feature of using density distribution can help achieve more uniform balancing rather than over-replicating samples in already dense areas.

ADASYN, prior to selecting the samples, calculates the density distribution for each sample in minority class based on the following equation (He et al., 2008):

$$\mathrm{r}_i = \frac{\Delta_i}{k} \qquad (eq.3.2)$$

Here the k and Δ_i are indicating the number of k -nearest neighbors of the selected sample (x_i) in the minority and majority classes, respectively. The overall number of samples that need to be generated for balancing the data is calculated based on the following equation:

$$G = (n_{majority} - n_{minority}) \times \alpha \qquad (eq.3.3)$$

Where α is again a number randomly assigned in a range of 0 and 1, determining the degree of generalization. However, ADASYN modifies the number of required data by:

$$g_i = r_i \times G \qquad (eq.3.4)$$

Therefore, for each sample (x_i) in the minority class g_i samples can be generated in the same way of SMOTE algorithm indicated in $(eq.3.1)$.

It should be mentioned that evaluating model performance using synthetic samples could be misleading since generated samples were not truly observed. As such, only actual observed data were included in the test set, and synthetic samples were only used for training purposes. Furthermore, the data being used to generate synthetic data should be carefully considered because the new samples are generated based on the real data and having wide representative samples likely impacts balancing algorithms in generating more conclusive data. Therefore, splitting the training and testing subsets of the data can be important. In this study, two approaches were used to derive a set of data used to generate synthetic data for balancing. First, all the samples with the initial degree of imbalance (the difference of major and minor class) have been used to synthesize the samples with ADASYN and SMOTE. As a first approach, the same minority class data used to generate samples was then reused in the test or validation set for evaluating performance, although exact samples were not reused between splits. This approach maximizes the ability to generate a wide range of representative synthetic samples since all minority samples were considered. However, bias may be introduced this way since synthetic data will be influenced by real data used in testing and accuracy specific to the minority class may be overemphasized. Therefore, to compare results and provide a measure of performance without the influence of the synthetic data generation on true test data, a second approach was taken where training and test data split is prior to synthesizing samples when 50% of samples in the minority and around 70% of samples in majority classes were separated and used for generation and training the model. As such, the remaining true data in the test were not either in synthesizing data or training the models. Figure 3.1 depicts a flow chart visualizing how the dataset is split and

used for generating synthetic data and training the models. These training and testing datasets, as well as the original unbalanced dataset, were used for developing BBN models.

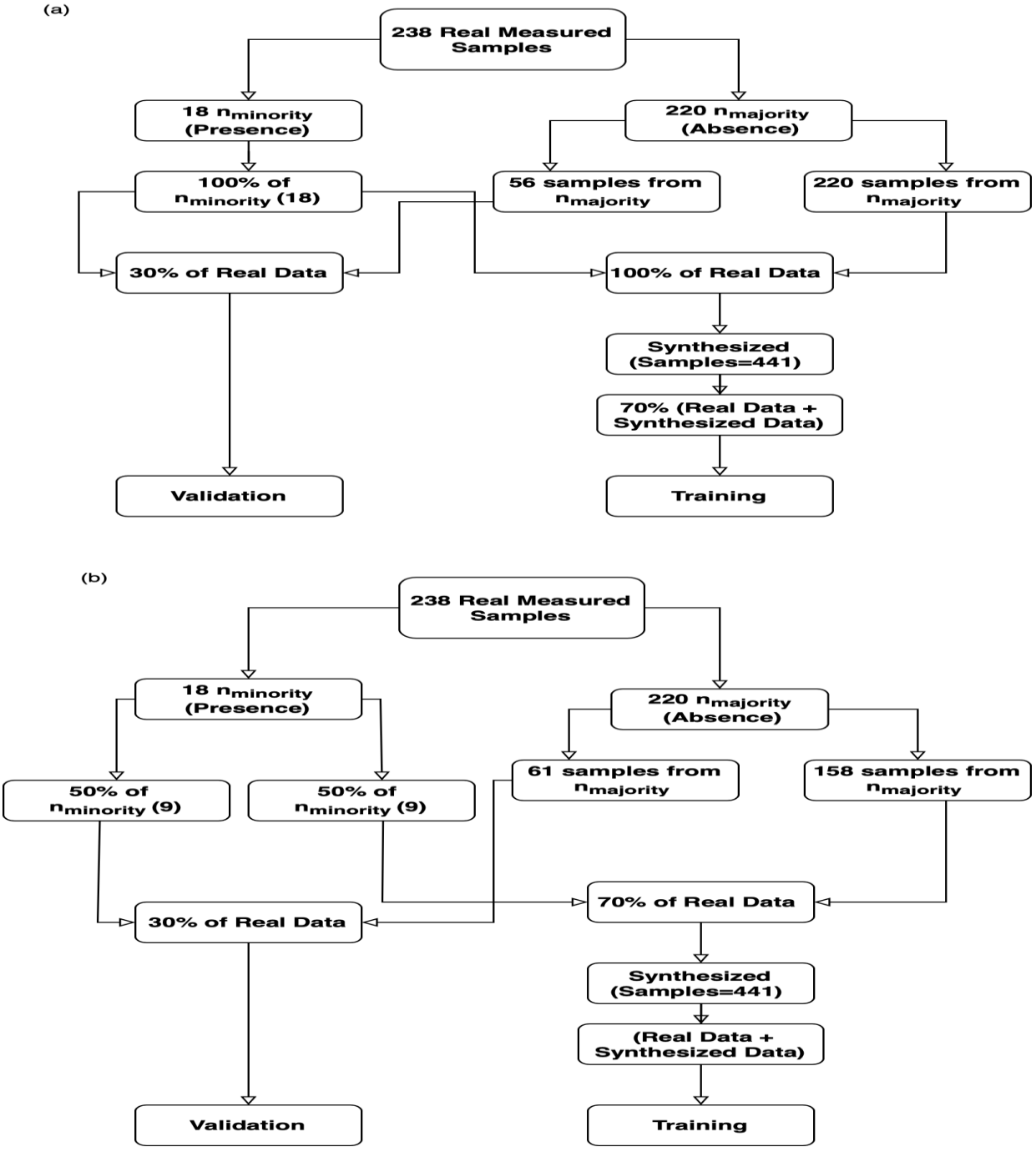


Figure 3.1 Flow chart of the modelling framework a) test data used in balancing b) test data not used in balancing. Notice that n minority and n majority indicate the number of samples in minority and majority classes, respectively.

3.2.3 Statistical Analysis of Variables

In order to better understand the flocculation of parameters over time and to identify any existing trend or pattern in data, the time series of each variable were depicted and presented in Figure 3.2. Time series of weather parameters shows repetitive pattern in temperature and precipitation which was expected due to the seasonal nature of climatic features. While these repetitive flocculation over time can be leveraged in predicting the future changes, microbial parameters (*Cryptosporidium* and fecal coliform) lack any meaningful pattern. This can be due to the diversity of parameters that can impact the presence or absence of pathogens in environmental systems.

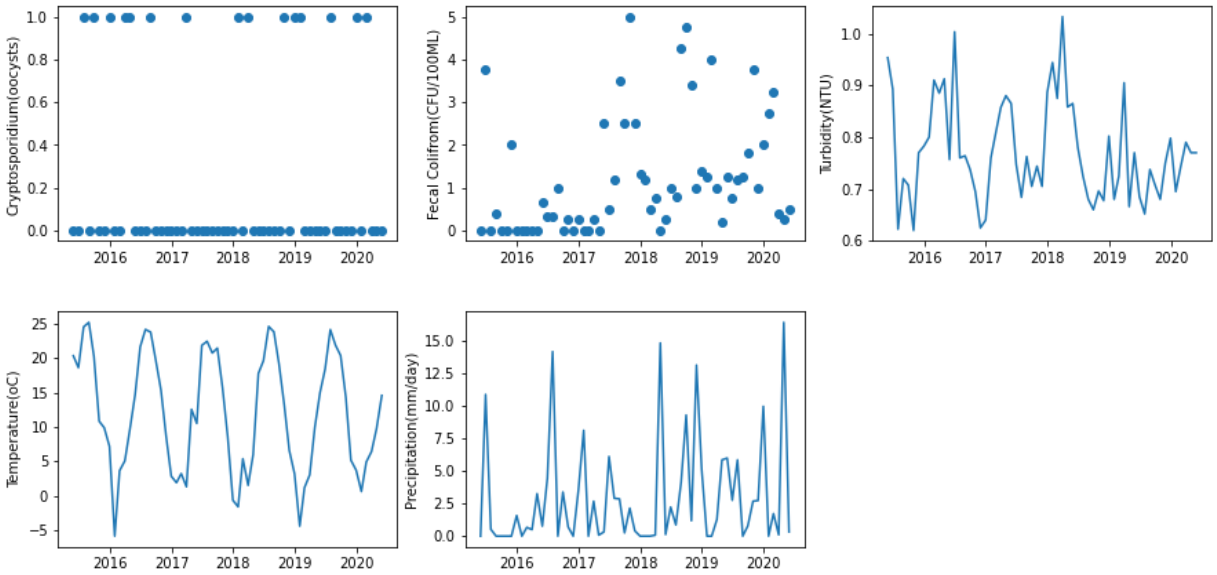


Figure 3.2 Visualization of the time series for weather and water quality parameters of Kensico reservoirs.

To gain more in-depth insight over parameters' historical variation, Autocorrelation Function (ACF) was used to explain the similarity between parameters observation in a function of lagged time. This function allows to estimate how the parameter's state in current time is affected by its observation in the previous time steps (Bocquet, 2022). ACF is a statistical tool to extract variable's feature based on their movement frequency (Zarei & Mohammadzadeh Asl, 2020). For example, a higher ACF can be indicator of a meaningful frequency over the function of time (Dwivedi et al., 2016) .The resulted analysis is illustrated in Figure 3.3 for all considered parameters.

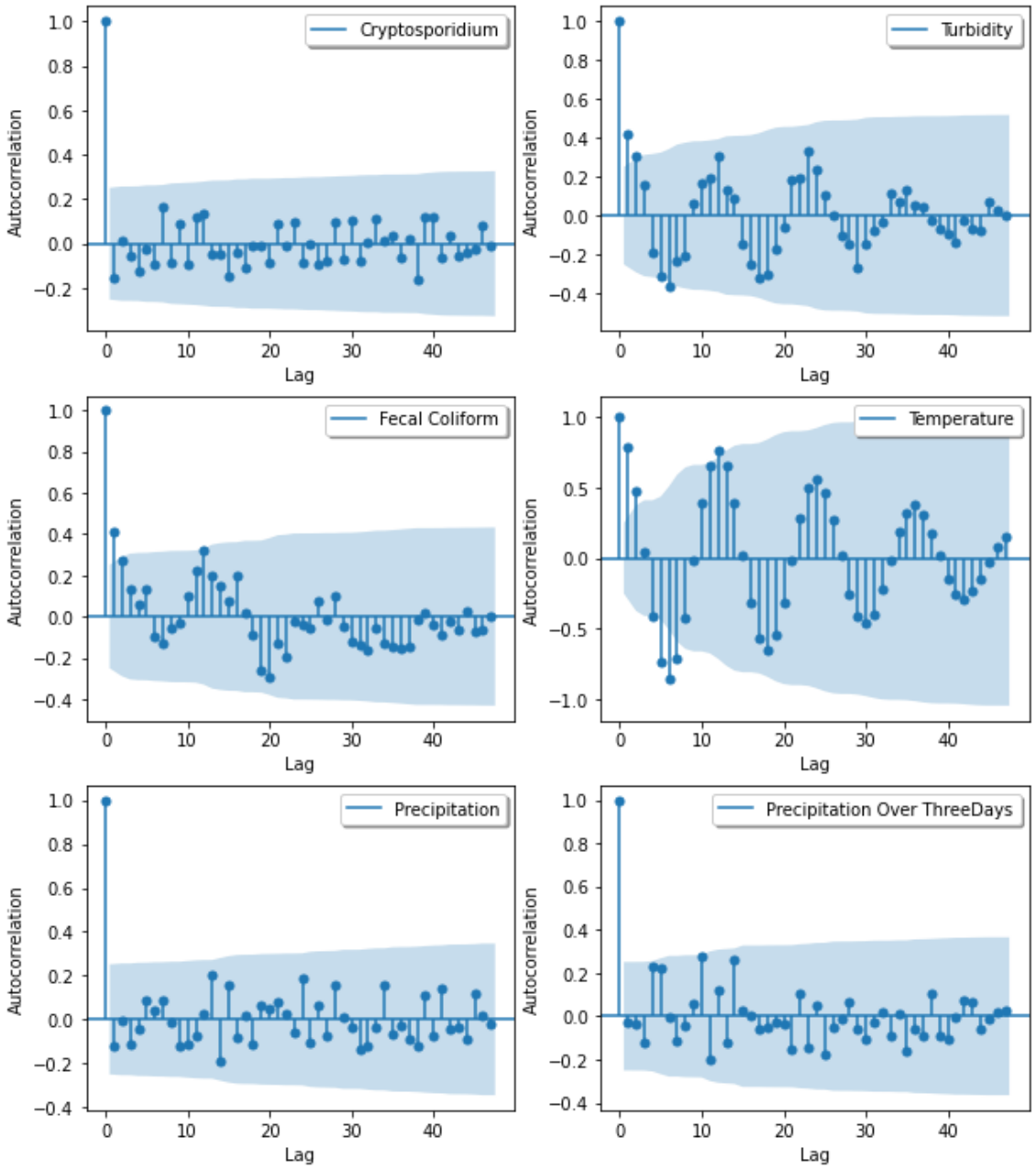


Figure 3.3 Autocorrelation of all parameters based on one month time lag and confidence interval of 95%.

As it can be observed from graph, temperature seems to have a repeating pattern over first 2 months and 12 months. Besides temperature, the repetitive pattern can be observed from ACF plot of turbidity and fecal coliform. A similar observation for *E. coli* (as a fecal coliform) seasonality was also reported by Dwivedi et al. (2016). Based on ACF analysis of *E. coli* occurrence in groundwater, the study indicated a repeating pattern of 7 and 10 months due to the

impact of climatic factors (Dwivedi et al., 2016). Similarly, since a strong ACF (>0.5) was observed here for temperature in one month, the repeating pattern for turbidity and fecal coliform can be due to these flocculation of weather parameters which seems to impact water quality indicators. Although Muchiri et al., (2009) has indicated a seasonal pattern for *Cryptosporidium* peaking in surface waters, in this study not any considerable repeating pattern was observed for *Cryptosporidium* based on ACF (Muchiri et al., 2009). This can be due to the difficulties in measuring *Cryptosporidium* that has resulted in a poor dataset when most of the samples include "non-detects" and the frequency of samples indicating *Cryptosporidium* is very low. Therefore, it is possible that correlation between observed *Cryptosporidium* with its lagged value be very poor.

In order to identify these underlying relationships between parameters the linear correlations between variables were also determined (Figure 3.4). Furthermore, to visualize the spread of variables and their dependency to each other, the scatterplot for each pair of variables have been illustrated. Due to the higher number of resulted graphs, Figure 3.7 only indicates the scatterplot of turbidity with all other parameters and the remaining graphs are presented in appendix.

As the scatterplot in Figure 3.6 and correlation results in Figure 3.3 indicates, linear dependency between all parameters and *Cryptosporidium* were poor ($R < 0.3$), and only fecal coliforms and turbidity were observed to have positive correlations with *Cryptosporidium* levels. It was expected that weather parameters (precipitation and temperature) could be correlated with *Cryptosporidium* based on the previous reports of changes in surface water protozoa concentrations during and after severe weather conditions (Duris et al., 2013a; Young et al., 2015). However, weak linear correlations between *Cryptosporidium* and weather parameters were found in this dataset. This finding is aligned with the result of the review study conducted by Young et al. (2015). This review has indicated that the correlation between precipitation and *Cryptosporidium* was non-consistent and site-specific in different studies. Therefore, it seems that linear observations cannot capture the complicated relationship between *Cryptosporidium* and weather events.

The strongest correlation compared to other variables was observed between precipitation over three days and fecal coliforms used as indicator organisms ($R = 0.27$). However, correlations between *Cryptosporidium* and fecal coliforms/turbidity were low ($R = 0.006/0.0045$). The observation of poor correlations between *Cryptosporidium* and fecal coliforms or turbidity supports the idea that these indicators are insufficient to assess microbial water quality. While stronger relationships were expected between these parameters, this observation is aligned with several

previous studies that concluded *E. coli* and fecal coliforms are not robust indicators of protozoa in surface waters (Lalancette et al., 2014a; Wohlsen et al., 2006b).

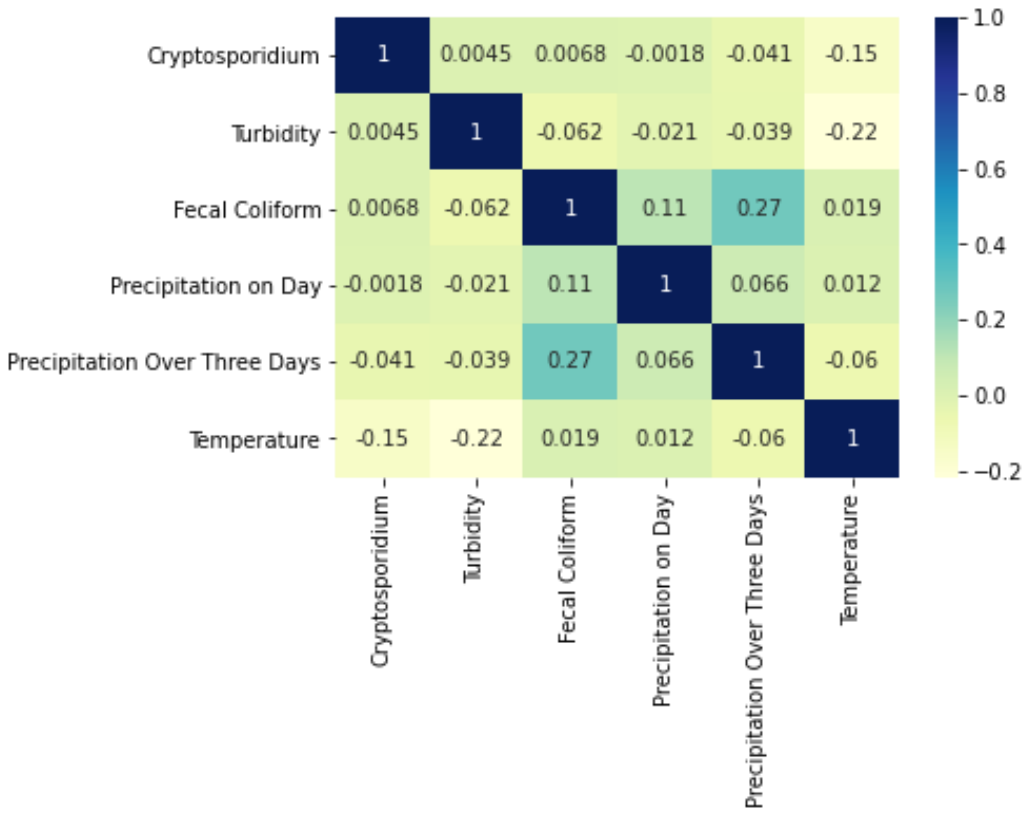


Figure 3.4 Correlation coefficients (R) of water quality and weather parameters.

It was hypothesized that the weak correlations observed in this dataset might be due to the design of the sampled reservoir, sample location within the reservoir, or the probability that samples were not taken during extreme weather events. The sampled reservoir has been designed and maintained to protect New York City's drinking water quality, even under severe weather conditions (National Academies of Sciences, Engineering, and Medicine. 2020). For example, flow into the reservoir is reduced to always maintain turbidity under 5 NTU (Nicholas et al., 2016). Other design features of the reservoir and the surrounding area, including soil type, vegetation, and the slope of the banks, can also affect how rainfall affects water quality (Duris et al., 2013a; Mavimbela et al., 2019). The sampling point within the reservoir should be considered, and previous reports have indicated the effects of sampling points on measured levels of *Cryptosporidium* by more than 58% (Ligda et al., 2020b). The sensitivity of the pathogen levels to the sampling location is not limited to *Cryptosporidium* but also to measured levels of *E. coli* and fecal coliforms that may be used as indicators (Herrig et al., 2019).

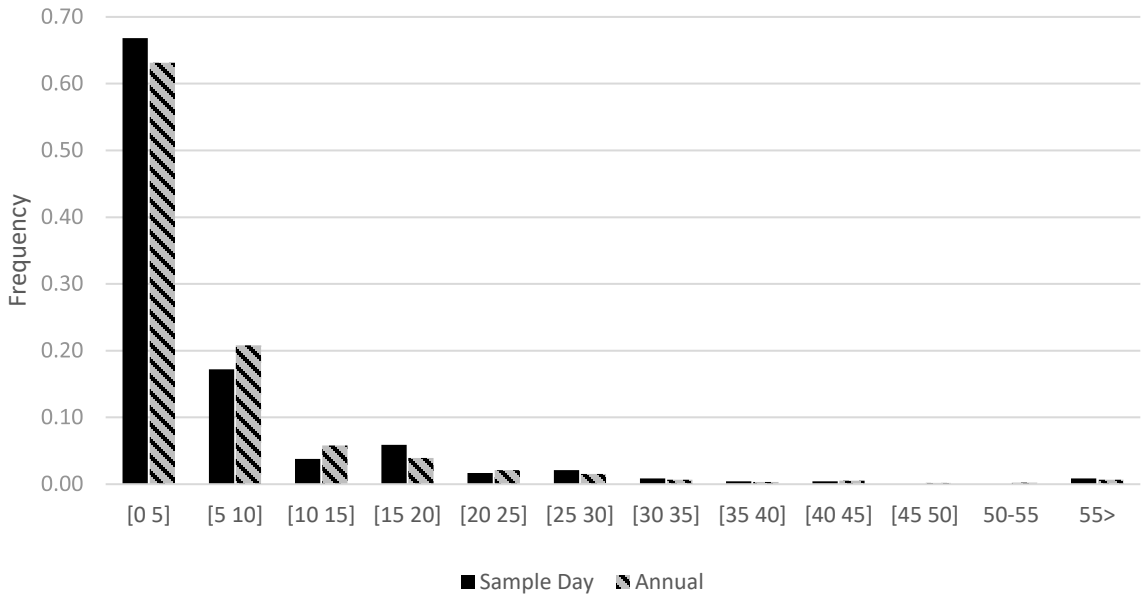


Figure 3.5 Frequency Histogram of normalized yearly and sampled precipitation level. (Although the precipitation on sampling day is a subset of annual precipitation, it should be noticed that the fraction and ratio are normalized.)

It is possible that the samples were taken during non-representative weather conditions (i.e., low precipitation days), which could influence the overall distribution of *Cryptosporidium* concentrations. Therefore, precipitation on all days over the sampling period (1977 days) was also collected to consider any possible differences between actual precipitation distribution and sampled precipitation distributions (from the 238 sample days). The two precipitation datasets were compared using the non-parametric Kruskal Wallis test to assess similarity. It was found that there was a difference between precipitation on sample days and expected precipitation during the sample period (95% confidence level). The normalized frequency of precipitation of both yearly and sampled datasets is shown in Figure 3.5. Sampled days slightly overrepresented low or no precipitation days (66% of sampled days vs 63% of all days) and overrepresented heavy rainfalls > 50 mm/day (10% of sampled days vs 7% of all days).

Furthermore, the poor correlation coefficient between *Cryptosporidium* with both weather and water parameters Indicates that dependency of pathogens to other variables can hardly be captured by linear analysis such as correlation coefficient or even linear regression coefficients analysis. For instance, regression coefficient tries to explain the variation in the *Cryptosporidium* based on the changes in turbidity or other parameters, while Figure 3.6 indicates that the association between the spread/variation of *Cryptosporidium* over turbidity is very poor and

finding an optimal regression coefficient that defines the best line for predicting future concentration of *Cryptosporidium* based on turbidity can be challenging.

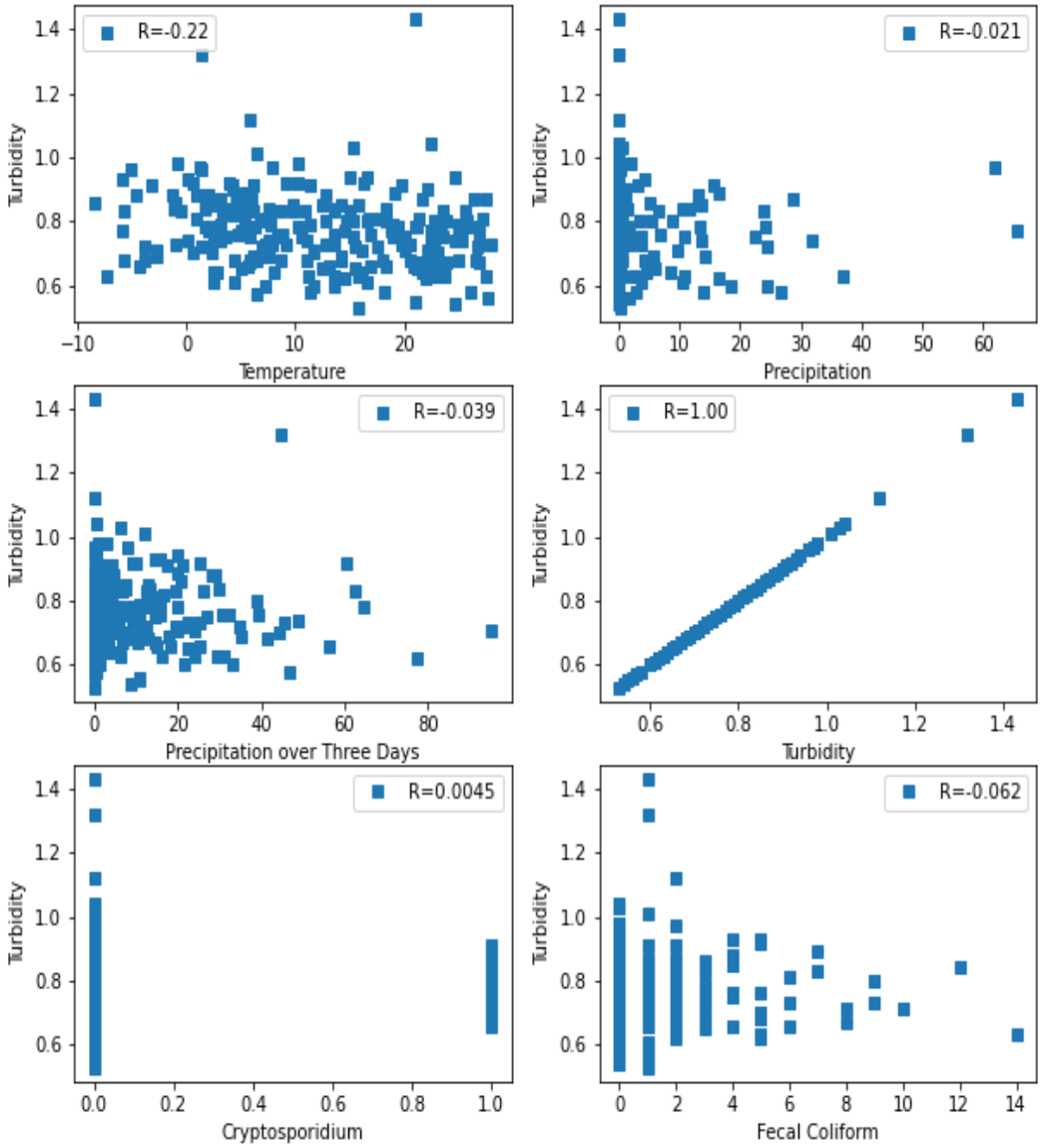


Figure 3.6 Example of scatterplots developed to visualize parameters and their relationship with turbidity as an example. The remaining graphs for all other parameters have been presented in appendix.

The analysis in this section were implemented using pandas, seaborn and matplot libraries of python 3.9. The matplot and seaborn libraries were used for generating plots from the results and the Pandas library was used for calculating the correlation coefficient and autocorrelation between variables.

3.2.4 Bayesian Belief Networks (BBNs)

Bayesian Belief Networks are constructed based on the Bayes theorem that can capture variable relationships in a probabilistic structure (Bertone et al., 2016b). Based on the conditional dependencies, BBNs can describe and calculate the probability distributions of variables and represent other factors' impact on the probability of an outcome. BBNs can be defined by a "Directed Acyclic Graphs (DAGs)" graph in which nodes represent the variables, and their connecting arcs indicates these variables' probabilistic influences. More precisely, arcs point out from the parent node toward a child node can define the conditional dependency of these two nodes. Therefore, the absence of an arc between two nodes indicates their independence. This graphical representation of variable relationships in DAGs is the main advantage of BBNs where each node (variable) can be connected to another node indicating a one-way dependence (Fasaee et al., 2021a). The number of connections or condition probabilities is dictated by the number of variables and the developed structure. Each variable contains two or more states, and each state has a prior probability $P(x_1)$ for an event x_1 and the summation of the prior probability of all states is equal to 1 for each node. These prior probabilities for each state and node normally are introduced to the model by training data (Woolf, 2009). The prior probability can be updated by introducing new evidence or observation of $x_1$like x_2 and the further probability would be $P(x_1|x_2)$ which is called posterior probability. Basically, BBNs calculate the probability of some variable of interest by introducing evidence or the condition on the parent nodes. Based on the conditions over the variables, conditional probability tables can be developed for each node simply based on the following formula (Han & Kamber, 2019) :

$$P(x_1|x_2) = \frac{P(x_2|x_1)P(x_1)}{P(x_2)} \qquad (eq.3.5)$$

Where $P(x_2|x_1)$ shows the conditional probability for x_2 given x_1 which can be re-written in terms of joint probability distribution as follows:

$$P(x_2, x_1) = P(x_2|x_1)P(x_1) \qquad (eq.3.6)$$

By identifying the "key parents", BBN calculates the joint probability only for these variables and simplifies the calculation required for developing the conditional probability for n number of variables as follow (Panidhapu et al., 2020):

$$P(x_1, \dots x_2) = \prod_{i=1}^{n} P\,(x_{\mathrm{i}} | Parents\,(x_{\mathrm{i}})) \qquad (eq.3.7)$$

In this study, GeNIe software was used for calculating and updating the conditional probabilities after establishing the variables, defining their connection, and developing the BBN structures. Creating BBN structures due to including the variable's connection can be critical because the CPTs are constructed based on the network's topology. The following section explains the detail of the employed approaches for developing the BBN structures and initiating the prediction models.

3.2.5 Structure Development of BBN

The primary step in developing BBN is learning the structures of the model and defining parent and child nodes based on the feature parameters because BBNs develop the conditional probability tables based on the established connection between nodes. Therefore, it is important to include all possible connections of variables in a correct way and create an optimal structure that is computationally feasible.

Developing the graph or learning structure is normally done with two main categories of constraint-based and score-based methods and hybrid algorithms (Scutari et al., 2019). The structures developed based on constraint-based methods identify the independencies. In the case of dealing with the data in small size (<20 variables), the constrain-based techniques are reported to be faster and more accurate. One of the common algorithms of the constraint-based method is the PC algorithm. This algorithm is shown to be one of the fastest algorithms commonly used for structure learning. The algorithm constructs a complete graph with a full edge between variables and removes the connections based on conditional independence of parameters (Mittal & Maskara, 2011b).

One of the structure learning algorithms that are known for having simple structures, easy construction and fast learning, is Naïve Bayes algorithms. As the name indicates, this method is a naïve method of developing BBN structures because instead of learning structures from data, Naïve Bayes builds the network's topology by assumptions. This approach considers the target node as the solo parent of all other predictors/variables and assumes that all features are independent. This feature makes it a good candidate when there is no significant correlation between variables (Menti et al., 2016). Although the Naïve Bayes is a straightforward and fast method, it misses and possible dependency between predictors. In order to overcome this

shortcoming, Tree Augmented Naïve Bayes was introduced and normally developed beside Naïve Bayes for developing BBN structures in the literature studies (Downs & Tang, 2004; Jiang et al., 2005; Li & Abdul Rahman, 2018). This algorithm creates the initial structure based on the Naïve Bayes but adds to the connection between parameters based on their dependency. This algorithm relaxes the assumption of no dependencies between features and builds the edge between parameters based on a tree structure (Jongsawat & Road, 2017). The class node (or target parameter) is defined as the root of the structure, and other edges are formed "pointing outwards" and form a tree structure in a way that the root node cannot have any parent. As mentioned, Naïve Bayes assumes one parent for all parameters (the class variable or target node) and Tree Augmented Naïve Bayes allows all parameters to have one more parent than class nodes. While these algorithms limit the number of parents for each variable to only two parents, Augmented Naïve Bayes algorithms lift the limit on having parent nodes. It should be mentioned that, although Augmented Naïve Bayes improves the accuracy more than the later ones, it can add to the complexity of the structure accordingly. More computational detail of these structure learning algorithms can be found in the study conducted by Singharoy, 2018.

In addition to the three approaches of Naïve Bayes, Tree Augmented Naïve Bayes and Augmented Naïve Bayes that are developed in this study as complementary algorithms, other two structure learning algorithms are employed: Bayesian Search and Greedy Thick Thinning Algorithm. Bayesian Search is one of earlier score-based algorithms that starts with a random structure and updates the connection based on the relative posterior probabilities (Heckerman et al., 1995). Initiated by Bayesian Search, Greedy Thick Thinning Algorithm includes two-steps of thickening the structures and the step of thinning. In the earlier step, the algorithm creates structures without any arcs and builds the arcs that increase the marginal likelihood. The algorithm continues to build arcs until the marginal likelihood connection stays unchanged. Then during the thinning step, the algorithm removes the arcs that cause increased marginal likelihood until the marginal likelihood remains unaffected.

In this study, structure learning algorithms are used to find the relatively best performance in predicting pathogen's presence and the objective was 1) to provide a base of comparison for the prediction models and 2) analyze the sensitivity of model performance and target variable to the connection between each parameter.

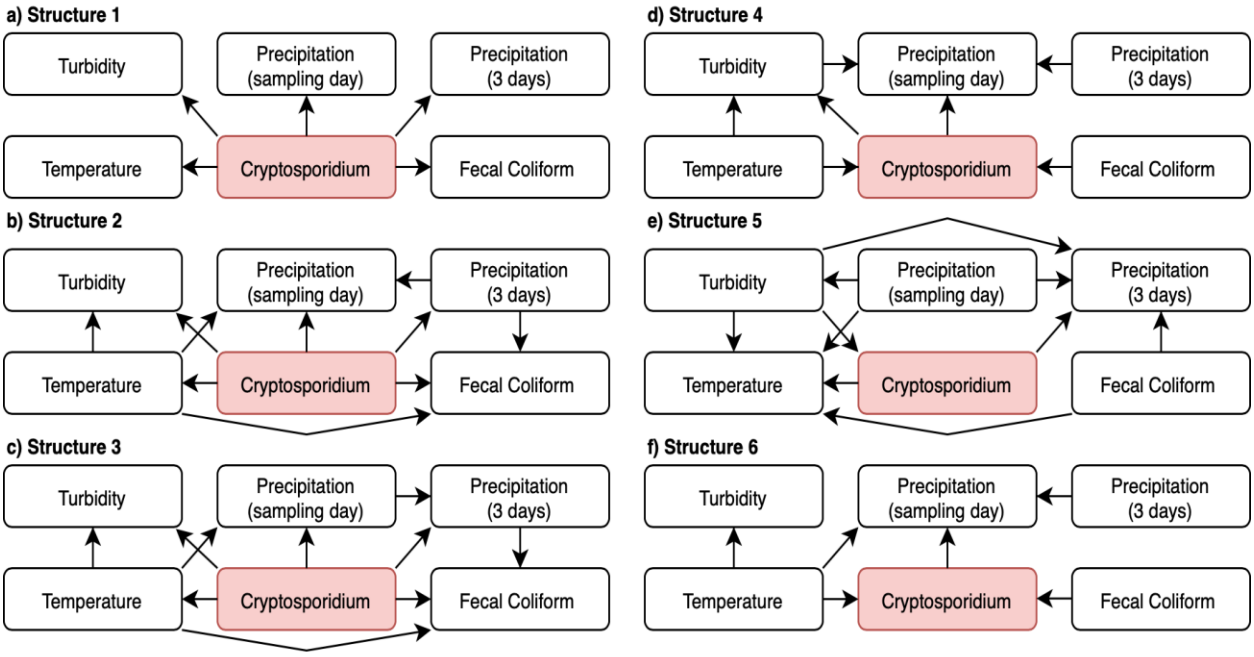


Figure 3.7 Learned BBN structures based on a) Naïve Bayes b) Augmented Naïve Bayes c) Augmented Tree d) Greedy Thick Thinning e) PC f) Bayesian Search algorithms.

Learning BBN's structure is considered an NP-hard problem and studying all possible algorithms and the detail of algorithms requires a separate study with an in-depth analysis of the algorithms. As such, there is still several studies and debate about the best method of learning and developing BBN structures (Jongh, 2014; Scutari et al., 2019) because the number of possible structures increases with number of variables, and each structural learning algorithm can identify varying connection. Therefore, selecting one graph for developing a model among many potential candidates can be a challenging practice (Mittal & Maskara, 2011b). Therefore, the models in this work were initially developed based on the popular structural learning algorithms (Figure 3.7) which was applicable through GenNIe software (BayesFusion LLC, Pittsburgh, PA): Bayesian Search, Naïve Bayes, Augmented Naïve Bayes, and Tree Augmented Naïve Bayes. In addition to structural learning, *prior*/ expert knowledge was also used in the current work to develop the BBNs and prediction models. Using expert knowledge allows the decision-makers and modelers to reflect the potential physical dependency and connection between variables for assessing their influence on the outcome of interest. The results of the implemented networks are presented in the results and discussion section.

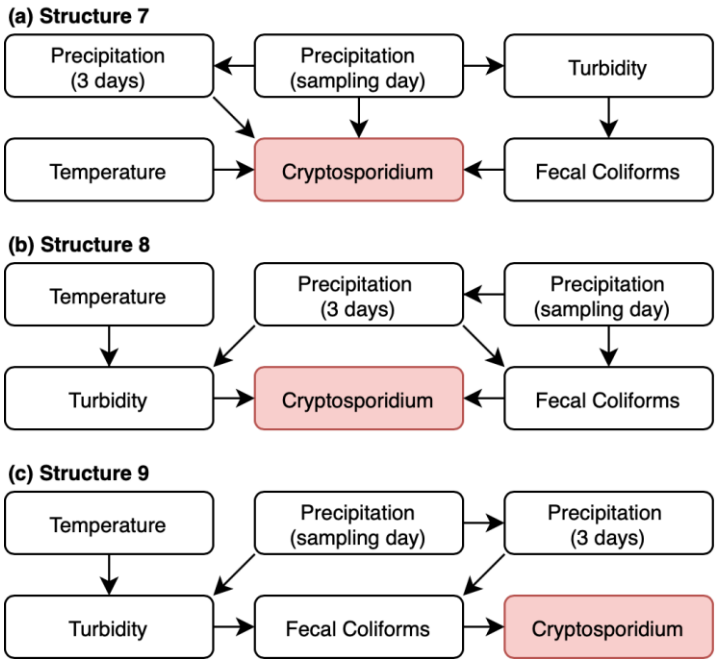


Figure 3.8 BBN Structures were developed based on expert knowledge.

As expected, the datasets generated by different balancing algorithms have developed distinct structures due to algorithmic and generated data differences. The exception to this was Naïve Bayes since this approach connects all variables independently to the dependent variable in all cases.

Besides the structures learned based on the given data and algorithms, two other structures were developed based on *a priori knowledge* or expert knowledge of expected relationships between variables. The dependencies of variables can be represented in BBNs' structures. For instance, it was anticipated that temperature impacts the turbidity regime because an increase in temperature can result in thawing permafrost or snow melting, which in turn can increase turbidity (Jolivel & Allard, 2017). Similar to the temperature, increased precipitation can also generate or wash out more sediments or particles to the flow and add to the turbidity level. Turbidity increase can also be an indicator of microorganisms' presence, including fecal coliforms or *Cryptosporidium*.

As shown in Figure 3.8 (a), direct connections have been defined between weather parameters and *Cryptosporidium. In contrast,* in the second structure (Figure 3.8. b), these connections have been indirectly applied to the target parameters. This approach has been employed to see if the weather condition directly contributes to the microbial concentration of source water or their main impact is on other characteristics of water such as turbidity and

hardness, while the variation on these characteristics provides an opportunity for microorganism's growth or activity.

In order to examine the approach of surrogating fecal bacteria for *Cryptosporidium*, a third structure has been developed for predicting *Cryptosporidium* in which the only predictor variable is fecal bacteria, and the other parameters only impact this variable. The modelling results for each structure and prediction of *Cryptosporidium* have been reported in the following sections.

3.2.6 Assessment of Model Performance

The available data for each case study was divided into two subsets of training and testing datasets. The models were tested based on predicting the target variable in testing data, and the performance of the model was assessed based on the correct number of these predictions (accuracy of the model). The model's accuracy can be computed based on a confusion matrix that describes the number of accurate predictions versus the total number of predictions, including the false predictions (W. Li & Guo, 2013). Table 3.1 indicates a binary example of a confusion matrix where the diagonal indicates all the correct predictions.

Table 3.1 Confusion Matrix to assess model performance (a binary example for predicting the absence or presence of pathogens)

		Real-Time Measurement	
		Presence	Absence
Observed / Prediction	Presence	True Positive	False Positive
	Absence	False Negative	True Negative

The accuracy of prediction is basically the diagonal of the confusion matrix or the total number of accurate predictions over the total number of predictions (N) and can be calculated based on the following formula:

$$Accuracy = \frac{Ture\ Positives + ture\ Negatives}{N} \qquad (eq.3.8)$$

3.3 Result and Discussion

3.3.1 Prediction of *Cryptosporidium* with Balanced and Un-Balanced Data using BBN

This section represents the result of predicting the presence and absence of *Cryptosporidium*. The learned and developed structures explained in the previous part were trained by unbalanced datasets. Then the models were tested by an unmodified test dataset, and the prediction results are shown in Table 3.2. The overall accuracy was high (> 90%); however, models generally did not predict the presence of *Cryptosporidium* for any sample.

Table 3.2 Prediction accuracy of *Cryptosporidium* with BBN using the unmodified dataset. Prediction results are based on a randomly separated test set (15%) of data

Model	Structure	Overall accuracy	Prediction accuracy of Absence	Prediction accuracy of Presence
BBN	Structure 1	92%	100%	0%
	Structure 2	92%	99%	0%
	Structure 3	90%	98%	0%
	Structure 4	90%	97%	0%
	Structure 5	90%	97%	0%
	Structure 6	90%	97%	0%

The reason for poor accuracy in predicting the presence of *Cryptosporidium* was that the dataset included a significant number of non-detects resulting in overfitting with absence samples. Furthermore, it cannot be concluded that the model was performing well in predicting the absence of *Cryptosporidium* either. Since both training and testing datasets has included a high proportion of absence samples and it is likely the model has been overfitted and only produces predictions of the majority class, regardless of information inputted from other variables.

It is unreasonable to expect that the model has learned the true decision boundaries or would generalize well to future data. The observed results using unbalanced data illustrate the common motivating challenge of this work, where models built on highly imbalanced data are overfitting to the majority class. Furthermore, the results demonstrate that simple performance metrics, such as overall accuracy, can mislead the interpretation of performance. The results shown in Table 3.3 indicate improvements in predicting the presence of *Cryptosporidium* using balancing algorithms. Regardless of data splits, balancing algorithms, and classification methods,

the overall accuracy could be lifted to more than 60% without having overfitting issues for either of the classes.

Table 3.3 The prediction of *Cryptosporidium* with ADASYN-BBN and SMOTE-BBN. Performance is evaluated both on (a) Scheme where all minority samples were used to generate synthetic training data. b) Subset of data not used to generate synthetic samples. The number of correct predictions over all made predictions are presented in parenthesis.

Model	Structure	**(a)** Test data used in balancing			**(b)** Test data not used in balancing		
		Overall Accuracy	Prediction Accuracy of Absence	Prediction Accuracy of Presence	Overall Accuracy	Prediction Accuracy of Absence	Prediction Accuracy of Presence
ADA-BBN	Naïve Bayes	69%($\frac{51}{74}$)	66% ($\frac{37}{56}$)	88%($\frac{16}{18}$)	58%($\frac{41}{70}$)	60%($\frac{37}{61}$)	22%($\frac{2}{9}$)
	Augmented	55%($\frac{41}{74}$)	50%($\frac{26}{56}$)	88%($\frac{16}{18}$)	66% ($\frac{46}{70}$)	79% ($\frac{48}{61}$)	22%($\frac{2}{9}$)
	Tree	66%($\frac{49}{74}$)	66% ($\frac{37}{56}$)	66%($\frac{12}{18}$)	61% ($\frac{43}{70}$)	64% ($\frac{48}{61}$)	22%($\frac{2}{9}$)
	Greedy	66%($\frac{49}{74}$)	62%($\frac{35}{56}$)	88%($\frac{16}{18}$)	58%($\frac{41}{70}$)	61% ($\frac{38}{61}$)	22%($\frac{2}{9}$)
	PC	66%($\frac{49}{74}$)	62%($\frac{35}{56}$)	88%($\frac{16}{18}$)	38%($\frac{27}{70}$)	39% ($\frac{24}{61}$)	22%($\frac{2}{9}$)
	Bayesian Search	69%($\frac{51}{74}$)	67%($\frac{38}{56}$)	77%($\frac{14}{18}$)	58%($\frac{41}{70}$)	61% ($\frac{38}{61}$)	22%($\frac{2}{9}$)
	Structure 7	52%($\frac{39}{74}$)	48%($\frac{27}{56}$)	77%($\frac{14}{18}$)	63%($\frac{45}{70}$)	66% ($\frac{41}{61}$)	33%($\frac{3}{9}$)
	Structure 8	69%($\frac{51}{74}$)	67%($\frac{38}{56}$)	77%($\frac{14}{18}$)	61%($\frac{43}{70}$)	63% ($\frac{38}{61}$)	33%($\frac{3}{9}$)
	Structure 9	37%($\frac{28}{74}$)	37%($\frac{21}{56}$)	33%($\frac{6}{18}$)	37%($\frac{26}{70}$)	37% ($\frac{22}{61}$)	33%($\frac{3}{9}$)
SMOTE-BBN	Naïve Bayes	54%($\frac{40}{74}$)	53%($\frac{30}{56}$)	55%($\frac{10}{18}$)	53%($\frac{37}{70}$)	55% ($\frac{34}{61}$)	22%($\frac{2}{9}$)
	Augmented	66%($\frac{49}{74}$)	64%($\frac{35}{56}$)	77%($\frac{14}{18}$)	60%($\frac{42}{70}$)	63% ($\frac{38}{61}$)	22%($\frac{2}{9}$)
	Tree	54%($\frac{40}{74}$)	51%($\frac{29}{56}$)	66%($\frac{12}{18}$)	63%($\frac{44}{70}$)	67% ($\frac{40}{61}$)	22%($\frac{2}{9}$)
	Greedy	67%($\frac{50}{74}$)	66% ($\frac{37}{56}$)	77%($\frac{14}{18}$)	55%($\frac{39}{70}$)	57% ($\frac{2}{61}$)	33%($\frac{3}{9}$)
	PC	50%($\frac{37}{74}$)	46%($\frac{24}{56}$)	77%($\frac{14}{18}$)	51%($\frac{36}{70}$)	53% ($\frac{32}{61}$)	22%($\frac{2}{9}$)
	Bayesian Search	64%($\frac{48}{74}$)	64%($\frac{35}{56}$)	66%($\frac{12}{18}$)	55%($\frac{44}{70}$)	57% ($\frac{35}{61}$)	33%($\frac{3}{9}$)
	Structure 7	66%($\frac{49}{74}$)	66% ($\frac{37}{56}$)	66%($\frac{12}{18}$)	65%($\frac{46}{70}$)	69% ($\frac{42}{61}$)	22%($\frac{2}{9}$)
	Structure 8	49%($\frac{36}{74}$)	46%($\frac{26}{56}$)	66%($\frac{12}{18}$)	63%($\frac{44}{70}$)	66% ($\frac{41}{61}$)	22%($\frac{2}{9}$)
	Structure 9	32%($\frac{24}{74}$)	25%($\frac{14}{52}$)	77%($\frac{14}{18}$)	13%($\frac{10}{70}$)	0% ($\frac{0}{61}$)	100%($\frac{9}{9}$)

In order to compare the accuracy result of BBN models with a simpler linear method, logistic regression model was developed with the same balanced data of the case study (using Sklearn library of python 3.9). This model classified the presence and absence of *Cryptosporidium* based on the linear function (Bishop, 2006). Logistic regression model trained with generated data by ADASYN and SMOTE resulted in the same overall accuracy of 60%, and accuracy of 55% and 72% for prediction of absence and presence of *Cryptosporidium* respectively. Achieving the same accuracy and performance regardless of how data was generated can be due to the reason that the model was not very sensitive to the difference of generated data. In addition, comparing the result indicates that while BBNs were able to promote the accuracy of *Cryptosporidium* presence to 88% in first generated data set, logistic regression resulted in 72% accuracy. Similarly, BBNs achieved 66% accuracy in predicting absence of *Cryptosporidium* while logistic regression indicated slightly weaker performance by 55% of accuracy. Although the change is not very large it should be considered that in the case of using generated data of second scenario (dataset balanced by excluding test set prior to data generation) by ADASYN, logistic regression resulted in only 10% accuracy while BBNs by using the same data set outperformed logistic regression. BBNs seems to outperform classical linear methods not only because of more accurate results, but also because they allow to assess different assumptions about the parameters relationship. While the volatile interaction between environmental factors can hardly be captured by linear models.

For instance, different structures were defined to assess the interaction of variables. Comparing the accuracy of the group of datasets, the developed structure did not show a considerable impact on the model performance. Furthermore, the performance of the models with specific structures was none consistent for each dataset. For instance, Naïve Bayes has resulted in 69% accuracy for the ADA-BBN model while considering SMOTE-BBN models with the same structures, the accuracy dropped to less than 60%. This observation indicated that the best model structure depends on the nature of data and any given structure learning algorithm is unlikely to be universally the best method for varying datasets.

The same observation was also seen in structures developed by expert knowledge. For instance, although Structure 8 showed a better performance than structure 7 for models developed by the ADA-BBN dataset, this model has shown less accurate results than structure 7 for models developed by SMOTE-BBN. However, structure 9 consistently performed poorly. Structure 9 has included only one parent node, fecal coliform, connected to the dependent

variable, *Cryptosporidium*. This approach was to consider the impacts of precipitation and turbidity on fecal coliforms and use fecal coliform counts to inform *Cryptosporidium* levels. This parent-to-child relationship implies that *Cryptosporidium* levels are independent of turbidity and precipitation, given fecal coliform counts. This observation denoted that the measures of fecal coliform counts cannot adequately predict *Cryptosporidium* alone, and regulations or guidelines that rely on only indicator bacteria to evaluate the microbial quality of source water could be misleading. It is also well established that indicator bacteria such as fecal coliforms and *E. coli* do not have a consistent and well-defined relationship with protozoa levels in the source waters (Lalancette et al., 2014b).

The probability of *Cryptosporidium* presence and how it reacts to variations in turbidity, fecal coliforms, and precipitation can be inferred from the BBN (Figure 3.9). The probability output of BBN was concurrent with the pattern of turbidity and fecal coliform. However, the output was also modulated lower during periods of low rainfall, which indicates that the model has been making the decision by considering all parameters, not only fecal coliform levels. For instance, the highest probability of presence was achieved in the day within the maximum level of experienced turbidity in the observed dataset. Also, Figure 3.10 shows that the probability was rarely increased for the days with low precipitation, while during the second half of 2018, the model has been estimating a higher probability for the presence of *Cryptosporidium* when the fecal coliform level was not elevated. As such 70% of presence prediction (more than 50% probability of presence) was for the dates that had at least 1 mm precipitation over three days, while only 46% of the presence predictions (more than 50% probability of presence) was for the days with fecal coliform presence. Furthermore, this concurrent increase in the probability of presence and precipitation level highlights *Cryptosporidium*'s dependency on rainfall or weather events. Therefore, in the case of extreme weather events considering a wide range of characteristics instead of relying on only fecal coliform seem more logical. However, it should be mentioned that the observation here is based on a historical and small record of a source water dataset. While the studied reservoir is designed by considering climate change scenarios and is developed in such a way to tolerate/alleviate its' impact. Therefore, it is possible that other normal water sources that experience a more severe impact of weather conditions have more microbial contamination or experience more variation on quality characteristics.

The approach of separating validation datasets before balancing has been shown to impact model accuracy. The main impact of this approach was reflected in the capability of the model in predicting the presence of *Cryptosporidium*. As previously mentioned, two different

approaches were used to decide which portion of real measurements and which combination of minor and major samples can be used to balance the data. Results in Table 3.3 demonstrate that the highest accuracy for the prediction of *Cryptosporidium* presence can be achieved by balancing the data with all samples in the minority and majority classes (Table 3.3 a). Although only real samples were used to test the model, data used in the training set has influenced the test set since the test data were used to synthesize training samples. Therefore, an alternative approach was used to alleviate possible issues with test set bias, where 30% of minority and majority samples are kept intact for validating the model and have never been used for data balancing (Table 3.3 b). Results using this second method showed that the accuracy stays high in predicting the absence of *Cryptosporidium* but is reduced for presence prediction. A reduction in performance could result from the lower number of samples in the minority class used for generating the data (only 9 samples). In this case, the model was being tested on a relatively smaller set of true test samples. The increase in performance when using all minority samples to generate data may be due to introducing carry-over influence between the test set and training set and/or the improved generation of synthetic samples.

In the case of having all minority classes included, the algorithm will consider the full range of parameters in generating the new samples. However, once using only 50% of samples, it is not possible to reflect the full range of parameters. For instance, the range of precipitation was between 0 and 24.5 in the dataset including all samples from minority classes, and the range of the same parameter in the dataset with 50% of samples was between 0 and 13. Therefore, increasing the number of presence samples to build synthetic data can result in the representation of a broader range of possible conditions such as heavy rainfall or severe microbial contamination.

Although the ADASYN algorithm generates data preferentially in low-density regions or at decision boundaries and was expected to result in a more realistic decision boundary (He et al., 2008), the observation here indicates that with regards to predicting *Cryptosporidium*, none of the pre-processing methods considerably outperforms the other one. The results do however show that among all models the three structures/models that were trained and tested by the data generated with the ADASYN algorithm could predict the presence of *Cryptosporidium* with ≥ 30% accuracy while only two models developed by SMOTE had comparable performance. Also, the highest accuracy in predicting *Cryptosporidium* presence was observed using ADASYN (accuracy of presence: 88%), showing a minor improvement over SMOTE (77%). Still, these observation does not imply that ADASYN can consistently be more favorable than SMOTE pre-processing method.

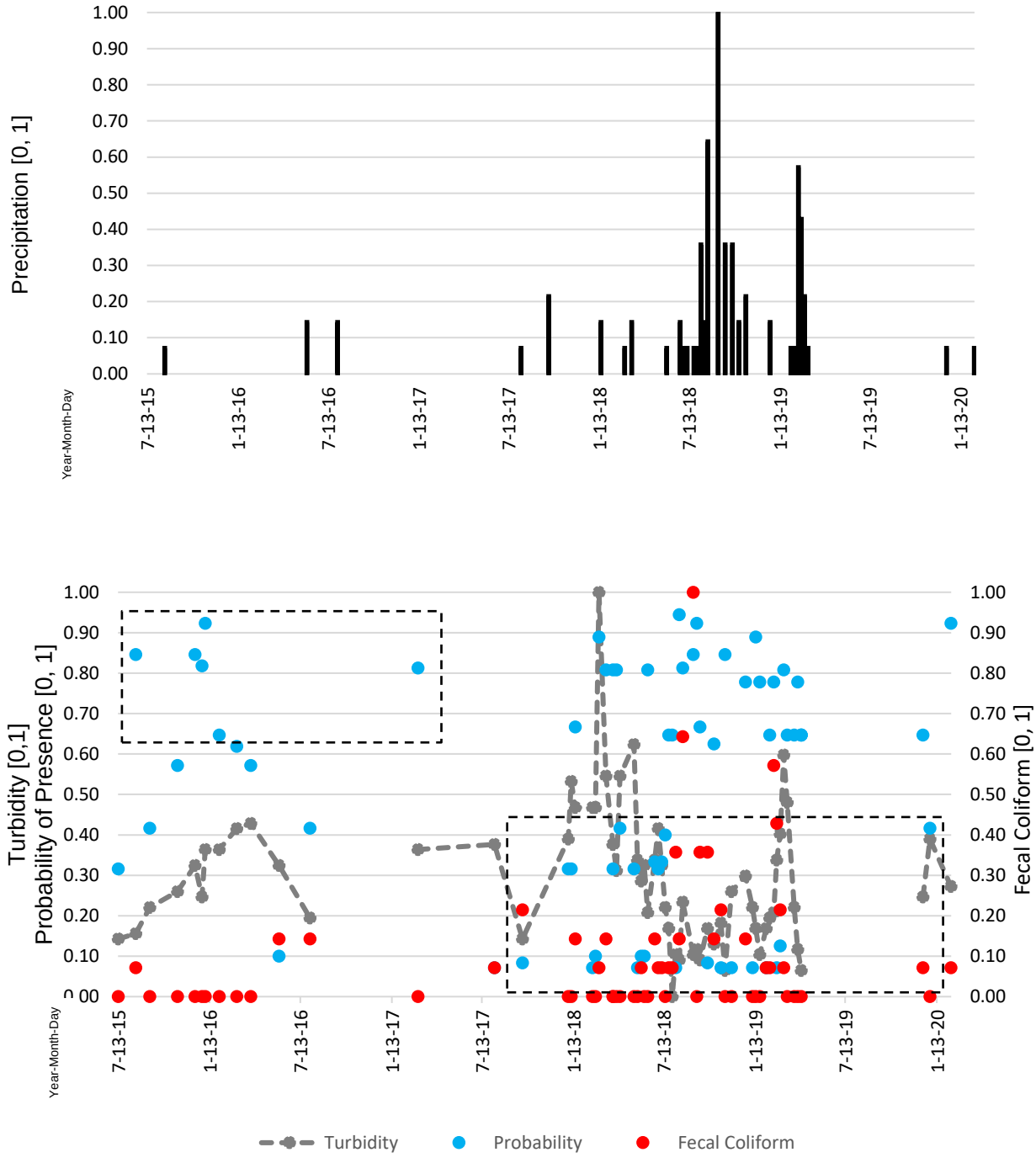


Figure 3.9 Probability distribution of *Cryptosporidium* presence and rainfall level versus normalized Turbidity and Fecal Coliform. The graph is based on structure 8 using data balanced by SMOTE. The probabilities indicated in the box denotes the correct predictions made by the model.

The observation of comparing the accuracy results of the model developed by data generated with SMOTE and ADASYN with different structures indicates that the efficiency of the employed balancing method or a specific structure is highly dependent on the employed dataset. It was observed that both SMOTE and ADASYN contributed to the improvement of the model performance, however, the degree of this improvement for both SMOTE and ADASYN was different for the dataset that contained all samples from minority classes and the ones including only 50% of the samples in this class. These findings are also aligned with the observation of (Chawla et al., 2002), where improvements to the model performance using SMOTE are highly dependent on the dataset and learning model applied. Therefore, it can be concluded that although the application of balancing algorithms can add to model performance, the most proper method should be selected based on the type/availability of the data intended to be studied. The specific pre-processing method does not seem to be consistently the best for all datasets because the ratio of unbalancing and the way that data is scattered can impact their function.

Furthermore, a specific balancing algorithm cannot consistently outperform the other algorithms for all classification/prediction models. For instance, a study conducted by Brandt and Lanzen (2020) has studied the performance of SMOTE and ADASYN techniques and compared their superiority considering different classification models. However, the study concludes that although both SMOTE and ADASYN improve the classification performance, none consistently outperformed the other for all models. For example, the study indicates that although SMOTE increases the classification accuracy of the SVM model, ADASYN contributes to better performance in developing random forests (Brandt & Lanzén, 2021). The focus of this study was to explore the capability of balancing algorithms in improving BBNs in predicting *Cryptosporidium* and in-depth investigation of how to choose the best pre-processing algorithms based on the dataset and for the desired models can be planned for future work by comparing different AI-based models.

3.4 Summary

The persistent challenges in the real-time measurement of *Cryptosporidium* have resulted in a lack of available data for developing prediction models for this pathogen. Two commonly applied data balancing algorithms, ADASYN, and SMOTE methods were developed to generate new samples based on observed real data. The BBN method was utilized to predict the level of *Cryptosporidium* in the Kensico reservoir. The model was trained by generated data and tested over the real measurements. The results indicated the capability of BBN approach in predicting

the *Cryptosporidium* and also denoted the efficiency of both ADASYN and SMOTE algorithms in lifting the performance of the BBN model. Coupled application of BBN and data balancing methods was indicated to address the time and cost challenges of indirect measurement of *Cryptosporidium* by predicting its absence of presence with more than 60% accuracy.

In addition, developing distinct structures of BBN besides analyzing the correlation coefficient between variables and the probabilistic output of BBN indicated the dependency of *Cryptosporidium* to temperature and precipitation as well as turbidity and fecal coliform. The resulted poor performance of the structure with the only connection between *Cryptosporidium* and fecal coliform denoted that the assumption of using fecal indicator bacteria as the surrogate for other pathogens measurement can be misleading and should be used with caution.

Chapter 4: Prediction of *E. coli* using Bayesian Networks and Balancing Algorithms

4.1 Introduction

The rapid changes in pathogen levels in source waters complicate both the real-time assessment of water quality and optimization of the treatment facilities (Sokolova et al., 2022b). *E. coli* is one of these pathogens of concerned recognized as the principal indicator organism in freshwaters (Jamieson et al., 2004). Several regulations consider *E. coli.* as a surrogate for other pathogens such as *Cryptosporidium* because although a low dosage of this bacteria cannot cause infection in the human body, the presence of *E. coli* indicates contamination of water to fecal pollution. At the same time, a high concentration of *E. coli* in the water supply is capable of posing serious risks to public health. During the history, the presence of *E. coli* was the source of several worldwide outbreaks (such as the Wyoming outbreak during 1998, the Swaziland outbreak during 1992 in South Africa or the *E. coli* outbreak in Grampian in Scotland during 1990 and the Walkerton outbreak in Ontario during 200).

Therefore, *E. coli* is considered one of the main microbial quality parameters required to be monitored. However, the obstacle in real-time assessment of *E. coli* level is the needed time to measure this bacteria because measurement of *E. coli* can take more than 24 hours. While its concentration can exceed the standard limit during this time gap. In order to overcome this time delay, data-driven methods can provide a real-time prediction of *E. coli* based on historical observation and recorded data. Furthermore, AI-based methods can provide an understanding of *E. coli* sensitivity to other weather and water quality parameters. In this section, similar to the previous chapter, the BBN method has been applied for predicting *E. coli*. BBN was used to capture the uncertainty of different environmental factors that are involved and can impact the pathogen level in the water. This section includes the implemented steps for modelling *E. coli* and the discussion of the observed results.

4.1 Material and Methods

4.2.1 Data Preparation and Discretization

In this chapter, the case studies were three monitoring sites located in British Columbia, Cheakamus, Salmon and Peace River. The data was selected based on the availability of *E. coli* records and consistent measurement of other water quality parameters. The water quality data for these three case studies in Canada was obtained from (*Environment and Climate Change Canada*, 2019).

The Cheakamus River is located in Daisy Lake Forest in British Columbia. This River is a tributary of the Squamish River and initiates from Cheakamus lake and enters Daisy Lake dam in the Whistler area (BC Hydro, 2012). The sampling site (BC08GA0010) is reported to be 200m upstream of Daisy Lake Forest Road bridge, with a latitude of 5005890 and longitude of -123.09678. One of the flows upstream of Cheakamus river is the Callaghan creek that enters 5.5 km upstream from Daisy Lake. Similarly, the whistler wastewater effluent discharge, 100 m above Millar Creek, enters Cheakamus River, which can possibly impact quality parameters. The weather data for this case study was also achieved from Callaghan Valley (Climate Id: 1101300) as the closest weather station.

Salmon River (BC08LE0004), located in the Shuswap region in British Columbia, has a 120 in km length and drains approximately 1500 km^2 area with an average of 1.17×108 m^3 water flow per year. The River extends from the Salmon watershed into the Salmon Arm and Shuswap Lake with a latitude of 50.6926 and longitude of -119.3304. The area is covered with a combination of mainly agricultural and forest lands, and the main activities besides agricultural activities are cattle and poultry farms and forage production. Also, one of the features of this watershed is the air barrier created by Coast, Cascade and Columbia mountains between Kelowna and Kamplooms city that prevents the flowing air from Pacific ocean and impacts the climatic condition of the region (Zhu et al., 2012). Similarly, weather parameters were obtained from the closest station of the case study, Salmon Arm Cs. (Climate ID: 116FRMN).

The last case study is Peace River originated in the borders of British Columbia and Alberta with a latitude of 56.1261. and longitude of -120.0564. The River flows into the Peace-Athabasca Delta from Willison Reservoir in northeast British Columbia and Bennett Dam draining an area of 118 000 km^2. The sampling location was above Alces River (BC07FD005), one of Peace River's tributary. The sewage discharge of the City of Fort St. John close to the monitoring station can possibly impact the water quality parameters in the region. Besides, the agricultural and irrigation activities can also affect this River's water, which supplies the drinking water for Taylor and Hudson's Hope (*Water Quality Assessment of Peace River Above Alces River*, 2003). The weather data for this case study was also achieved from Peace River A. station (Climate id of 3075041).

The data for each three location were used for developing BBN model using GeNIe software (Bayesfusion LLC, Pittsburgh,PA). Also, before developing the BBN models, the following steps were employed to prepare the models; Step 1) Excel Microsoft software was used to clean and analyze the dataset of each location and the date without the complete record of the

desired parameters was removed from the dataset. In addition, to prevent the loss of information and keep more samples of *E. coli* some parameters that only had a smaller number of samples for a shorter period were also excluded from the dataset. It is worth mentioning that *E. coli* is a specific species of fecal contaminates and are normally used as an alternative pathogen for the measurement of fecal coliform. However, it was assumed that considering a wider diversity of parameters (including fecal coliform information) from available data can result in more comprehensive models that allow for investigating their relationship and evaluate the efficiency of considering only fecal coliform or *E. coli* as the fecal contamination indicator. Finally, besides the *E. coli* record, turbidity, pH, hardness, fecal coliform, and weather parameters including temperature, precipitation on sampling day and over three days formed the datasets for this chapter. The *E. coli* and fecal coliform were reported based on CFU/100 mL in all of three locations. The number of samples and recorded dates for each location are presented in Table 4.1.

Table 4.1 Details of the utilized dataset for *E. coli* Prediction

Location/Data Source	Total number of samples	Recorded Dates
Cheakamus River	349 Samples	2004-2021
Salmon River	406 Samples	2000-2021
Peace River	290 Samples	2000-2021

Step 2) Before training the model with the dataset, each of the variables was discretized into different classes. Although some information can be lost in the case of discretizing the variables, developing BBN with continuous data requires assuming continuous probability distribution, which is unavailable in current BBN software due to complicated the computation of probability tables complicated and the drop in speeds of calculation (Nojavan A. et al., 2017; Panidhapu et al., 2020b). Furthermore, the objective of this chapter was a binary prediction of *E. coli*. Therefore, the measured and reported levels of *E. coli* were divided into two classes. The threshold of this classification was chosen 20 CFU *E. coli*/100 ml due to the filtration deferral regulations in British Columbia (B.C.Ministry of Health, 2012). The days with less than 20 CFU *E. coli*/100 ml were considered the absence class, and the days having above the 20 thresholds implied the presence of *E. coli*. Based on this policy, the source water monitoring station can differ the filtration process if 90% of weekly samples report less than 20 CFU *E. coli*/100 ml over six months (Panidhapu et

al., 2020b). All other parameters were discretized in a way to have an equal number of data in each bin. The parameters were discretized into 3 bins, except precipitation on the day and over three days. These two parameters were discretized to two bins because recorded data were mainly zero for most of the samples, and it was aimed to have an equal number of samples in each bin; Step 3) discretized data for each monitoring site were used for training the structures that were developed based on the following two sections. Each parameters' data in training set were assigned to the appropriate node and model was trained based on discretized training data set. The performance of the model was assessed using testing dataset and comparing the number of accurate predictions to the measured values.

4.2.2 Relationship between Water quality, Weather and *E. coli*

One of the main challenges in understanding both water quality or hydrological phenomena is that these environmental systems are impacted by several stochastic processes (Stegen et al., 2012). Sensitivity of these models to the wide range of parameters such as population growth, land cover or economic development has made it difficult to anticipate future changes (Taheri Tizro et al., 2014). Therefore, prior to developing BBN models for *E. coli* prediction, the time series of available data for all three locations were illustrated to assess the determinant trends or long - term dynamics in last 4 years. As an example, the spread of all parameters in last 4 years in Salmon River are depicted in Figure 4.1.

While in previous section both of *Cryptosporidium* and fecal coliform as the microbial characteristics of water, lacked any meaningful time specific changes. Figure 4.1 indicates that there is a seasonal trend in *E. coli* and fecal coliform concentration, with a promoted *E. coli* concentration in summer and autumn and decrease in winter and spring time. This observation also aligned with the result of a study conducted by Oliver and Page (2016) which shows the peak of *E. coli* in both fall and summer due to optimal weather conditions. For instance, the higher temperature in summer and "rehydration" from precipitation can promote the fecal contaminant concentration (Oliver & Page, 2016).

The reason that this seasonality was not observed in time series of *Cryptosporidium* in previous section can be due to the high number of "non-detects" and zero value of *Cryptosporidium*. Including mainly the absence of a parameter in a time series makes it difficult to extract meaningful trends over time. The observed seasonal flocculation in temperature, hardness and turbidity can be identified here as well. Also, an increasing trend can be seen from

the precipitation data that can be explained by the impacts of climate changes and the reported promoting rainfall over the time.

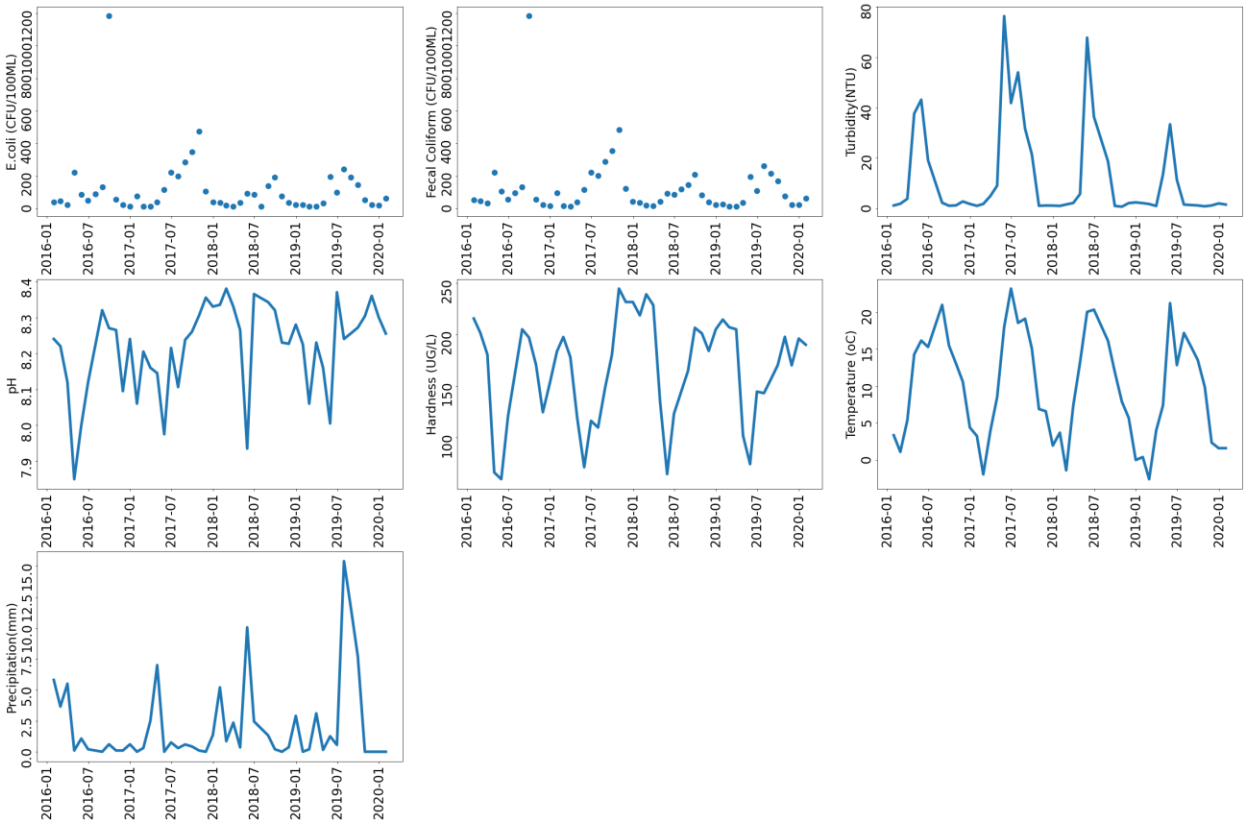


Figure 4.1 Visualization of the time series for weather and water quality parameters of Salmon River. The time interval is equal (one month) for all parameters. The time series for Cheakamus River and Peace River can be found in appendix.

Similar to the section 3, the linear correlation between the earlier value and present value of parameters were analyzed through ACF graphs for Salmon River. As it can be seen from Figure 4.2 while there is a repetitive pattern between observed turbidity, hardness, and pH in one month, the value of *E. coli* was not affected by previous values. However, Dwivedi et al. (2016) have indicated a correlation between *E. coli* in one, 7 and 10 months (Dwivedi et al., 2016). Although the repeating patter was also observed in over 6 months in the simple time series graph (Figure 4.1), it should be considered that ACF measures the linear correlation between parameters value in each time lag, and due to the complexity of microbial activities, linear analysis fail to reflect the dynamic of the system. However, comparing the result of ACF for each location indicates that the behavior of pathogens such as *E. coli* can be highly site specific. For instance, the correlation between current *E. coli* and its concentration in one month earlier, were more than 0.5 in Peace

and Cheakamus Rivers. This observation again emphasizes the dependency of microbial communities to other site-specific parameters such as landcover, climatic conditions or human activates around the area.

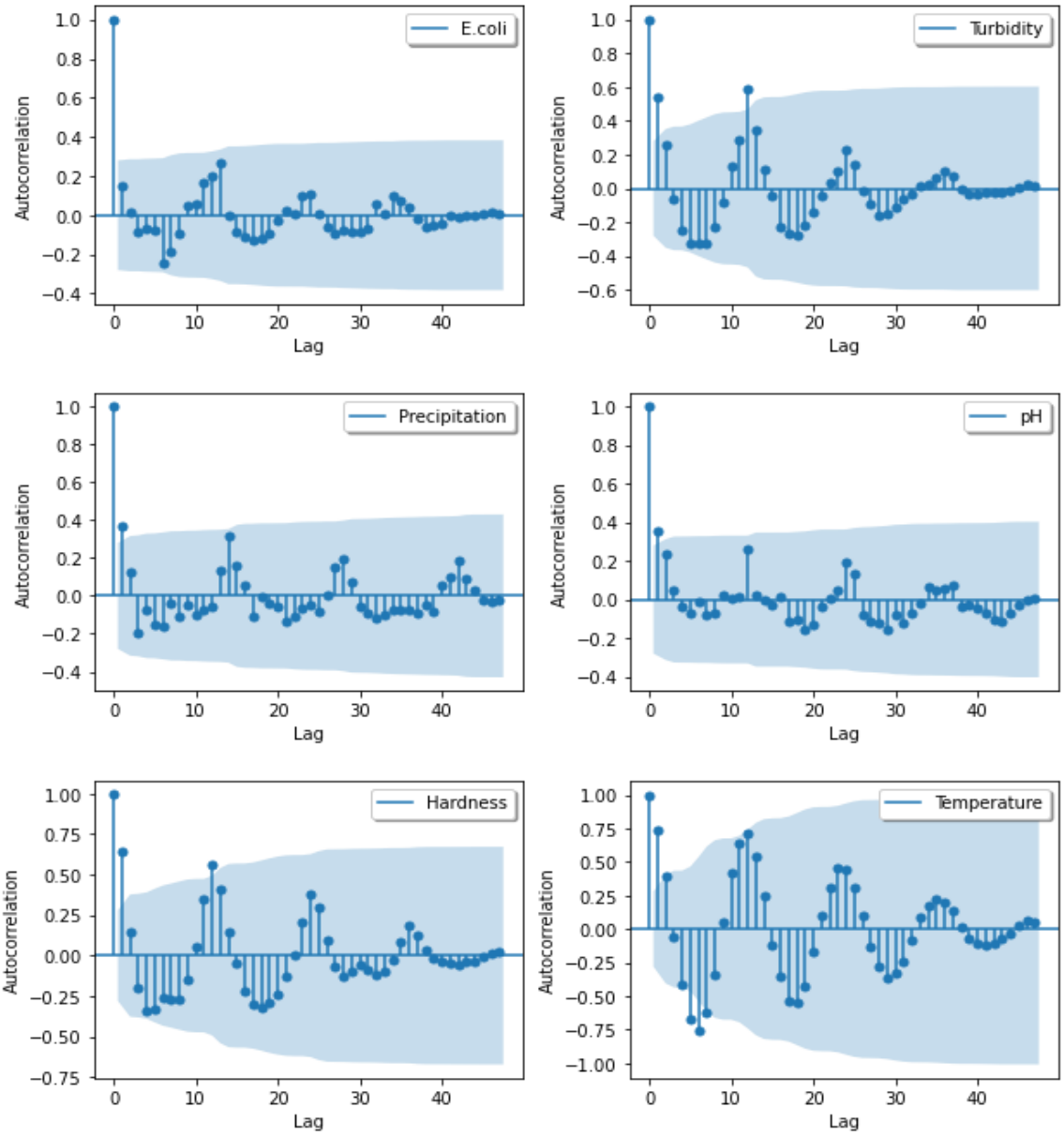
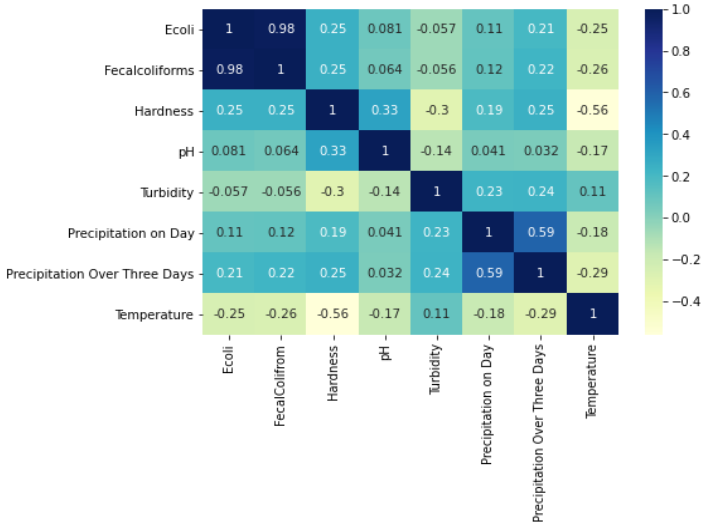


Figure 4.2 Autocorrelation of all parameters based on monthly lag and confidence interval of 95% for Salmon River. The autocorrelation graph of Peace and Cheakamus River can be found in Appendix.

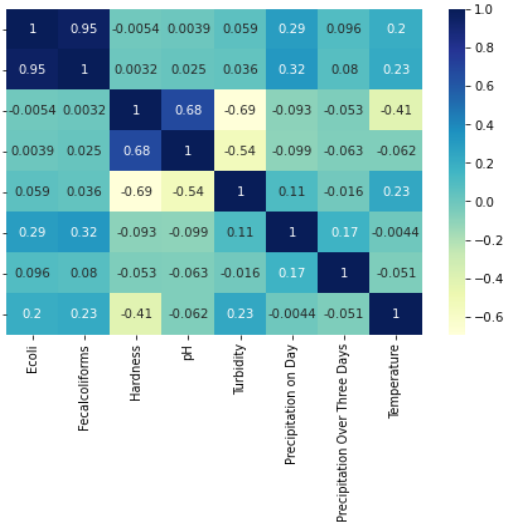
In order to see dependencies between parameters, the linear correlations between Cheakamus, Peace and Salmon River variables were analyzed. While BBN models were based on the discretized class of variables, the correlation coefficient has been calculated for each

parameter based on the continuous data. Figure 4.3 indicates the correlation matrix for the three monitored locations. As was expected, the results show a very strong correlation between *E. coli* and fecal coliform for all three sites (>80%). However, none of the other correlations between water quality parameters and *E. coli* were consistent for each location. For instance, the correlation of turbidity with *E. coli* or fecal coliform ranged from -0.06 to 0.33 for Cheakamus and Peace River, respectively. Similarly, the correlation between hardness and turbidity was strong for the Salmon River (-0.69), while it was very poor (0.04) for Peace River.

(a)



(b)



(c)

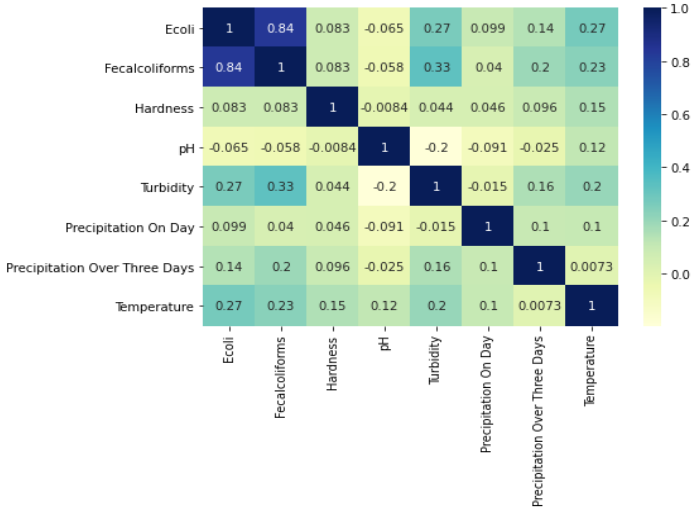


Figure 4.3 Correlation coefficients (R) of water quality and weather parameters in the a) Cheakamus River, b) Salmon River, c) Peace River.

Figure 4.3 Correlation coefficients (R) of water quality and weather parameters in the a) Cheakamus River, b) Salmon River, c) Peace River.

Other variables were observed to have a variable positive and negative correlation with *E. coli* and fecal coliform. The observation in this study aligns with investigations in the study by (Panidhapu et al., 2020b), which reported non-consistent correlations for similar monitoring sites in British Columbia. These site-specific results have indicated that several parameters such as climatic patterns or land usage in the region can contribute to the dynamics of the water quality.

For instance, while a strong positive correlation between temperature and *E. coli* was expected because of the possible increase in the activities of microorganisms (Cha et al., 2016), a moderate positive correlation (0.20 – 0.27) was observed for Peace and Salmon River and a moderate negative (-0.25) for Cheakamus River. When the surrounding area of the Peace and Salmon River is agricultural, and the Cheakamus River is mainly covered by forest.

Despite the site-specific variation of other parameters, the correlation between fecal coliform and precipitation (both on the sampling day and over three days) was positive for all three sites, with differing strengths. For example, for the Salmon River, the correlation of precipitation on the day and fecal coliform was 0.32, which is relatively strong considering the data being analyzed. The positive correlation between fecal coliform and precipitation was also noted in the analysis of *Cryptosporidium* data. Therefore, it can be included that there is an underlying relationship between the characteristics of weather events.

Still, due to the uncertainty and complexity of climatic systems and the interaction between pathogens in a watershed, their relationships cannot be reflected thoroughly by simplistic analysis, such as linear correlations.

4.2.3 Structure Development of BBN for *E. coli* Prediction

BBN structures can be developed based on *prior* and expert knowledge or defined structure development algorithms. While both approaches have been widely applied and have been reported to result in accurate models, there is still a debate on the best approach for defining the BBN structures (Scutari et al., 2019). The focus of this chapter was to evaluate the *E. coli* level as a microbial quality indicator and explore the capability of BBN models in predicting quality indicators of source water. A more in-depth investigation on how the structure developments can impact the prediction capability of BBN models is another field of study that can be of future interest.

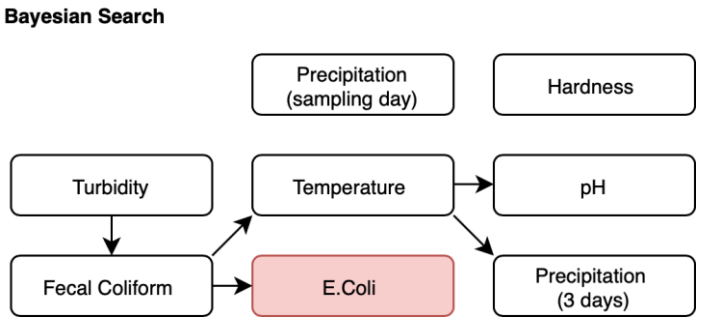


Figure 4.4 Learned BBN structures based on Bayesian Search structure-learning algorithm.

Therefore, in this work, similar to the models developed for predicting *Cryptosporidium*, both algorithm-based structures besides expert knowledge were developed. Although structure learning algorithms resulted in topologies that were logical and reflected the expected relationships based on expert knowledge for *Cryptosporidium*, these algorithms produced unexpected structures when built based on the *E. coli* dataset. Furthermore, the same algorithms excluded some of the weather and water quality variables, and this exclusion was observed in all three employed datasets. For instance, Figure 4.4 indicates a sample structure developed by Bayesian search and shows that the precipitation on day and Hardness are excluded. Also, pH and precipitation over three days are in no connection with *E. coli*.

Therefore, the structures for predicting *E. coli* were developed only based on *prior* knowledge and similar structures that have been reported to have a good performance in previous studies (Panidhapu et al., 2020a). As it can be observed from Figure 4.5, two other structures were developed for *E. coli*. In the first structure, (shown in Figure 4.5 a), the connection between weather parameters and *E. coli* has been defined indirectly and through turbidity. While in the second structure (Figure 4.5 b) a direct connection between all of the considered parameters and target variable have been defined. This approach has been employed to see if the weather condition directly contributes to the microbial concentration of source water or their main impact is on other characteristics of water such as turbidity and hardness, while the variation on these characteristics provides an opportunity for microorganism's growth or activity.

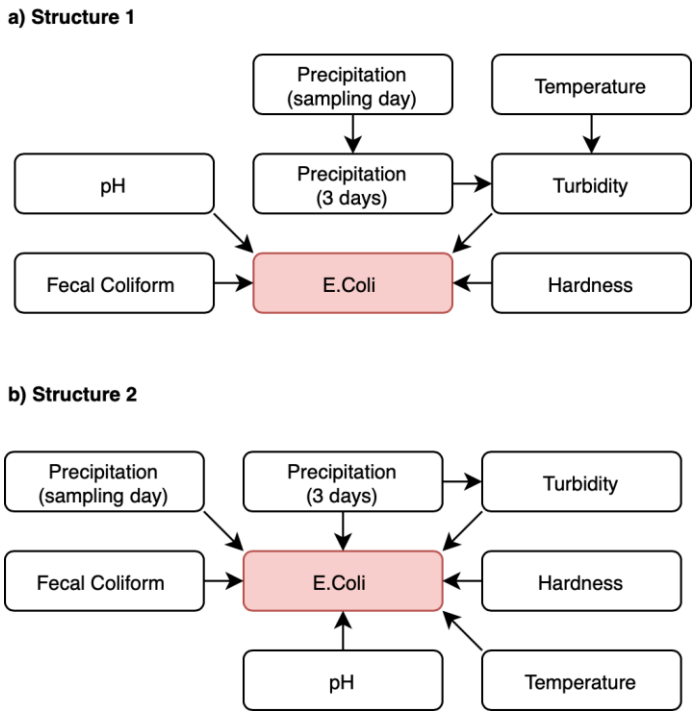


Figure 4.5 Learned BBN structures for prediction of *E. coli* based on expert knowledge.

4.3 Result and Discussion

4.3.1 Prediction of *E. coli* using Bayesian Networks

The previous chapter indicated the capability of BBNs in predicting *Cryptosporidium* presence/absence in a drinking water reservoir. This section similarly evaluates the capability of BBNs in predicting *E. coli* in three different monitoring sites (Cheakamus, Salmon, and Peace River) in BC, Canada.

Similar to modelling *Cryptosporidium*, for predicting *E. coli*, both weather and water quality parameters were included and discretized based on details in section 4.2.1. The sites were selected based on the availability of both weather and water quality data as well as *E. coli* records for the considered period. The data was used to train the BBN models that were developed based on the shown structures in Figure 4.5 and section 4.2.3. The employed structures were developed based on expert knowledge and logical relationships between parameters. As mentioned earlier, two structures have been used for all model developments to compare the results for each dataset and location. Table 4.2 indicates the result of prediction models for each site. Besides, BBNs the

same data were used for developing with a simple logistic regression model. While the overall accuracy of model performance was higher than 50% for all structures and locations, using BBN prediction model, simple models like logistic regression failed to predict the absence of *E. coli*. For instance, despite BBN which resulted in more than 70% accuracy in predicting both *E. coli* presence and absence, logistic regression model resulted in less than 10 % accuracy for predicting *E. coli* absence (<20 CFU/100 L). Logistic regression resulted in an acceptable performance only for Peace River with overall prediction accuracy of 88%, and accuracy of 92% and 50% for predicting *E. coli* higher and lower than >20 CFU/100 mL respectively. However, BBN was successful in achieving accurate predictions not only for Peace River, but also for all monitoring sites regardless of their data and location difference. This observation indicates that the complexity of interaction between environmental parameters and microbial activities cannot be captured by linear models in all instances. BBNs not only seems to be stronger in reflecting such dynamic systems but also allows for assessing possible interactions.

Regarding the optimum structure, it can be seen form Table 4.2 that structure 1 resulted in a better performance for all three datasets than structure 2. Structure 1 consistently resulted in the prediction of presence accuracies greater than 50%; however, there was a decrease in prediction accuracy of presence with structure 2 (3% - 11% decrease). Overall, it was concluded that structure 1 is a better approach to developing BBN models for *E. coli* prediction since the improvements in presence prediction (>20 CFU/100 L), were more significant compared to the minor deterioration in the accuracy of absence prediction.

As described in the previous section, structure 1 (indicated in Figure 4.5 a) has included the indirect connection between precipitation on day/ precipitation over three days prior and *E. coli*. It was expected that the direct impacts of precipitation on *E. coli* would be more representative. This can be due to the reason that high rainfall intensities can wash out more fecal contaminants from the surrounding landscape. Also, studies have shown the increase in fecal contamination after several stream flows in regions covered by farmlands (Kay et al., 2008; Lyautey et al., 2007) and the agricultural cover was reported to be one of the factors involved in the presence of fecal bacteria (Laurent & Mazumder, 2012). While the Salmon and Peace River were surrounded by agricultural land and likely to be affected by feces and agricultural waste after rainfalls, the result indicated that the model makes a better decision considering the reflection of precipitation on turbidity and hardness and not directly on the target variable.

Table 4.2 Prediction accuracy of *E. coli* with BBN. Prediction results are based on a randomly separated test set (15%) of data.

Model	Location	Structure	Overall accuracy	Prediction Accuracy of *E.coli*>20 CFU/100	Prediction Accuracy of *E.coli*<20 CFU/100
BBN	Cheakamus River	Structure 1	69%	51%	88%
		Structure 2	55%	22%	92%
	Salmon River	Structure 1	72%	75%	70%
		Structure 2	56%	29%	81%
	Peace River	Structure 1	77%	52%	86%
		Structure 2	74%	19%	92%

It was observed in the previous chapter that in terms of predicting *Cryptosporidium*, the model presented lower accuracy for the structures that had the indirect connection between weather parameters and *Cryptosporidium*. Regarding this contrast, it should be noticed that the fecal coliform was the only predictor in predicting *Cryptosporidium* (Figure 3.8 c). In contrast, here for predicting *E. coli* (Figure 4.5 a), all other water quality parameters, besides fecal coliform, were connected to the target parameters. Therefore, it seems that the presence of water quality parameters is crucial in achieving a reliable model for predicting microbial quality. This observation is also aligned with the observation of a previous study conducted by Pandihapu et al. (2020) that reports a 25 % reduced accuracy in the case of excluding water quality parameters in predicting the *E. coli* (Panidhapu et al., 2020a).

It is worth mentioning that the objective of machine learning approaches is minimizing the required time for direct measurement of pathogens and enabling a timely manner risk assessment. The distinct feature of BBNs that can be leveraged for this objective is their capability in using incomplete data. This advantage of BBNs can compensate for the incompleteness in measuring some parameters, which is a persistent challenge in modeling environmental systems. In order to assess this ability of BBNs, and to predict *E. coli* threshold only based on easy-to-measure parameters, the developed models were used for predicting the *E. coli* exceeding the 20 CFU/100mL threshold by using incomplete observation of fecal coliform parameter. The results in Table 4.3 indicates that excluding the fecal coliform impacted the predicting accuracy, which was expected due to the same origin of E. coli and fecal coliform parameters. However, this impact was different for each monitoring sites and structures. For instance, despite the previous

observation in which structure 1 consistently resulted in a higher accuracy, here, structure 2 resulted in a lower reduction in prediction accuracy (less than 10%) compared to the reduced accuracy (>20%) by structure 1.

Table 4.3 Prediction accuracy of *E. coli* with BBN

Model	Location	Structure	Overall Accuracy	Prediction Accuracy of *E.coli*>20 CFU/100	Prediction Accuracy of *E.coli*<20 CFU/100
BBN	Cheakamus River	Structure 1	51%	77%	24%
		Structure 2	55%	72%	40%
	Salmon River	Structure 1	49%	3%	100%
		Structure 2	47%	48%	45%
	Peace River	Structure 1	69%	78%	42%
		Structure 2	67%	81%	23%

Although the accuracy of the *E. coli* absence was observed to be reduced by structure 2, the model's performance in predicting the presence (*E. coli*>20 CFU/100) was improved compared to the models tested with complete data. As shown in Figure 5.3 b structure, 2 contains direct connection from all variables, including weather parameters to the *E. coli*, while structure 1 only had direct connections between water quality parameters. Excluding the observation of fecal coliforms reduced the number of direct predictors in structure 1. Therefore, the limited number of predictors could be responsible for the observed sharp accuracy reduction in this structure compared to the minor accuracy drops in structure 2. Furthermore, the model's sensitivity to each monitoring site can be observed from the accuracy result of structure 2. For instance, the overall accuracy has not been changed for data of Cheakamus River while it has dropped in a similar range (7% and 9%) for Salmon and Peace River, with a similar land cover pattern.

These observations have confirmed the capability of BBNs in making predictions by using incomplete data, which implies the compatibility of this approach in modeling environmental systems with missing observations. Still, the results of using incomplete data for predicting *E. coli* threshold besides the previous observations in this chapter, emphasize the complexity of the relationship between variables and the diversity of factors such as land cover that potentially impact the pathogens' level in source waters. In order to have a better understanding of the impact of each predictor and variable on target parameters of *E. coli* and *Cryptosporidium*, the analysis of parameters relationship has been presented in the next section.

4.4 Summary

E. coli is one of the leading causes of water-borne illness and worldwide outbreaks. The presence of this pathogen is the basis for microbial quality assessment in some regulations. For instance, the source water monitoring policy in British Columbia differs the filtration process if 90% of weekly samples of *E. coli* report less than 20 CFU *E. coli*/100 ml over six months (Panidhapu et al., 2020b). This filtration differation has been a base for this section, and the BBNs method has been developed to predict the *E. coli* absence and presence based on 20 CFU *E. coli*/100 ml threshold in three monitoring sites in Canada. Successful performance of BBN was observed in predicting *E. coli* presence in all three monitoring sites. Although the prediction results were close for all three case studies, the correlation coefficient between employed weather and water quality parameters was observed to be very site-specific and dependent on different land or climatic characteristics of each site. Therefore, it seems broader parameters representing a wide feature of each case study, such as land cover, can provide even more information on the *E. coli* variation in source waters, and the probabilistic feature of BBN can aid in reflecting these involved uncertainties.

Chapter 5: Variable Importance in Assessment of the Microbial Quality of Source Waters

5.1 Introduction

The previous chapters were an attempt to improve the quality assessment of drinking water sources by predicting two important pathogens in a timely manner. However, the persistent challenge in terms of modelling environmental systems is their complexity. Natural processes, such as variation of microorganism's level in source waters, are sensitive to a wide range of factors including climatic parameters. Therefore, instead of facing steady-state processes, environmental systems demonstrate a dynamic behavior that was tried to be captured by the applied models (Rutten et al., 2020).

The observed results in previous chapters indicated that the correlation coefficient between E. *coli* and other considered parameters were different for each case study depending on the features of the locations. Also, it was observed that besides water quality parameters, weather features such as precipitation could impact the model's decision in predicting *Cryptosporidium*. Besides the observation of this study, several other pieces of research have unfolded the dependency of waterborne pathogens such as *E. coli* and *Cryptosporidium* to climatic conditions and other water quality indicators (Atherholt et al., 1998; M. M. M. Islam et al., 2017; Kleinheinz et al., 2010; Young et al., 2015). In this matter, awareness of the factors that potentially impact the level of pathogens and quality indicators can provide a better insight into the steps required to mitigate the possible risks.

Therefore, this chapter was developed to draw a more detailed understanding of each parameter's contribution to *E. coli* or *Cryptosporidium* prediction. This has been done by evaluating their strength of influence on the model's decision and analyzing the model's sensitivity to each of the variables. In addition, the developed probability distributions of BBN models have been interpreted to evaluate the projected variations in microbial quality of water bodies under different scenarios.

5.2 Materials and Methods

5.2.1 Strength of Influence and Sensitivity Analysis

The DAG or BBN structures, as described earlier, include several connections between variables. The strength of variables and sensitivity analysis allows quantifying these connections and relationships between variables (Canova Calori et al., 2007). The strength of influence measures

the distance of conditional probability distributions. More precisely, this method by calculating the distance between the probability distribution of the child node with and without evidence or condition of its parent indicates the magnitude of influence that parent node may have on child node (Koiter, 2006). Therefore, the influence of all predictors on the target node (presence or absence of pathogens) can be quantified, and the most impactful parameters can be identified. GeNIe software allows the calculation of the distance between CPT of nodes using different methods (Euclidean, Hellinger, J-Divergence and CDF methods). In this study Euclidean approach was used to compute the strength of influence of two random variables. Assuming A as one node and random variable with probabilities of a_i $([0,1])$ and similarly B with probabilities of b_i the Euclidean distance can be computed as:

$$Euclidean\ Distance\ (A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad (eq.5.1)$$

Similar to the strength of Influence, the sensitivity analysis is associated with validation of parameters probability and assesses the impact of changes in variables probability on the outcome of interest. The sensitivity analysis computes this impact by calculating the derivative of the target node's posterior probability distributions over each one of the predictors/variables (Chan & Darwiche, 2004). The resulting small derivative for a random variable indicates that the changes in this variable are unlikely to impact the probability distribution of the target parameter. On the other hand, any changes on variables with larger derivatives can result in significant changes in the posterior probabilities (Kjærulff & Van Der Gaag, 2013). Both sensitivity analysis and strength of influence were developed after training prediction models to explore and identify the most critical factors in predicting *E. coli* and *Cryptosporidium*. The criteria and description of the selected models are discussed in the following section.

5.2.2 Selection of BBN Model and Structure

BBN models were developed in the previous chapters to predict *Cryptosporidium* and *E. coli*. Two of these developed BBN models were chosen for evaluating the impact of the considered water quality and weather parameters on predicting the target node. Since each model was different in terms of structures and applied data, the models that resulted in a relatively better performance were selected. Among these chosen models, those structures that have included connections/arc from all of the predictors to the target node were used since these structures could provide an equal opportunity for all variables to be considered in the sensitivity analysis. For instance,

structures 7 and 8 that were developed by data generated with SMOTE algorithm had similar performance in predicting *Cryptosporidium* (Figure 3.2 a and b); however, structure 7 has included all the links/arcs from the predictor to the target parameters. Therefore, this structure was selected for the analysis in this chapter for *Cryptosporidium* analysis. Similarly, Structure 1, indicated in Figure 4.2 a, was chosen for investigating the impact of variables on predicting *E. coli*.

5.3 Result and Discussion

5.3.1 Factors affecting *Cryptosporidium* Assessment

In the previous section, considering both weather and water quality parameters, BBN models could predict the presence of *Cryptosporidium*. To garner insight into how impactful each variable was in predicting the target pathogen, the impact of varying each parameter individually on the probability distribution of *Cryptosporidium* was calculated.

Being DAGs, BBNs allow to investigate the strength of the variable's influence on the *Cryptosporidium* and quantity the sensitivity of the pathogen to each predictor parameter. Both methods can be computed based on the probability distributions of the variables. The strength of influence measures the magnitude of impact that parent node leaves on *Cryptosporidium* by calculating the distance between the probability distribution of the child node with and without evidence or condition of its parent. However, the sensitivity analysis computes this impact by calculating the derivative of the target node's posterior probability distributions over each one of the predictors/variables (Chan & Darwiche, 2004). The results of both strength of influence and sensitivity analysis are presented in Figure 5.1 and Figure 5.2, respectively. As the effects of both methods are in a relative base, the results were normalized between 0 and 1 to easily compare and interpret the rank of all parameters.

It was expected that fecal coliform would show a higher impact as this bacteria's presence can be an indicator of a favorable environment for microorganisms' presence or growth. However, both sensitivity and strength of influence analysis show a poor interconnection between these two pathogens. It should be mentioned that besides some literature research that indicates an existing relationship between indicator organisms and parasites like *Cryptosporidium* (e.g., Lipp et al., 2001; Burnet et al., 2014; Devane et al., 2014; Korajkic et al., 2018;), others have established a poor dependency between these two pathogens. For instance, Hogan et al. (2012) have shown that the presence of FIB can be associated with *Cryptosporidium*'s presence in wetlands in California surrounded by agricultural activities. In contrast, Lemarchand and Lebaron (2003) have

indicated that fecal indicator bacteria cannot be a good representative for *Cryptosporidium* presence on the coastal watershed in France (Lemarchand & Lebaron, 2003).

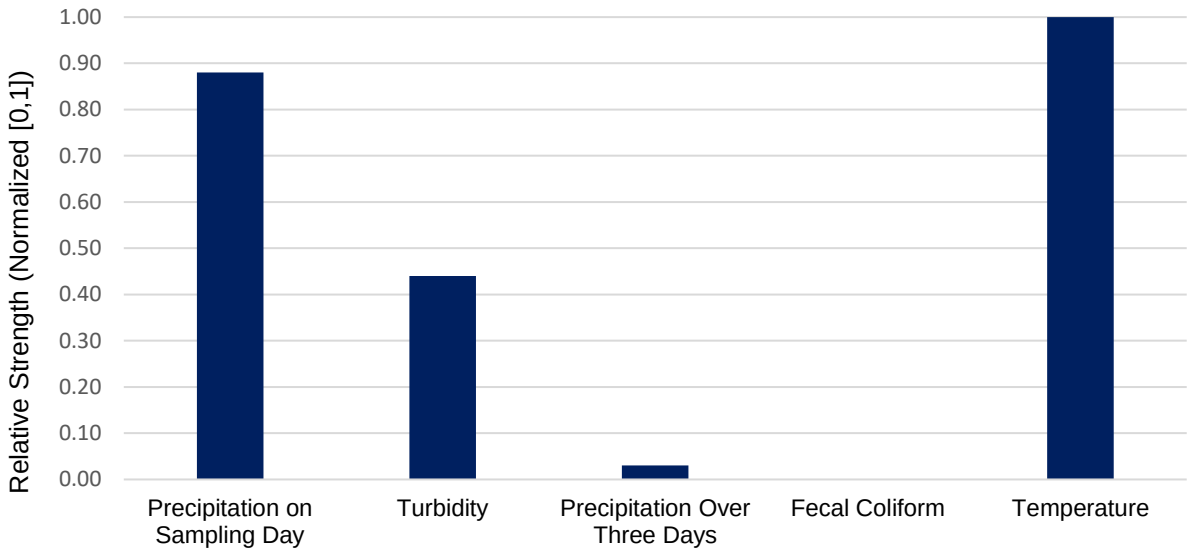


Figure 5.1 Importance of Variables Based on their Strength of Influence in Predicting *Cryptosporidium.*

Therefore, the sensitivity of pathogens' relationship to the considered site or even the season makes it challenging to come to a uniform or inclusive conclusion. Thus, due to the different behavior of the pathogens in environmental systems (also indicated in the study by Wilkes et al. (2009)), the fecal indicator microorganism seems to be a poor surrogate for other pathogen's presence, such as *Cryptosporidium* (Wilkes et al., 2009).

Although the result of both methods was very close and almost the same for all parameters, turbidity was observed to have a higher strength of influence while showing a lower sensitivity. It should be considered that the strength of influence considers two probability distributions for the child node, with and without the evidence of parent node and then measures the distance of these probability distributions. However, the sensitivity analysis validates the variable's probability and investigates the impact of predictors' probability variation on the probability of child node. Sensitivity analysis does this by computing the derivatives of the child nods' probability distribution over the probability distribution of other predictors. This subtle computational difference can be responsible for the distinct observation of turbidity. It was expected that turbidity would have a more substantial influence on the prediction of *Cryptosporidium* since turbidity can provide the required nutrients for microbiological activity

(Lozano et al., 2019; Muylaert et al., 2002). However, the lack of strong connections between turbidity and *Cryptosporidium,* based on influence analysis, aligned with the previously discussed poor linear correlations between turbidity and *Cryptosporidium.*

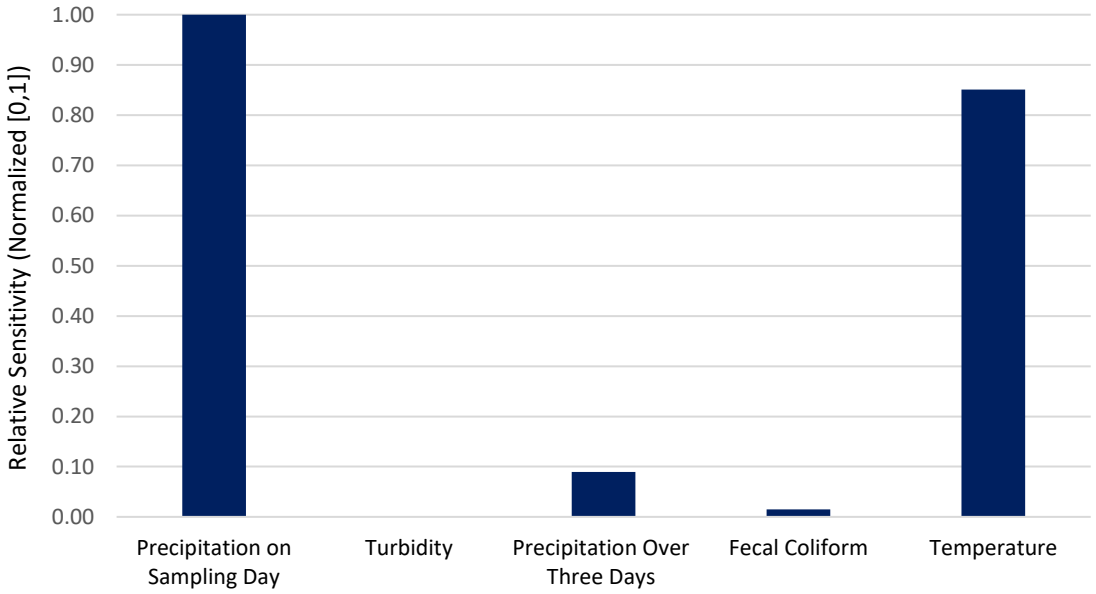


Figure 5.2 Importance of Variables in Predicting *Cryptosporidium* based on Sensitivity Analysis

These observations suggest that regulations and guidelines that solely rely on water quality indicators, such as turbidity, to define the requirement and disinfection processes in the water treatment system (Farrell et al., 2018) can be misleading.

Despite the methodology differences, the temperature and precipitation on sampling day have been observed to leave the greatest influence on the *Cryptosporidium*'s prediction based on both analysis methods. This result reveals that including heavy rainfalls can be a critical parameter in changing the parasite concentration. This finding is also aligned with the literature observation that has reported the increase of pathogens such as *Cryptosporidium* in surface waters after heavy rainfalls (e.g., Duris et al., 2013b; Masina et al., 2018). Therefore, it seems more promising to consider both water quality and weather variables to maximize the prediction accuracy. An observation of interest was that weather parameters (i.e., precipitation, temperature) were shown to have relatively high and similar levels of impact on *Cryptosporidium* levels, despite previous observations that linear correlations with *Cryptosporidium* were weak ($R < 0.5$). Thus, the relationship of the weather or water quality parameters with microbial factors is unlikely to be

linear, and the influence of several factors in combination likely drives a higher or lower probability of protozoa presence.

5.3.2 Factors affecting E. *coli* Assessment

Similar to the previous section, the strength of influence for the considered connections/arcs between predictors and *E. coli* as the target variable was investigated for each case study (Figure 5.3). In order to compare the results from a different perspective, the sensitivity of the *E. coli* to each of the variables was also depicted in Figure 5.4. The output of both strengths of influence and sensitivity analysis were normalized in a location-based between 0 and 1 (parameter's influence were ranked and considered for each location separately).

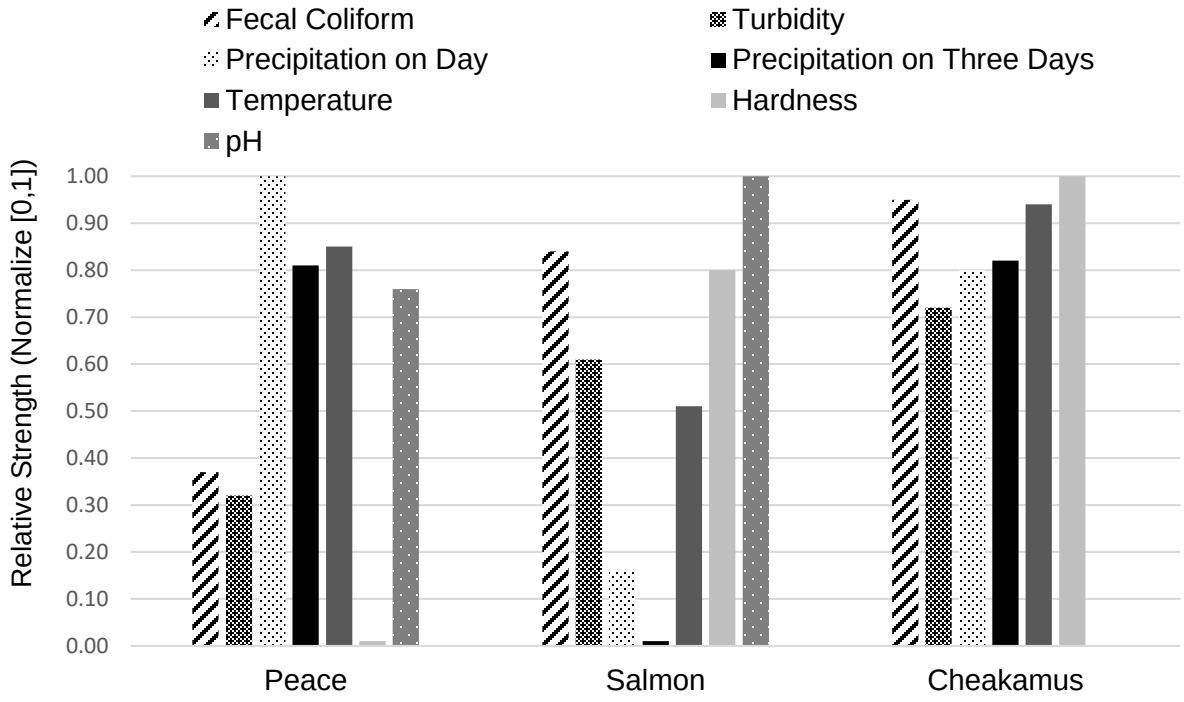


Figure 5.3 Importance of Variables Based on their Strength of Influence in Predicting *E. coli* for three surface water bodies in BC.

Similar to the observation of parameter's impact on *Cryptosporidium* prediction, weather parameters appeared to considerably influence *E. coli* concentrations based on both strengths of influence and sensitivity analysis shown in Figure 5.3 and Figure 5.4, respectively. Especially for the Peace and Cheakamus Rivers, precipitation on the day showed a very high impact on *E. coli* concentration. Similarly, precipitation over three days and temperature as weather variables were

observed to strongly influence the target variable in these locations. However, the result denoted that the variables' impact is specific to each monitoring site, as observed for linear correlation too.

As indicated in Figure 5.3 and Figure 5.4, precipitation showed a small impact in the Salmon River based on both methods while highly impacting *E. coli* concentration in the other two sites. The excessive amounts of nutrients (nitrate/phosphates) and pesticides usually are generated from agricultural activity that can be washed out by rainfalls in the region (Sasakova et al., 2018). Therefore, due to the number of farmland and high agricultural activities around the Salmon River, precipitation on the sampling day and over three days were expected to contribute to the microbial contamination in this region. However, compared to Cheakamus, which is surrounded mainly by forests, weather events have the lowest influence on *E. coli* prediction in the Salmon River.

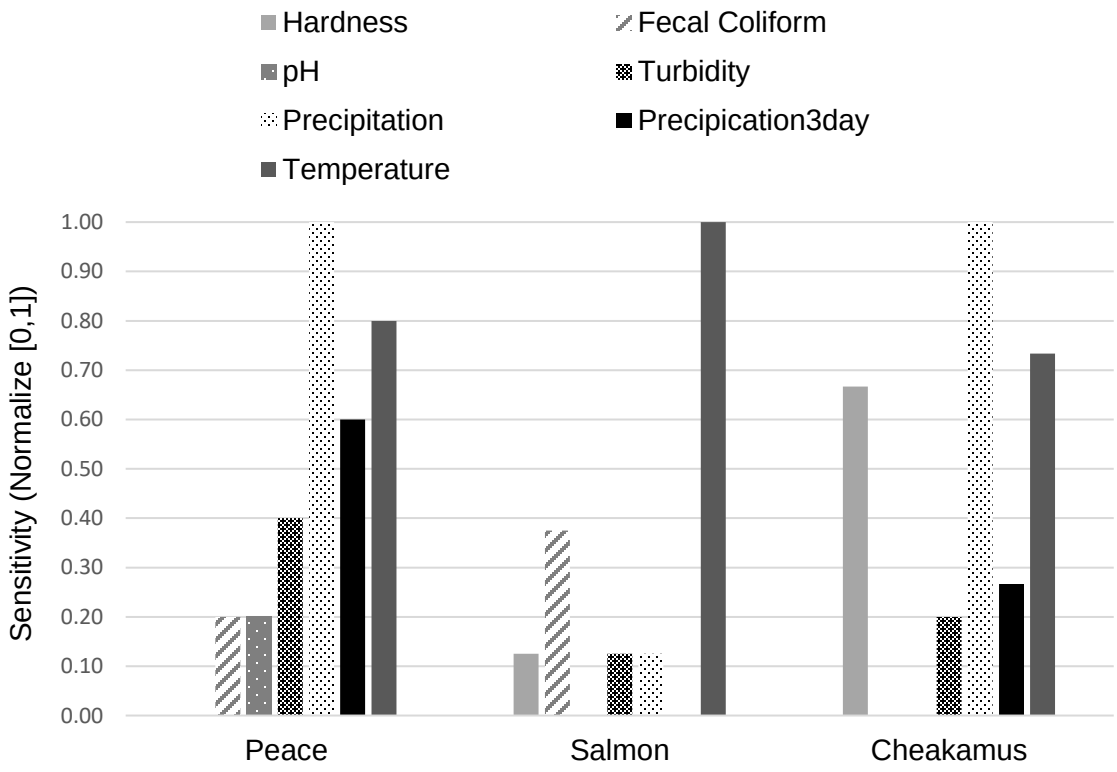


Figure 5.4 Importance of Variables Based on their Analysis of Sensitivity in Predicting *E. coli* for three surface water bodies in BC.

The relative importance of water quality parameters also showed significant variation over each monitoring site. For instance, hardness was denoted to be very impactful (>0.5) for Cheakamus River and moderately impactful for Salmon River but very poor (<0.1) for Peace River. On the other side, *E. coli* in Peace River seemed to be more affected by pH than hardness.

Although based on the sensitivity analysis, pH's impact in Salmon and Cheakamus River was observed to be ignorable (Figure 5.4), considering the strength of influence (Figure 5.3), pH seemed to be critical for Salmon River. Therefore, it is possible that *E. coli* was not significantly affected by pH only in the Cheakamus River. The mentioned difference in land cover around the Cheakamus river and wide industrial and tourism activities in its forestry area (BC Conservation Foundation, 2009; Clague et al., 2003) can be responsible for the distinct reaction of *E. coli* and pH in this region. For instance, incidents such as the sodium hydroxide spill in 2005 due to industrial activities could highly influence the chemical and, consequently, the biological water balance in testing points.

Furthermore, the relative impact of fecal coliform was observed for all monitoring sites based on the observations for the strength of influence (Figure 5.3). Considering the fact that *E. coli* is a major species of fecal coliforms this strong relationship was expected for all sites while only Peace and Salmon River have shown sensitivity to fecal coliform. The variation of fecal coliform and *E. coli* relationship in different case studies and locations has also been reported in the literature (Oliveira et al., 2017) which can be due to the different features of each monitoring site or measurement errors in the lab. For instance, the stronger influence of fecal coliform in the Peace and Salmon River was expected due to the agricultural activities and the higher livestock contamination of these areas. The impact of farming activities on the variation of *E. coli* and water quality parameters has also been studied in the literature. As such, Namugize etal. (2018) have studied *E. coli* and water-quality variations in different locations of Midmar Dam's stream and indicated *E. coli* increase in the lands surrounded by agricultural practices (Namugize et al., 2018).

Similarly, Kibena et al. (2014) have indicated the contribution of land-use changes in increasing total phosphate and biological oxygen demand in the Zimbabwe River (Kibena et al., 2014). The study also reported that the changes or impact of land cover seems to be more highlighted in the wet season than the dry season. The observations align with the variables' linear correlation, indicating that each site's surrounding land usage/cover can alleviate or maximize the weather and water quality parameter impact. Therefore, besides the finding in the literature, the results here imply that the normally observed water quality factors are not a universal indicator of surface waters quality and cannot solely explain the complex relationship of the aquatic processes. Therefore, a more comprehensive analysis that can include most of the factors such as land usage, weather condition or seasonal changes is required for making decisions regarding the contamination of these watersheds.

5.3.3 Pathogen Levels under Different Scenarios

The application of BBNs made it possible to quantify the interaction and the impact of different parameters using sensitivity and influence analysis in the previous section. Although the parameters and their impact on both *Cryptosporidium* and *E. coli* levels were captured earlier, the probabilistic nature of BBN models allows further analysis to see how changes in the recognized affecting factors will cause changes in the level of the pathogen. This feature of BBNs provides a robust tool for decision-makers to develop different scenarios and observe the reaction of source water quality indicators to the outcomes of the scenario. Therefore, being aware of the system's behavior in diverse circumstances will garner a better preparation for risk managers.

As weather parameters were observed in this study to affect the level of the considered pathogens, different scenarios have been defined in this section, and the probability of pathogen's present under each scenario was examined. It should be mentioned that scenarios were basically assumptions made based on the available data, annual weather observation and climatic perspectives provided in the literature. The trained BBN models in the previous sections were used as the base models, and their probability distribution was updated based on the new scenarios. The specific feature of these scenarios was propagated throughout the model, and the variations on the probability of the pathogen's presence were compared to the base models.

Considering *Cryptosporidium* and *E. coli* as the target pathogens, the first scenario was an observation of heavy rainfalls (>10mm) since this parameter was observed to influence both target parameters based on sensitivity and influence analysis. Also, several studies in the literature have shown pathogen's variation after a high precipitation rate (e.g., Duris et al., 2013b; Masina et al., 2018; Young et al., 2015). The second scenario was based on having the fecal coliform presence as the evidence; this scenario was defined only to evaluate the *Cryptosporidium*'s presence and examine how its probability distribution reacts to the presence or absence of fecal indicators. Finally, the third scenario was described to investigate the probability of both *E. coli* and *Cryptosporidium*'s presence to the projected climate changes over temperature and precipitation. Since the reported expected changes were significantly different based on several climate models, a 1.9 ^{o}C increase in temperature and 7.5% increase in precipitation level were assumed for this scenario (McClure et al., 2022). Also, the annual precipitation and temperature of the case study were utilized to reflect these expected climatic changes over the parameters.

The probability of the *Cryptosporidium* and *E. coli* was observed under each of these scenarios. Propagating the impact of the first scenarios on the *Cryptosporidium* model, the

probability of *Cryptosporidium*'s presence increased from 15% (probability of presence in the initial trained model) to 27%. While this increase of pathogens level under severe rainfall was expected, the reaction of *E. coli* was observed to be different under this scenario for each case study. As the land cover of Peace and Salmon River is close, a similar observation was expected for these two sites. However, the probability of *E. coli*'s presence increased from 43% to 50% only for Peace River and reduced from 55 to 49 and from 56 to 50 for Cheakamus and Salmon Rivers, respectively. These specific results, here again, could highlight the influence of land cover on the interconnection of parameters and implies that expecting the same behavior in pathogens transmission for the case studies with similar land features can be misleading. Since there was no other *Cryptosporidium* data available with the different case studies, it can be of future studies to compare the result of this study with varying sources of *Cryptosporidium* data to see if this pattern changes for each location or not.

Reflecting the second scenario, increased the *Cryptosporidium*'s presence probability from 15% to 16%. This minor increase in *Cryptosporidium*'s presence probability by only evidencing the fecal coliform's presence, besides the previous observations of sensitivity and dependency analysis emphasized the need for a direct assessment of parasites such as *Cryptosporidium* rather than relying only on the presence of fecal indicator bacteria.

The projected climate change scenario and propagating the intensified temperature and rainfall, resulted in an increased presence's probability for *Cryptosporidium* from 15% to 29%. A sharper increase in presence probability was expected under this scenario because the increased temperature can consequently increase microorganism activity or provide a more favorable environment for their growth, and higher precipitation facilitates the transportation of the pathogen. A more pronounced increase in presence could be expected for future real-time cases since the scenarios in this work were developed based on projections from large scale data, and the real projected variations can be different. Also, the way of reflecting the projected change under climate change can likely impact the probability distribution of *Cryptosporidium*. While a linear modification on data was used to develop climate change scenario in this work, the availability of real data from projected changes extracted from climate change models possibly allow for a more accurate reflection.

It can also be due to the same reason that the changes in *E. coli* presence probability for the third scenario were the same as the results observed for the first scenarios. Although severe rainfall was the only assumption in the first scenario, both temperature and precipitation were assumed to be changed in the third scenario. The requirement of discretizing the data for

developing BBNs can be another issue of reflecting the changes since the data will be classified into a specific number of bins. Therefore, further studies that can use continuous data to prevent information loss can help develop more practical scenario analysis. Also, taking the benefit of other climate change models that could generate real data proportion to the projected changes can be helpful in more accurate probability predictions.

5.4 Summary

The factors affecting the *Cryptosporidium* and *E. coli* level in source water were assessed using BBNs in this study. The developed BBN prediction models were used in this section to explore the relationship between weather and water quality parameters as the predictors and the *Cryptosporidium* and *E. coli* as the target variables. The strength of influence within the BBN models and sensitivity analysis were investigated in this section to garner an insight into how each parameter impacts the presence of *Cryptosporidium* and *E. coli*. Also, CPTs of BBN models allowed for analyzing different climatic scenarios to see how projected changes can worsen source water qualities in terms of pathogen's presence.

Regarding the assessment of *Cryptosporidium*, both sensitivity analysis and strength of influence observations emphasized the importance of weather parameters on the decision made by the model for predicting the presence or absence of *Cryptosporidium*. Despite expecting a high dependency on fecal coliform, a poor interconnection between *Cryptosporidium* and fecal coliform implied that these indicators bacteria seem to be a poor universal indicator for other parasites' presence. The observed increase in *Cryptosporidium*'s presence probability to more than 10%, under severe rainfall and climate change scenarios also aligned with the sensitivity analysis and strength of influence observation and reinforced the importance of considering weather parameters in water quality assessments.

The relationship between *E. coli* and other water quality/ weather parameters were observed to be very site-specific based on the employed approaches. For instance, the land cover around the case study seemed to impact the parameters' dependency. The *E. coli* level in both Peace and Salmon Rivers were observed to be more sensitive to fecal coliform, possibly due to the surrounding agricultural lands. On the other hand, the pathogen's concentration in the Cheakamus River presented a high dependency on hardness and weather parameters, as located in forestry areas. The analysis of scenarios also supported the site-specific variation of *E. coli* and the necessity of considering a wide range of parameters for evaluating the microbial quality of source waters.

Chapter 6: Conclusion

6.1 Summary of Contributions

In this thesis, *Cryptosporidium* and *E. coli* levels were assessed in different source waters using Bayesian Belief Network (BBN) models. The impacting parameters on the capability of the models in predicting microbial quality indicators was investigated to introduce a better approach in source water quality estimations. The lack of available data for microbial contamination, which was a limiting factor in developing prediction models, have been addressed. The contributions of this thesis can be summarized as follows.

1. **BBNs successfully applied for predicting the microbial quality indicators, and the application of data balancing algorithms was observed to improve the accuracy of prediction in the case of lacking complete data.**

   Results from applying BBNs on four distinct source waters with different datasets, demonstrated the ability of these models to predict *E. coli* and *Cryptosporidium* as two pathogens of concern. However, it is well known that complete and extensive historical data is required to train data-driven models. The observations in this thesis indicate that data balancing and synthetic sampling algorithms, such as ADASYN and SMOTE, can address the lack of data and contribute to improving the accuracy of prediction models for source water quality assessments.

   Two different splitting approaches were tried in this work to see if including the test size in synthesizing the training dataset can cause an overfitting situation for the model or not. The observation showed that this inclusion could be deceptive since excluding test data in the data generation step showed a drop in the prediction capacity of the model developed with generated data. Although the application of data balancing algorithms was shown to improve the prediction accuracy for BBN models, the higher accuracy achieved for predicting *E. coli* than *Cryptosporidium* showed that these methods cannot completely compensate for the lack of real observed data.

2. **The impact of weather characteristics on *E. coli* and *Cryptosporidium* level in source waters investigated.**

The underlying relationships between climatic parameters (temperature and precipitation) were analyzed through the linear method and BBN's performance. The observation indicated the inefficiency of monitoring fecal bacteria as a representative of microbial contamination. The high dependency that denoted between precipitation and temperature for prediction on *E. coli* and *Cryptosporidium* indicated the necessity of considering weather events besides water quality parameters in microbial quality assessment of source waters before intaking to treatment systems. Also, the non-consistency of the relationship between *E. coli* and other variables for each monitoring site indicated that different characteristics of case studies should be considered and reflected in developing models for microbial quality assessments. Developing different climatic scenarios also reinforced these observations by indicating the distinct reaction of each pathogen to the projected variations of weather parameters such as precipitation. Although the presence's probability of *Cryptosporidium* improved under heavy rainfalls and increased temperature and the likelihood of *E. coli's* presence was observed to reduce for two monitoring sites. These findings highlight the impact of other features of the case study on pathogen concentration in source waters.

6.2 Limitations and Suggestions for Future Works

The *Cryptosporidium* dataset used in this study was from a reservoir with multiple regulation systems for alleviating the impact of severe weather on source water quality. However, applying the same models for different case studies with lower standards can give a more realistic dataset and observations in future works. The only available dataset for long-term observation of *Cryptosporidium* was used in this study. The difficulties in collecting *Cryptosporidium* data make it rare to find a comprehensive dataset that would probably be even more challenging for regions with lower standards. Therefore, it could be of future interest to consider other available datasets worldwide to assess the capability of BBN models in predicting *Cryptosporidium*.

BBNs allow for quantifying the relationship of variables and investigating the target parameters' probability variation under different scenarios. However, the need for discretizing the data prior to developing the model can reduce the accuracy of the model when the real data are not available, and the scenarios should be reflected by modifying the data. Discretizing the modified data into a specific number of bins will result in missing some part of information and disable reflecting the associated changes of scenarios. Therefore, developing BBNs that are

capable of using continuous data can improve the capability of the model in the assessment of pathogens under different scenarios.

Furthermore, the impact of weather characteristics on *E. coli* was observed to be very site-specific. It seems that the land cover of the monitoring site can play an essential role in microbial contamination of the surrounding source waters. Therefore, besides considering the weather characteristics of the case study, propagating the feature of the land cover can contribute to developing more accurate and realistic BBN models. This can be of future interest to improve the BBN models developed here by considering other variables reflecting broader properties of the source waters.

Bibliography

Abia, A. L. K., Ubomba-Jaswa, E., & Momba, M. N. B. (2015). Impact of seasonal variation on Escherichia coli concentrations in the riverbed sediments in the Apies River, South Africa. *Science of The Total Environment*, *537*, 462–469. https://doi.org/10.1016/j.scitotenv.2015.07.132

A.F. Pressini. (2018). Causation, Proability, and the Continuity Bind. *The Birtish Journal for the Philosophy of Science*, *69*, 881–909.

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling and Software*, *26*(12), 1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Aguilera, P. A., Fernández, A., Ropero, R. F., & Molina, L. (2013). Groundwater quality assessment using data clustering based on hybrid Bayesian networks. *Stochastic Environmental Research and Risk Assessment*, *27*(2), 435–447.

Al-Adhaileh, M., & Alsaade, F. (2021). Modelling and Prediction of Water Quality by Using Artificial Intelligence. *Sustainability*, *13*, 4259. https://doi.org/10.3390/su13084259

Alameddine, I., Cha, Y., & Reckhow, K. H. (2011). An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics. *Environmental Modelling & Software*, *26*(2), 163–172. https://doi.org/10.1016/j.envsoft.2010.08.007

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

Allende, A., Castro-Ibáñez, I., Lindqvist, R., Gil, M. I., Uyttendaele, M., & Jacxsens, L. (2017). Quantitative contamination assessment of Escherichia coli in baby spinach primary production in Spain: Effects of weather conditions and agricultural practices.

International Journal of Food Microbiology, *257*, 238–246. https://doi.org/10.1016/j.ijfoodmicro.2017.06.027

Andriyas, S., & McKee, M. (2015). Development of a bayesian belief network model framework for analyzing farmers' irrigation behavior. *Journal of Agricultural Science*, *7*(7), 1.

Atherholt, T. B., LeChevallier, M. W., Norton, W. D., & Rosen, J. S. (1998). *Effect of rainfall on giardia and crypto*. https://doi.org/10.7282/T3X92F7V

Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018a). Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management*, *206*, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049

Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018b). Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management*, *206*, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049

Avilés, A., Célleri, R., Solera, A., & Paredes, J. (2016). Probabilistic forecasting of drought events using Markov chain-and Bayesian network-based models: A case study of an Andean regulated river basin. *Water*, *8*(2), 37.

Baldock, T. E., Shabani, B., & Callaghan, D. P. (2019). Open access Bayesian Belief Networks for estimating the hydrodynamics and shoreline response behind fringing reefs subject to climate changes and reef degradation. *Environmental Modelling & Software*, *119*, 327–340. https://doi.org/10.1016/j.envsoft.2019.07.001

Baldursson, S., & Karanis, P. (2011a). Waterborne transmission of protozoan parasites: Review of worldwide outbreaks—An update 2004-2010. *Water Research*, *45*(20), 6603–6614. https://doi.org/10.1016/j.watres.2011.10.013

Baldursson, S., & Karanis, P. (2011b). Waterborne transmission of protozoan parasites: Review of worldwide outbreaks–an update 2004–2010. *Water Research*, *45*(20), 6603–6614.

Ballesté, E., Demeter, K., Masterson, B., Timoneda, N., & Meijer, W. (2019). *Implementation and Integration of Microbial Source Tracking in a River Watershed Monitoring Plan*. https://doi.org/10.1101/514257

Ban, S. S., Pressey, R. L., & Graham, N. A. J. (2014). Assessing interactions of multiple stressors when data are limited: A Bayesian belief network applied to coral reefs. *Global Environmental Change*, *27*, 64–72. https://doi.org/10.1016/j.gloenvcha.2014.04.018

Barragán, J. L. M., Cuesta, L. D. I., & Susa, M. S. R. (2021). Quantitative microbial risk assessment to estimate the public health risk from exposure to enterotoxigenic E. coli in drinking water in the rural area of Villapinzon, Colombia. *Microbial Risk Analysis*, *18*, 100173. https://doi.org/10.1016/j.mran.2021.100173

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

BC Conservation Foundation. (2009). *Cheakamus/Squamish River Engineered Log Jam Pilot Project*.

BC Hydro. (2012). *Cheakamus River Water Use Plan*.

B.C.Ministry of Health. (2012). *Part B: Drinking Water Treatment Objectives for Surface Water in B.C., Drinking Water Officers' Guide 2017*.

Beaudequin, D., Harden, F., Roiko, A., & Mengersen, K. (2016a). Utility of Bayesian networks in QMRA-based evaluation of risk reduction options for recycled water. *Science of The Total Environment*, *541*, 1393–1409. https://doi.org/10.1016/j.scitotenv.2015.10.030

Beaudequin, D., Harden, F., Roiko, A., & Mengersen, K. (2016b). Utility of Bayesian networks in QMRA-based evaluation of risk reduction options for recycled water. *Science of The Total Environment*, *541*, 1393–1409. https://doi.org/10.1016/j.scitotenv.2015.10.030

Benham, B. L. (2006). Modeling Bacteria Fate and Transport in Watersheds to Support TMDLs (B. L. Benham, C. Baffaut, R. W. Zeckoski, K. R. Mankin, Y. A. Pachepsky, A. M. Sadeghi, K. M. Brannan, M. L. Soupir, & M. J. Habersack, Trans.). *Transactions of the ASABE*, *v. 49*(4), 987–1002. PubAg.

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R., & Ramazzotti, D. (2018). Learning the Structure of Bayesian Networks: A Quantitative Assessment of the Effect of Different Algorithmic Schemes. *Complexity*, *2018*, 1591878. https://doi.org/10.1155/2018/1591878

Bergion, V., Sokolova, E., Åström, J., Lindhe, A., Sörén, K., & Rosén, L. (2017). Hydrological modelling in a drinking water catchment area as a means of evaluating pathogen risk reduction. *Journal of Hydrology*, *544*, 74–85. https://doi.org/10.1016/j.jhydrol.2016.11.011

Bertone, E., Sahin, O., Richards, R., & Roiko, A. (2016a). Extreme events, water quality and health: A participatory Bayesian risk assessment tool for managers of reservoirs. *Journal of Cleaner Production*, *135*, 657–667. https://doi.org/10.1016/j.jclepro.2016.06.158

Bertone, E., Sahin, O., Richards, R., & Roiko, A. (2016b). Extreme events, water quality and health: A participatory Bayesian risk assessment tool for managers of reservoirs. *Journal of Cleaner Production*, *135*, 657–667. https://doi.org/10.1016/j.jclepro.2016.06.158

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, *128*(9).

Bocquet, S. (2022). Ocean wave autocorrelation function. *Applied Mathematics and Computation*, *426*, 127114. https://doi.org/10.1016/j.amc.2022.127114

Brandt, J., & Lanzén, E. (2021). *A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*.

Breternitz, B. S., Barbosa da Veiga, D. P., Pepe Razzolini, M. T., & Nardocci, A. C. (2020). Land use associated with Cryptosporidium sp. And Giardia sp.in surface water supply in the

state of São Paulo, Brazil. *Environmental Pollution*, *266*, 115143. https://doi.org/10.1016/j.envpol.2020.115143

Brion, G., Neelakantan, T. R., & Lingireddy, S. (2001). Using neural networks to predict peak Cryptosporidium concentrations. *Journal American Water Works Association - J AMER WATER WORK ASSN*, *93*, 99–105. https://doi.org/10.1002/j.1551-8833.2001.tb09103.x

Brooks, W., Corsi, S., Fienen, M., & Carvin, R. (2016). Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental Modelling & Software*, *76*, 81–94. https://doi.org/10.1016/j.envsoft.2015.10.012

Canova Calori, I., Stålhane, T., & Ziemer, S. (2007). Robustness analysis using fmea and bbn: Case study for a web-based application. *Webist 2007 - 3rd International Conference on Web Information Systems and Technologies, Proceedings*.

CAO, S. K., JIANG, Y. Y., YUAN, Z. Y., YIN, J. H., XU, M., XUE, J. B., TANG, L. H., SHEN, Y. J., & CAO, J. P. (2021). Quantitative Microbial Risk Assessment of Cryptosporidium and Giardia in Public Drinking Water in China. *Biomedical and Environmental Sciences*, *34*(6), 493–498. https://doi.org/10.3967/bes2021.068

Carmena, D. (2010). Waterborne transmission of Cryptosporidium and Giardia: Detection, surveillance and implications for public health. *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology*, *20*, 3–4.

Cha, Y., Park, M.-H., Lee, S.-H., Kim, J. H., & Cho, K. H. (2016). Modeling spatiotemporal bacterial variability with meteorological and watershed land-use characteristics. *Water Research*, *100*, 306–315. https://doi.org/10.1016/j.watres.2016.05.024

Chan, H., & Darwiche, A. (2004). Sensitivity Analysis in Bayesian Networks: From Single to Multiple Parameters. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 67–75.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). snopes.com: Two-Striped Telamonia Spider. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357.

Chen, W.-B., & Liu, W.-C. (2015). Water Quality Modeling in Reservoirs Using Multivariate Linear Regression and Two Neural Network Models. *Advances in Artificial Neural Systems*, *2015*, 1–12. https://doi.org/10.1155/2015/521721

Cheng, J., Niu, S., & Kim, Y. (2013). Relationship between water quality parameters and the survival of indicator microorganisms – Escherichia coli – in a stormwater wetland. *Water Science and Technology*, *68*(7), 1650–1656. https://doi.org/10.2166/wst.2013.386

Christophersen, A., Deligne, N., Hanea, A., Chardot, L., Fournier, N., & Aspinall, W. (2018). Bayesian Network Modeling and Expert Elicitation for Probabilistic Eruption Forecasting: Pilot Study for Whakaari/White Island, New Zealand. *Frontiers in Earth Science*, *6*, 211. https://doi.org/10.3389/feart.2018.00211

Clague, J., Turner, R., & Reyes, A. (2003). Record of recent river channel instability, Cheakamus Valley, British Columbia. *Geomorphology*, *53*, 317–332. https://doi.org/10.1016/S0169-555X(02)00321-5

Coffey, R., Cummins, E., Bhreathnach, N., Flaherty, V. O., & Cormican, M. (2010). Development of a pathogen transport model for Irish catchments using SWAT. *Agricultural Water Management*, *97*(1), 101–111. https://doi.org/10.1016/j.agwat.2009.08.017

Coffey, R., Cummins, E., O'Flaherty, V., & Cormican, M. (2010). Analysis of the soil and water assessment tool (SWAT) to model Cryptosporidium in surface water sources. *Biosystems Engineering*, *106*(3), 303–314. https://doi.org/10.1016/j.biosystemseng.2010.04.003

Craun, G. F., Hubbs, S. A., Frost, F., Calderon, R. L., & Via, S. H. (1998). Waterborne outbreaks of cryptosporidiosis. *Journal-American Water Works Association*, *90*(9), 81–91.

Cryptosporidium: Drinking Water Health Advisory. (2001). 31.

Daniel, D., Iswarani, W. P., Pande, S., & Rietveld, L. (2020). A Bayesian Belief Network model to link sanitary inspection data to drinking water quality in a medium resource setting in rural Indonesia. *Scientific Reports*, *10*(1), 18867. https://doi.org/10.1038/s41598-020-75827-7

de Vries, J., Kraak, M. H. S., Skeffington, R. A., Wade, A. J., & Verdonschot, P. F. M. (2021). A Bayesian network to simulate macroinvertebrate responses to multiple stressors in lowland streams. *Water Research*, *194*, 116952. https://doi.org/10.1016/j.watres.2021.116952

Death, R., Death, F., Stubbington, R., Joy, M., & Belt, M. (2015). How good are Bayesian belief networks for environmental management? A test with data from an agricultural river catchment. *Freshwater Biology*, *60*. https://doi.org/10.1111/fwb.12655

DeFlorio-Barker, S., Wing, C., Jones, R. M., & Dorevitch, S. (2018). Estimate of incidence and cost of recreational waterborne illness on United States surface waters. *Environmental Health*, *17*(1), 3. https://doi.org/10.1186/s12940-017-0347-9

Deng, W., Wang, G., & Zhang, X. (2015). A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting. *Chemometrics and Intelligent Laboratory Systems*, *149*, 39–49. https://doi.org/10.1016/j.chemolab.2015.09.017

Donald, M., Cook, A., & Mengersen, K. (2009). Bayesian network for risk of diarrhea associated with the use of recycled water. *Risk Analysis: An International Journal*, *29*(12), 1672–1685.

Dondeynaz, C., López Puga, J., & Carmona Moreno, C. (2013). Bayesian networks modelling in support to cross-cutting analysis of water supply and sanitation in developing countries. *Hydrology and Earth System Sciences*, *17*(9), 3397–3419.

Dorner, S. M., Huck, P. M., & Slawson, R. M. (2004). Estimating potential environmental loadings of Cryptosporidium spp. And Campylobacter spp. From livestock in the Grand River Watershed, Ontario, Canada. *Environmental Science and Technology*, *38*(12), 3370–3380. https://doi.org/10.1021/es035208+

Downs, T., & Tang, A. (2004). Boosting the Tree Augmented Naïve Bayes Classifier. In Z. R. Yang, H. Yin, & R. M. Everson (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2004* (pp. 708–713). Springer Berlin Heidelberg.

Drinking Water Officers' Guide,gov,BC. (2017). *Drinking Water Treatment Objctives(Microbiological) for Surface Water Supplies in Biritish Columbia* (Part B).

Duris, J. W., Reif, A. G., Krouse, D. A., & Isaacs, N. M. (2013a). Factors related to occurrence and distribution of selected bacterial and protozoan pathogens in Pennsylvania streams. *Water Research*, *47*(1), 300–314. https://doi.org/10.1016/j.watres.2012.10.006

Duris, J. W., Reif, A. G., Krouse, D. A., & Isaacs, N. M. (2013b). Factors related to occurrence and distribution of selected bacterial and protozoan pathogens in Pennsylvania streams. *Water Research*, *47*(1), 300–314. https://doi.org/10.1016/j.watres.2012.10.006

Dwivedi, D., Mohanty, B., & Lesikar, B. (2016). Impact of the Linked Surface Water-Soil Water-Groundwater System on Transport of E. coli in the Subsurface. *Water, Air, & Soil Pollution*, *227*, 351. https://doi.org/10.1007/s11270-016-3053-2

Edberg, S. C., Rice, E. W., Karlin, R. J., & Allen, M. J. (2000). Escherichia coli: The best biological drinking water indicator for public health protection. *Journal of Applied Microbiology*, *88*(S1), 106S-116S. https://doi.org/10.1111/j.1365-2672.2000.tb05338.x

Edge, T. A., Boyd, R. J., Shum, P., & Thomas, J. L. (2021). Microbial source tracking to identify fecal sources contaminating the Toronto Harbour and Don River watershed in wet and dry weather. *Journal of Great Lakes Research*, *47*(2), 366–377. https://doi.org/10.1016/j.jglr.2020.09.002

Efstratiou, A., Ongerth, J. E., & Karanis, P. (2017a). Waterborne transmission of protozoan parasites: Review of worldwide outbreaks—An update 2011–2016. *Water Research*, *114*, 14–22. https://doi.org/10.1016/j.watres.2017.01.036

Efstratiou, A., Ongerth, J. E., & Karanis, P. (2017b). Waterborne transmission of protozoan parasites: Review of worldwide outbreaks-an update 2011–2016. *Water Research*, *114*, 14–22.

Environment and Climate Change Canada. (2019). https://open.canada.ca/data/en/dataset/67b44816-9764-4609-ace1-68dc1764e9ea

Farrell, C., Hassard, F., Jefferson, B., Leziart, T., Nocker, A., & Jarvis, P. (2018). Turbidity composition and the relationship with microbial attachment and UV inactivation efficacy. *Science of The Total Environment*, *624*, 638–647. https://doi.org/10.1016/j.scitotenv.2017.12.173

Fasaee, M. A. K., Berglund, E., Pieper, K. J., Ling, E., Benham, B., & Edwards, M. (2021a). Developing a framework for classifying water lead levels at private drinking water systems: A Bayesian Belief Network approach. *Water Research*, *189*, 116641. https://doi.org/10.1016/j.watres.2020.116641

Fasaee, M. A. K., Berglund, E., Pieper, K. J., Ling, E., Benham, B., & Edwards, M. (2021b). Developing a framework for classifying water lead levels at private drinking water systems: A Bayesian Belief Network approach. *Water Research*, *189*, 116641. https://doi.org/10.1016/j.watres.2020.116641

Feng, Y., Barr, W., & Harper Jr, W. (2013). Neural network processing of microbial fuel cell signals for the identification of chemicals present in water. *Journal of Environmental Management*, *120*, 84–92.

Fenton, N., & Neil, M. (2012). Risk assessment and decision analysis with bayesian networks. In *Risk Assessment and Decision Analysis with Bayesian Networks* ((2nd ed.)). https://doi.org/10.1201/b21982

Fletcher, S. M., Stark, D., Harkness, J., & Ellis, J. (2012). Enteric protozoa in the developed world: A public health perspective. *Clinical Microbiology Reviews*, *25*(3), 420–449.

Forio, M. A. E., Burdon, F. J., De Troyer, N., Lock, K., Witing, F., Baert, L., De Saeyer, N., Rîșnoveanu, G., Popescu, C., Kupilas, B., Friberg, N., Boets, P., Johnson, R. K., Volk, M., McKie, B. G., & Goethals, P. (2021). A Bayesian Belief Network learning tool integrates multi-scale effects of riparian buffers on stream invertebrates. *Science of The Total Environment*, 152146. https://doi.org/10.1016/j.scitotenv.2021.152146

Francy, D. S., Brady, A. M. G., Cicale, J. R., Dalby, H. D., & Stelzer, E. A. (2020). Nowcasting methods for determining microbiological water quality at recreational beaches and drinking-water source waters. *Journal of Microbiological Methods*, *175*, 105970. https://doi.org/10.1016/j.mimet.2020.105970

Francy, D. S., Brady, A. M. G., & Zimmerman, T. M. (2019). *Real-time assessments of water quality—A nowcast for Escherichia coli and cyanobacterial toxins* (Report No. 2019–3061; Fact Sheet, p. 4). USGS Publications Warehouse. https://doi.org/10.3133/fs20193061

Francy, D. S., Stelzer, E. A., Duris, J. W., Brady, A. M. G., Harrison, J. H., Johnson, H. E., & Ware, M. W. (2013). Predictive models for Escherichia coli concentrations at inland lake beaches and relationship of model variables to pathogen detection. *Applied and Environmental Microbiology*, *79*(5), 1676–1688. https://doi.org/10.1128/AEM.02995-12

Frizzle, C., Fournier, R. A., Trudel, M., & Luther, J. E. (2022). Towards sustainable forestry: Using a spatial Bayesian belief network to quantify trade-offs among forest-related ecosystem services. *Journal of Environmental Management*, *301*, 113817. https://doi.org/10.1016/j.jenvman.2021.113817

Gallas-Lindemann, C., Sotiriadou, I., Plutzer, J., & Karanis, P. (2013). Prevalence and distribution of Cryptosporidium and Giardia in wastewater and the surface, drinking and ground waters in the Lower Rhine, Germany. *Epidemiology & Infection*, *141*(1), 9–21.

Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine Pollution Bulletin*, *64*(11), 2409–2420. https://doi.org/10.1016/j.marpolbul.2012.08.005

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining Explanations: An Overview of Interpretability of Machine Learning* (p. 89). https://doi.org/10.1109/DSAA.2018.00018

Gleick, P. H. (2002). *Dirty-water: Estimated deaths from water-related diseases 2000-2020*. Citeseer.

Goderniaux, P., Brouyere, S., Blenkinsop, S., Burton, A., Fowler, H. J., Orban, P., & Dassargues, A. (2011). Modeling climate change impacts on groundwater resources using transient stochastic climatic scenarios. *Water Resources Research*, *47*(12).

Goh, S., Reacher, M., Casemore, D. P., Verlander, N. Q., Chalmers, R., Knowles, M., Williams, J., Osborn, K., & Richards, S. (2004). Sporadic cryptosporidiosis, North Cumbria, England, 1996-2000. *Emerging Infectious Diseases*, *10*(6), 1007–1015. https://doi.org/10.3201/10.3201/eid1006.030325

Gonzalez, R. A., Conn, K. E., Crosswell, J. R., & Noble, R. T. (2012). Application of empirical predictive modeling using conventional and alternative fecal indicator bacteria in eastern North Carolina waters. *Water Research*, *46*(18), 5871–5882.

Gordon, D. M. (2001). Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology*, *147*(5), 1079–1085.

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 79–85. https://doi.org/10.1109/ICACCI.2017.8125820

Goulding, R., Jayasuriya, N., & Horan, E. (2012). A Bayesian network model to assess the public health risk associated with wet weather sewer overflows discharging into waterways. *Water Research*, *46*(16), 4933–4940.

Gronewold, A. D., Myers, L., Swall, J. L., & Noble, R. T. (2011). Addressing uncertainty in fecal indicator bacteria dark inactivation rates. *Water Research*, *45*(2), 652–664.

Grover, J. (2013). A Literature Review of Bayes' Theorem and Bayesian Belief Networks (BBN). In J. Grover (Ed.), *Strategic Economic Decision-Making: Using Bayesian Belief Networks to Solve Complex Problems* (pp. 11–27). Springer New York. https://doi.org/10.1007/978-1-4614-6040-4_2

Guo, H., & Li, H. (2022). A decomposition structure learning algorithm in Bayesian network based on a two-stage combination method. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00623-3

Haibo He, Yang Bai, E. A. G. and S. L. (2008). *Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE*

International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322-1328, d.

Haibo He, Yang Bai, E. A. Garcia, & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hamilton, K. A., Waso, M., Reyneke, B., Saeidi, N., Levine, A., Lalancette, C., Besner, M., Khan, W., & Ahmed, W. (2018). *Cryptosporidium* and *Giardia* in Wastewater and Surface Water Environments. *Journal of Environmental Quality*, *47*(5), 1006–1023. https://doi.org/10.2134/jeq2018.04.0132

Han, H., Wang, W.-Y., & Mao, B.-H. (2005a). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In D.-S. Huang, X.-P. Zhang, & G.-B. Huang (Eds.), *Advances in Intelligent Computing* (pp. 878–887). Springer Berlin Heidelberg.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005b). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning BT - Advances in Intelligent Computing* (D.-S. Huang, X.-P. Zhang, & G.-B. Huang, Eds.; pp. 878–887). Springer Berlin Heidelberg.

Han, J., & Kamber, M. (2019). *Classification: Advanced Methods Classification: Advanced Methods*. https://doi.org/10.1016/B978-0-12-381479-1.00009-5

Harvey, R., Lye, L., & Khan, A. (2009). *Regression models for predicting water temperatures and dissolved oxygen concentrations in Newfoundland rivers*. *1*, 295–304.

Harwood Valerie J., Levine Audrey D., Scott Troy M., Chivukula Vasanta, Lukasik Jerzy, Farrah Samuel R., & Rose Joan B. (2005). Validity of the Indicator Organism Paradigm for Pathogen Reduction in Reclaimed Water and Public Health Protection. *Applied and Environmental Microbiology*, *71*(6), 3163–3170. https://doi.org/10.1128/AEM.71.6.3163-3170.2005

Hassall, K. L., Dailey, G., Zawadzka, J., Milne, A. E., Harris, J. A., Corstanje, R., & Whitmore, A. P. (2019). Facilitating the elicitation of beliefs for use in Bayesian Belief modelling. *Environmental Modelling & Software*, *122*, 104539. https://doi.org/10.1016/j.envsoft.2019.104539

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, *20*, 197–243. https://doi.org/10.1007/BF00994016

Herrig, I., Seis, W., Fischer, H., Regnery, J., Manz, W., Reifferscheid, G., & Böer, S. (2019). Prediction of fecal indicator organism concentrations in rivers: The shifting role of environmental factors under varying flow conditions. *Environmental Sciences Europe*, *31*(1). https://doi.org/10.1186/s12302-019-0250-9

Hesar, A. (2013). Structure learning of bayesian belief networks using simulated annealing algorithm. *Middle East Journal of Scientific Research*, *18*, 1343–1348. https://doi.org/10.5829/idosi.mejsr.2013.18.9.12375

Hogan, J. N., Daniels, M. E., Watson, F. G., Conrad, P. A., Oates, S. C., Miller, M. A., Hardin, D., Byrne, B. A., Dominik, C., Melli, A., Jessup, D. A., & Miller, W. A. (2012). Longitudinal Poisson regression to evaluate the epidemiology of Cryptosporidium, Giardia, and fecal indicator bacteria in coastal California wetlands. *Applied and Environmental Microbiology*, *78*(10), 3606–3613. PubMed. https://doi.org/10.1128/AEM.00578-12

Hong, H., Zhang, Z., Guo, A., Shen, L., Sun, H., Liang, Y., Wu, F., & Lin, H. (2020). Radial basis function artificial neural network (RBF ANN) as well as the hybrid method of RBF ANN

and grey relational analysis able to well predict trihalomethanes levels in tap water. *Journal of Hydrology*, *591*, 125574. https://doi.org/10.1016/j.jhydrol.2020.125574

Howard, G., Charles, K., Pond, K., Brookshaw, A., Hossain, R., & Bartram, J. (2010). Securing 2020 vision for 2030: Climate change and ensuring resilience in water and sanitation services. *Journal of Water and Climate Change*, *1*(1), 2–16.

Hrudey, S. E., Huck, P. M., Payment, P., Gillham, R. W., & Hrudey, E. J. (2002). Walkerton: Lessons learned in comparison with waterborne outbreaks in the developed world. *Journal of Environmental Engineering and Science*, *1*(6), 397–407. https://doi.org/10.1139/s02-031

Hunter, P., Anderle de Sylor, M., Risebro, H., Nichols, G., Kay, D., & Hartemann, P. (2010). Quantitative Microbial Risk Assessment of Cryptosporidiosis and Giardiasis from Very Small Private Water Supplies. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *31*, 228–236. https://doi.org/10.1111/j.1539-6924.2010.01499.x

Hunter, P. R. (2003). Drinking water and diarrhoeal disease due to Escherichia coli. *Journal of Water and Health*, *1*(2), 65–72. https://doi.org/10.2166/wh.2003.0008

Islam, M. M. M., Hofstra, N., & Islam, Md. A. (2017). The Impact of Environmental Variables on Faecal Indicator Bacteria in the Betna River Basin, Bangladesh. *Environmental Processes*, *4*(2), 319–332. https://doi.org/10.1007/s40710-017-0239-6

Islam, Md. S., Hassan-uz-Zaman, Md., Islam, Md. S., Clemens, J. D., & Ahmed, N. (2020). Chapter 3—Waterborne pathogens: Review of outbreaks in developing nations. In M. N. Vara Prasad & A. Grobelak (Eds.), *Waterborne Pathogens* (pp. 43–56). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-818783-8.00003-7

Jacklin, D. (2016). *A study of the factors involved in Escherichia coli contamination of surface water, a case study on the Groenkloof natural springs*. https://doi.org/10.13140/RG.2.1.2445.4008

Jamieson, R., Gordon, R., Joy, D., & Lee, H. (2004). Assessing microbial pollution of rural surface waters: A review of current watershed scale modeling approaches. *Agricultural Water Management*, *70*(1), 1–17. https://doi.org/10.1016/j.agwat.2004.05.006

Jang, J., Hur, H.-G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T., & Ishii, S. (2017). Environmental Escherichia coli: Ecology and public health implications—A review. *Journal of Applied Microbiology*, *123*(3), 570–581. https://doi.org/10.1111/jam.13468

Jiang, L., Zhang, H., Cai, Z., & Su, J. (2005). Learning Tree Augmented Naive Bayes for Ranking. In L. Zhou, B. C. Ooi, & X. Meng (Eds.), *Database Systems for Advanced Applications* (pp. 688–698). Springer Berlin Heidelberg.

Jolivel, M., & Allard, M. (2017). Impact of permafrost thaw on the turbidity regime of a subarctic river: The Sheldrake River, Nunavik, Quebec. *Arctic Science*, *3*. https://doi.org/10.1139/AS-2016-0006

Jongh, M. De. (2014). *ALGORITHMS FOR CONSTRAINT-BASED LEARNING OF BAYESIAN NETWORK STRUCTURES WITH LARGE NUMBERS OF by Martijn de Jongh M. S. Computer Science , Delft University of Technology , 2007 B. E. Electrical Engineering , Technische Hogeschool Rijswijk , 2003 Submi*.

Jongsawat, N., & Road, R. (2017). *Using Tree Augmented Naïve Bayes Classifiers to Learn Restaurant Reviews Data*. *18*(1), 31–41.

Kalin, S.Isik, & B.G.Lockaby. (2010). Predicting water quality in unmonitored watersheds using artificial neural networks. *Journal of Environmental Quality*, *39*(4), 1429–1440.

Karim, M. R., Manshadi, F. D., Karpiscak, M. M., & Gerba, C. P. (2004). The persistence and removal of enteric pathogens in constructed wetlands. *Water Research*, *38*(7), 1831–1837. https://doi.org/10.1016/j.watres.2003.12.029

Kay, D., Crowther, J., Stapleton, C., Wyer, M., Fewtrell, L., Anthony, S., Bradford, M., Edwards, A., Francis, C., & Hopkins, M. (2008). Faecal indicator organism concentrations and catchment export coefficients in the UK. *Water Research*, *42*(10–11), 2649–2661.

Khalil, I. A., Troeger, C., Rao, P. C., Blacker, B. F., Brown, A., Brewer, T. G., Colombara, D. V., De Hostos, E. L., Engmann, C., & Guerrant, R. L. (2018). Morbidity, mortality, and long-term consequences associated with diarrhoea from Cryptosporidium infection in children younger than 5 years: A meta-analyses study. *The Lancet Global Health*, *6*(7), e758–e768.

Khan, F. M., Gupta, R., & Sekhri, S. (2021). Superposition learning-based model for prediction of E.coli in groundwater using physico-chemical water quality parameters. *Groundwater for Sustainable Development*, *13*, 100580. https://doi.org/10.1016/j.gsd.2021.100580

Kibena, J., Nhapi, I., & Gumindoga, W. (2014). Assessing the relationship between water quality parameters and changes in landuse patterns in the Upper Manyame River, Zimbabwe. *Physics and Chemistry of the Earth, Parts A/B/C*, *67–69*, 153–163. https://doi.org/10.1016/j.pce.2013.09.017

Kim, J. H., Shin, J.-K., Lee, H., Lee, D. H., Kang, J.-H., Cho, K. H., Lee, Y.-G., Chon, K., Baek, S.-S., & Park, Y. (2021). Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. *Water Research*, *207*, 117821. https://doi.org/10.1016/j.watres.2021.117821

Kim, J., Seo, D., Jang, M., & Kim, J. (2021). Augmentation of limited input data using an artificial neural network method to improve the accuracy of water quality modeling in a large lake. *Journal of Hydrology*, *602*, 126817. https://doi.org/10.1016/j.jhydrol.2021.126817

Kim, J.-H., von Gunten, U., & Mariñas, B. J. (2004). Simultaneous Prediction of Cryptosporidium parvum Oocyst Inactivation and Bromate Formation during Ozonation of Synthetic

Waters. *Environmental Science & Technology*, *38*(7), 2232–2241. https://doi.org/10.1021/es034760w

King, S. L., Bennett, K. P., & List, S. (2000). Modeling noncatastrophic individual tree mortality using logistic regression, neural networks, and support vector methods. *Computers and Electronics in Agriculture*, *27*(1–3), 401–406.

Kjærulff, U., & Van Der Gaag, L. C. (2013). Making sensitivity analysis computationally efficient. *ArXiv Preprint ArXiv:1301.3868*.

Kleinheinz, G. T., McDermott, C. M., Hughes, S., & Brown, A. (2010). Effects of Rainfall on E. coli Concentrations at Door County, Wisconsin Beaches. *International Journal of Microbiology*, *2009*, 876050. https://doi.org/10.1155/2009/876050

Koiter, J. R. (2006). Visualizing inference in Bayesian networks. *Master of Science Thesis University of Technology*.

Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., Wu, Y., Sow, S. O., Sur, D., Breiman, R. F., Faruque, A. S. G., Zaidi, A. K. M., Saha, D., Alonso, P. L., Tamboura, B., Sanogo, D., Onwuchekwa, U., Manna, B., Ramamurthy, T., … Levine, M. M. (2013). Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. *The Lancet*, *382*(9888), 209–222. https://doi.org/10.1016/S0140-6736(13)60844-2

KRISHNAMURTI, K., & KATE, S. R. (1951). Changes in Electrical Conductivity during Bacterial Growth. *Nature*, *168*(4265), 170–170. https://doi.org/10.1038/168170b0

Lake, I. R., Nichols, G., Bentham, G., Harrison, F. C. D., Hunter, P. R., & Kovats, R. S. (2007). Cryptosporidiosis decline after regulation, England and Wales, 1989-2005. *Emerging Infectious Diseases*, *13*(4), 623–625. https://doi.org/10.3201/eid1304.060890

Lal, A., Baker, M. G., Hales, S., & French, N. P. (2013). Potential effects of global environmental changes on cryptosporidiosis and giardiasis transmission. *Trends in Parasitology*, *29*(2), 83–90.

Lal, A., Fearnley, E., & Kirk, M. (2015). The risk of reported cryptosporidiosis in children aged <5 years in Australia is highest in very remote regions. *International Journal of Environmental Research and Public Health*, *12*(9), 11815–11828. https://doi.org/10.3390/ijerph120911815

Lalancette, C., Papineau, I., Payment, P., Dorner, S., Servais, P., Barbeau, B., Di Giovanni, G. D., & Prévost, M. (2014a). Changes in Escherichia coli to Cryptosporidium ratios for various fecal pollution sources and drinking water intakes. *Water Research*, *55*, 150–161. https://doi.org/10.1016/j.watres.2014.01.050

Lalancette, C., Papineau, I., Payment, P., Dorner, S., Servais, P., Barbeau, B., Di Giovanni, G. D., & Prévost, M. (2014b). Changes in Escherichia coli to Cryptosporidium ratios for various fecal pollution sources and drinking water intakes. *Water Research*, *55*, 150–161. https://doi.org/10.1016/j.watres.2014.01.050

Landuyt, D., Broekx, S., D'hondt, R., Engelen, G., Aertsens, J., & Goethals, P. L. (2013). A review of Bayesian belief networks in ecosystem service modelling. *Environmental Modelling & Software*, *46*, 1–11.

Laurent, J., & Mazumder, A. (2012). The influence of land-use composition on fecal contamination of riverine source water in southern British Columbia. *Water Resources Research*, *48*. https://doi.org/10.1029/2012WR012455

Laurila-Pant, M., Mäntyniemi, S., Venesjärvi, R., & Lehikoinen, A. (2019). Incorporating stakeholders' values into environmental decision support: A Bayesian Belief Network approach. *Science of The Total Environment*, *697*, 134026. https://doi.org/10.1016/j.scitotenv.2019.134026

Lechevallier, M. W., & Au, K.-K. (2004). 3.1 Factors Affecting Disinfection. *WHO Water Treatment and Pathogen Control: Process Efficiency in Achieving Safe Drinking Water*, 41–65.

Leland, D., McAnulty, J., Keene, W., & Stevens, G. (1993). A Cryptosporidiosis Outbreak in a Filtered-Water Supply. *Journal (American Water Works Association)*, *85*(6), 34–42. JSTOR.

Lemarchand, K., & Lebaron, P. (2003). Occurrence of Salmonella spp. And Cryptosporidium spp. In a French coastal watershed: Relationship with fecal indicators. *FEMS Microbiology Letters*, *218*(1), 203–209. https://doi.org/10.1111/j.1574-6968.2003.tb11519.x

Levy, K., Hubbard, A. E., Nelson, K. L., & Eisenberg, J. N. S. (2009a). Drivers of Water Quality Variability in Northern Coastal Ecuador. *Environmental Science & Technology*, *43*(6), 1788–1797. https://doi.org/10.1021/es8022545

Levy, K., Hubbard, A. E., Nelson, K. L., & Eisenberg, J. N. S. (2009b). Drivers of Water Quality Variability in Northern Coastal Ecuador. *Environmental Science & Technology*, *43*(6), 1788–1797. https://doi.org/10.1021/es8022545

Li, L. X., & Abdul Rahman, S. S. (2018). Students' learning style detection using tree augmented naive Bayes. *Royal Society Open Science*, *5*(7), 172108.

Li, R. A., McDonald, J. A., Sathasivan, A., & Khan, S. J. (2021). A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems. *Water Research*, *190*, 116712. https://doi.org/10.1016/j.watres.2020.116712

Li, W., & Guo, Q. (2013). How to assess the prediction accuracy of species presence–absence models without absence data? *Ecography*, *36*. https://doi.org/10.1111/j.1600-0587.2013.07585.x

Ligda, P., Claerebout, E., Kostopoulou, D., Zdragas, A., Casaert, S., Robertson, L. J., & Sotiraki, S. (2020a). Cryptosporidium and Giardia in surface water and drinking water: Animal sources and towards the use of a machine-learning approach as a tool for predicting contamination. *Environmental Pollution*, *264*, 114766. https://doi.org/10.1016/j.envpol.2020.114766

Ligda, P., Claerebout, E., Kostopoulou, D., Zdragas, A., Casaert, S., Robertson, L. J., & Sotiraki, S. (2020b). Cryptosporidium and Giardia in surface water and drinking water: Animal sources and towards the use of a machine-learning approach as a tool for predicting contamination. *Environmental Pollution*, *264*, 114766. https://doi.org/10.1016/j.envpol.2020.114766

Lisnyj, K. T., & Dickson-Anderson, S. E. (2018). Community resilience in Walkerton, Canada: Sixteen years post-outbreak. *International Journal of Disaster Risk Reduction*, *31*, 196–202. https://doi.org/10.1016/j.ijdrr.2018.05.001

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., & Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012*, *58*(3), 458–465. https://doi.org/10.1016/j.mcm.2011.11.021

Liu, W., An, W., Jeppesen, E., Ma, J., Yang, M., & Trolle, D. (2018). Modelling the fate and transport of Cryptosporidium, a zoonotic and waterborne pathogen, in the Daning River watershed of the Three Gorges Reservoir Region, China. *Journal of Environmental Management*, *232*, 462–474. https://doi.org/10.1016/j.jenvman.2018.10.064

Lozano, V. L., Miranda, C. E., Vinocur, A. L., González, C., Unrein, F., Wolansky, M. J., & Pizarro, H. N. (2019). Turbidity matters: Differential effect of a 2,4-D formulation on the

structure of microbial communities from clear and turbid freshwater systems. *Heliyon*, *5*(8), e02221. https://doi.org/10.1016/j.heliyon.2019.e02221

Luengo, J., Fernández, A., García, S., & Herrera, F. (2011a). Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, *15*(10), 1909–1936.

Luengo, J., Fernández, A., García, S., & Herrera, F. (2011b). Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, *15*(10), 1909–1936. https://doi.org/10.1007/s00500-010-0625-8

Lyautey, E., Lapen, D. R., Wilkes, G., McCleary, K., Pagotto, F., Tyler, K., Hartmann, A., Piveteau, P., Rieu, A., & Robertson, W. J. (2007). Distribution and characteristics of Listeria monocytogenes isolates from surface waters of the South Nation River watershed, Ontario, Canada. *Applied and Environmental Microbiology*, *73*(17), 5401–5410.

Ma, J.-Y., Li, M.-Y., Qi, Z.-Z., Fu, M., Sun, T.-F., Elsheikha, H. M., & Cong, W. (2022). Waterborne protozoan outbreaks: An update on the global, regional, and national prevalence from 2017 to 2020 and sources of contamination. *Science of The Total Environment*, *806*, 150562. https://doi.org/10.1016/j.scitotenv.2021.150562

Mac Kenzie, W. R., Hoxie, N. J., Proctor, M. E., Gradus, M. S., Blair, K. A., Peterson, D. E., Kazmierczak, J. J., Addiss, D. G., Fox, K. R., Rose, J. B., & Davis, J. P. (1994). A Massive Outbreak in Milwaukee of Cryptosporidium Infection Transmitted through the Public Water Supply. *New England Journal of Medicine*, *331*(3), 161–167. https://doi.org/10.1056/NEJM199407213310304

MACDONALD, A. M., CALOW, R. C., MACDONALD, D. M. J., DARLING, W. G., & DOCHARTAIGH, B. É. Ó. (2009). What impact will climate change have on rural

groundwater supplies in Africa? *Hydrological Sciences Journal*, *54*(4), 690–703. https://doi.org/10.1623/hysj.54.4.690

Mahmud, Z. H., Islam, M. S., Imran, K. M., Hakim, S. A. I., Worth, M., Ahmed, A., Hossan, S., Haider, M., Islam, M. R., Hossain, F., Johnston, D., & Ahmed, N. (2019). Occurrence of Escherichia coli and faecal coliforms in drinking water at source and household point-of-use in Rohingya camps, Bangladesh. *Gut Pathogens*, *11*(1), 52. https://doi.org/10.1186/s13099-019-0333-6

Mainali, J., & Chang, H. (2018). Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. *Journal of Hydrology*, *564*, 26–40. https://doi.org/10.1016/j.jhydrol.2018.06.074

Mälzer, H.-J., aus der Beek, T., Müller, S., & Gebhardt, J. (2016). Comparison of different model approaches for a hygiene early warning system at the lower Ruhr River, Germany. *Safe Ruhr*, *219*(7, Part B), 671–680. https://doi.org/10.1016/j.ijheh.2015.06.005

Markó, L. (2005). Book Review: Book Review. *Chirality*, *17*(3), 168–168. https://doi.org/10.1002/chir.20146

Masina, S., Shirley, J., Allen, J., Sargeant, J. M., Guy, R. A., Wallis, P. M., Scott Weese, J., Cunsolo, A., Bunce, A., & Harper, S. L. (2018). Weather, environmental conditions, and waterborne Giardia and Cryptosporidium in Iqaluit, Nunavut. *Journal of Water and Health*, *17*(1), 84–97. https://doi.org/10.2166/wh.2018.323

Mason, B. W., Chalmers, R. M., Carnicer-Pont, D., & Casemore, D. P. (2010). A Cryptosporidium hominis outbreak in North-West Wales associated with low oocyst counts in treated drinking water. *Journal of Water and Health*, *8*(2), 299–310. https://doi.org/10.2166/wh.2009.184

Mavimbela, S. S. W., Dlamini, P., & van Rensburg, L. D. (2019). Infiltration-excess runoff properties of dryland floodplain soil types under simulated rainfall conditions. *Arid Land*

Research and Management, *33*(3), 235–254. https://doi.org/10.1080/15324982.2018.1531441

McCann, R. K., Marcot, B. G., & Ellis, R. (2006). Bayesian belief networks: Applications in ecology and natural resource management. *Canadian Journal of Forest Research*, *36*(12), 3053–3062. https://doi.org/10.1139/x06-238

McClure, M. L., Hranac, C. R., Haase, C. G., McGinnis, S., Dickson, B. G., Hayman, D. T. S., McGuire, L. P., Lausen, C. L., Plowright, R. K., Fuller, N., & Olson, S. H. (2022). Projecting the compound effects of climate change and white-nose syndrome on North American bat species. *Climate Change Ecology*, *3*, 100047. https://doi.org/10.1016/j.ecochg.2021.100047

Menti, E., Lanera, C., Lorenzoni, G., Giachino, D. F., Marchi, M. De, Gregori, D., & Berchialla, P. (2016). Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2016*, 884–893.

Mittal, S., & Maskara, S. (2011a). *A review of some Bayesian Belief Network structure learning algorithms*. https://doi.org/10.1109/ICICS.2011.6173579

Mittal, S., & Maskara, S. L. (2011b). A review of some Bayesian Belief Network structure learning algorithms. *ICICS 2011 - 8th International Conference on Information, Communications and Signal Processing*, *December*. https://doi.org/10.1109/ICICS.2011.6173579

Mohammed, H., Hameed, I., & Seidu, R. (2017). *Random forest tree for predicting fecal indicator organisms in drinking water supply* (p. 6). https://doi.org/10.1109/BESC.2017.8256398

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Data-Based Engineering Science and Technology*, *347*(11), 845–855. https://doi.org/10.1016/j.crme.2019.11.009

Muchiri, J. M., Ascolillo, L., Mugambi, M., Mutwiri, T., Ward, H. D., Naumova, E. N., Egorov, A. I., Cohen, S., Else, J. G., & Griffiths, J. K. (2009). Seasonality of Cryptosporidium oocyst detection in surface waters of Meru, Kenya as determined by two isolation methods followed by PCR. *Journal of Water and Health*, *7*(1), 67–75. PubMed. https://doi.org/10.2166/wh.2009.109

Muylaert, K., Van Der Gucht, K., Vloemans, N., Meester, L. D., Gillis, M., & Vyverman, W. (2002). Relationship between bacterial community composition and bottom-up versus top-down variables in four eutrophic shallow lakes. *Applied and Environmental Microbiology*, *68*(10), 4740–4750. PubMed. https://doi.org/10.1128/AEM.68.10.4740-4750.2002

Namugize, J. N., Jewitt, G., & Graham, M. (2018). Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. *The 17th WaterNet/WARFSA/GWPSA Symposium: Integrated Water Resources Management: Water Security, Sustainability and Development in Eastern and Africa Southern*, *105*, 247–264. https://doi.org/10.1016/j.pce.2018.03.013

National Academies of Sciences, Engineering, and Medicine. 2020. Review of the New York City Watershed Protection Program. Washington, D. T. N. A. Press. https://doi. org/10. 17226/25851. (n.d.). *No Title*.

Nevers, M. B., & Whitman, R. L. (2011). Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. *Water Research*, *45*(4), 1659–1668. https://doi.org/10.1016/j.watres.2010.12.010

Newton, A. (2009). Bayesian Belief Networks in environmental modelling: A review of recent progress. In *Environmental Modelling: New Research* (pp. 13–50).

Nicholas, R., Leslie, D., Scott, S., Casey, B., & Richard, P. (2016). Potential Impacts of Changes in Climate on Turbidity in New York City's Ashokan Reservoir. *Journal of Water Resources Planning and Management*, *142*(3), 4015066. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000614

Nojavan A., F., Qian, S. S., & Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling & Software*, *87*, 64–71. https://doi.org/10.1016/j.envsoft.2016.10.007

NYC Environmental Protection. (n.d.). https://www1.nyc.gov/site/dep/water/kensico-reservoir.page

O'Connor, D. R. (2002). *Part one: A summary: Report of the Walkerton Inquiry: The events of May 2000 and related issues*.

Odonkor, S. T., & Mahami, T. (2020). Escherichia coli as a Tool for Disease Risk Assessment of Drinking Water Sources. *International Journal of Microbiology*, *2020*, 2534130. https://doi.org/10.1155/2020/2534130

Oliveira, M., Freire, D., & Pedroso, N. (2017). Escherichia coli is not a suitable fecal indicator to assess water fecal contamination by otters. *Brazilian Journal of Biology*, *78*. https://doi.org/10.1590/1519-6984.167279

Oliver, D. M., & Page, T. (2016). Effects of seasonal meteorological variables on E. coli persistence in livestock faeces and implications for environmental and human health. *Scientific Reports*, *6*, 37101–37101. PubMed. https://doi.org/10.1038/srep37101

Omarova, A., Tussupova, K., Berndtsson, R., Kalishev, M., & Sharapatova, K. (2018). Protozoan parasites in drinking water: A system approach for improved water, sanitation

and hygiene in developing countries. *International Journal of Environmental Research and Public Health*, *15*(3), 1–18. https://doi.org/10.3390/ijerph15030495

Palani, S., Liong, S.-Y., & Tkalich, P. (2008). An ANN Application for Water Quality Forecasting. *Marine Pollution Bulletin*, *56*, 1586–1597. https://doi.org/10.1016/j.marpolbul.2008.05.021

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., & Singh, V. P. (2014). Contamination of water resources by pathogenic bacteria. *Amb Express*, *4*(1), 1–16.

Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2020a). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water Research*, *170*. https://doi.org/10.1016/j.watres.2019.115349

Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2020b). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water Research*, *170*, 115349. https://doi.org/10.1016/j.watres.2019.115349

Payment, P., & Locas, A. (2011). Pathogens in water: Value and limits of correlation with microbial indicators. *Groundwater*, *49*(1), 4–11.

Peleato, N. M., Legge, R. L., & Andrews, R. C. (2018). Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. *Water Research*, *136*, 84–94.

Petterson, S. R., & Ashbolt, N. J. (2016). QMRA and water safety management: Review of application in drinking water systems. *Journal of Water and Health*, *14*(4), 571–589. https://doi.org/10.2166/wh.2016.262

Phan, T. D., Smart, J. C. R., Stewart-Koster, B., Sahin, Oz., Hadwen, W. L., Dinh, L. T., Tahmasbian, I., & Capon, S. J. (2019). Applications of Bayesian Networks as Decision Support Tools for Water Resource Management under Climate Change and Socio-

Economic Stressors: A Critical Appraisal. *Water*, *11*(12). https://doi.org/10.3390/w11122642

Poonam, T., Tanushree, B., & Sukalyan, C. (2013). Water quality indices-important tools for water quality assessment: A review. *International Journal of Advances in Chemistry*, *1*(1), 15–28.

Price, R., & Wildeboer, D. (2017). *E. coli as an Indicator of Contamination and Health Risk in Environmental Waters*. https://doi.org/10.5772/67330

Qiu, J., Shen, Z., Leng, G., Xie, H., Hou, X., & Wei, G. (2019). Impacts of climate change on watershed systems and potential adaptation through BMPs in a drinking water source area. *Journal of Hydrology*, *573*, 123–135. https://doi.org/10.1016/j.jhydrol.2019.03.074

Reichert, P., Borchardt, D., Henze, M., Rauch, W., Shanahan, P., Somlyody, L., & Vanrolleghem, P. A. (2001). *River water quality model* (Issue 1). IWA publishing.

Reinoso, R., Torres, L. A., & Bécares, E. (2008). Efficiency of natural systems for removal of bacteria and pathogenic parasites from wastewater. *Science of The Total Environment*, *395*(2), 80–86. https://doi.org/10.1016/j.scitotenv.2008.02.039

Rezaei Tabar, V., Mahdavi, M., Heidari, S., & Naghizadeh, S. (2016). Learning Bayesian Network Structure Using Genetic Algorithm with Consideration of the Node Ordering via Principal Component Analysis. *Journal of the Iranian Statistical Society*, *15*, 45–62. https://doi.org/10.18869/acadpub.jirss.15.2.45

Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., Hebert, R. J., Olcott, E. S., Johnson, L. M., Hargrett, N. T., Blake, P. A., & Cohen, M. L. (1983). Hemorrhagic Colitis Associated with a Rare Escherichia coli Serotype. *New England Journal of Medicine*, *308*(12), 681–685. https://doi.org/10.1056/NEJM198303243081203

Rosell, R., Gomez-Codina, J., Camps, C., Maestre, J. A., Padille, J., Cantó, A., Mate, J. L., Li, S., Roig, J., Olazábal, A., Canela, M., Ariza, A., Skagel, Z., Morera-Prat, J., & Abad, A.

(1994). The New England Journal of Medicine Downloaded from nejm.org at NovartisLibrary on November 11, 2013. For personal use only. No other uses without permission. Copyright © 1994 Massachusetts Medical Society. All rights reserved. *The New England Journal of Medicine*, *330*(3), 153–158.

Rossi, A., Wolde, B., Lee, L., & Wu, M. (2020). Prediction of recreational water safety using Escherichia coli as an indicator: Case study of the Passaic and Pompton Rivers, New Jersey. *Science of The Total Environment*, *714*, 136814. https://doi.org/10.1016/j.scitotenv.2020.136814

Rutten, G., Cinderby, S., & Barron, J. (2020). Understanding Complexity in Freshwater Management: Practitioners' Perspectives in The Netherlands. *Water*, *12*(2). https://doi.org/10.3390/w12020593

S. Gwanikar, S.Cross, D.MacDonald, J.R. Brown, D.Q.Tao, & T.Berger. (1998). *Salmon River: Water Quality Assessment and Recommended Objectives*. Environment Canada-Fraser River Action Plan.

Sadik, N. J., Uprety, S., Nalweyiso, A., Kiggundu, N., Banadda, N. E., Shisler, J. L., & Nguyen, T. H. (2017). Quantification of multiple waterborne pathogens in drinking water, drainage channels, and surface water in Kampala, Uganda, during seasonal variation. *GeoHealth*, *1*(6), 258–269.

Santos, P. R. dos, & Daniel, L. A. (2017). Occurrence and removal of Giardia spp. Cysts and Cryptosporidium spp. Oocysts from a municipal wastewater treatment plant in Brazil. *Environmental Technology*, *38*(10), 1245–1254.

Sarkar, J., Prottoy, Z. H., Bari, Md. T., & Al Faruque, M. A. (2021). Comparison of ANFIS and ANN modeling for predicting the water absorption behavior of polyurethane treated polyester fabric. *Heliyon*, *7*(9), e08000. https://doi.org/10.1016/j.heliyon.2021.e08000

Sasakova, N., Gregova, G., Takacova, D., Mojzisova, J., Papajova, I., Venglovsky, J., Szaboova, T., & Kovacova, S. (2018). Pollution of Surface and Ground Water by Sources Related to Agricultural Activities. *Frontiers in Sustainable Food Systems*, *2*. https://www.frontiersin.org/article/10.3389/fsufs.2018.00042

Sato, M. I. Z., Galvani, A. T., Padula, J. A., Nardocci, A. C., de Souza Lauretto, M., Razzolini, M. T. P., & Hachich, E. M. (2013). Assessing the infection risk of Giardia and Cryptosporidium in public drinking water delivered by surface water systems in Sao Paulo State, Brazil. *Science of The Total Environment*, *442*, 389–396.

Scutari, M., Graafland, C. E., & Gutiérrez, J. M. (2019). Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, *115*, 235–253. https://doi.org/10.1016/j.ijar.2019.10.003

Seo, I. won, Yun, S. H., & Choi, S. Y. (2016). Forecasting Water Quality Parameters by ANN Model Using Pre-processing Technique at the Downstream of Cheongpyeong Dam. *12th International Conference on Hydroinformatics (HIC 2016) - Smart Water for the Future*, *154*, 1110–1115. https://doi.org/10.1016/j.proeng.2016.07.519

Singharoy, K. (2018). *Tree-Augmented Naïve Bayes Methods for Real-Time Training and Classification of Streaming Data*. 1–72.

SMITH, H. V. (1992). Cryptosporidium and water: A review. *Water and Environment Journal*, *6*(4), 443–451.

Sokolova, E., Ivarsson, O., Lilliesträm, A., Speicher, N. K., Rydberg, H., & Bondelind, M. (2022a). Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data. *Science of The Total Environment*, *802*, 149798. https://doi.org/10.1016/j.scitotenv.2021.149798

Sokolova, E., Ivarsson, O., Lilliesträm, A., Speicher, N. K., Rydberg, H., & Bondelind, M. (2022b). Data-driven models for predicting microbial water quality in the drinking water

source using E. coli monitoring and hydrometeorological data. *Science of The Total Environment*, *802*, 149798. https://doi.org/10.1016/j.scitotenv.2021.149798

Solgi, A., Pourhaghi, A., Bahmani, R., & Zarei, H. (2017). Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD). *Ecohydrology & Hydrobiology*, *17*(2), 164–175. https://doi.org/10.1016/j.ecohyd.2017.02.002

Staley, C., Reckhow, K. H., Lukasik, J., & Harwood, V. J. (2012). Assessment of sources of human pathogens and fecal contamination in a Florida freshwater lake. *Water Research*, *46*(17), 5799–5812. https://doi.org/10.1016/j.watres.2012.08.012

Stegen, J. C., Lin, X., Konopka, A. E., & Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *The ISME Journal*, *6*(9), 1653–1664. https://doi.org/10.1038/ismej.2012.22

Swaffer, B., Abbott, H., King, B., van der Linden, L., & Monis, P. (2018). Understanding human infectious Cryptosporidium risk in drinking water supply catchments. *Water Research*, *138*, 282–292. https://doi.org/10.1016/j.watres.2018.03.063

Sylvestre, É., Burnet, J. B., Dorner, S., Smeets, P., Medema, G., Villion, M., Hachad, M., & Prévost, M. (2021). Impact of Hydrometeorological Events for the Selection of Parametric Models for Protozoan Pathogens in Drinking-Water Sources. *Risk Analysis*, *41*(8), 1413–1426. https://doi.org/10.1111/risa.13612

Taheri Tizro, A., Ghashaghaie, M., & Georgiou, P. (2014). Time series analysis of water quality parameters. *Journal of Applied Research in Water and Wastewater*, *1*, 43–52.

Tang, J., McDonald, S., Samadder, S., Murphy, T., & Holden, N. (2011). Modelling Cryptosporidium oocysts transport in small ungauged agricultural catchments. *Water Research*, *45*, 3665–3680. https://doi.org/10.1016/j.watres.2011.04.013

Tarr, Joel A. (1996). *The Search for the Ultimate Sink: Urban Pollution in Historical Perspective.*

Taylor, R., Miret-Gaspa, M., Tumwine, J., Mileham, L., Flynn, R., Howard, G., & Kulabako, R. (2009). *Increased risk of diarrhoeal diseases from climate change: Evidence from urban communities supplied by groundwater in Uganda.* 15–19.

Telci, I. T., Nam, K., Guan, J., & Aral, M. M. (2009). Optimal water quality monitoring network design for river systems. *Journal of Environmental Management*, *90*(10), 2987–2998.

Tolouei, S., Burnet, J.-B., Autixier, L., Taghipour, M., Bonsteel, J., Duy, S. V., Sauvé, S., Prévost, M., & Dorner, S. (2019). Temporal variability of parasites, bacterial indicators, and wastewater micropollutants in a water resource recovery facility under various weather conditions. *Water Research*, *148*, 446–458. https://doi.org/10.1016/j.watres.2018.10.068

Tornevi, A., Bergstedt, O., & Forsberg, B. (2014). Precipitation Effects on Microbial Pollution in a River: Lag Structures and Seasonal Effect Modification. *PloS One*, *9*, e98546. https://doi.org/10.1371/journal.pone.0098546

Tousi, E. G., Duan, J. G., Gundy, P. M., Bright, K. R., & Gerba, C. P. (2021). Evaluation of E. coli in sediment for assessing irrigation water quality using machine learning. *Science of The Total Environment*, *799*, 149286. https://doi.org/10.1016/j.scitotenv.2021.149286

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78. https://doi.org/10.1007/s10994-006-6889-7

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

Usepa. (2005). Method 1623: *Control*, *December*.

USEPA (1986). (n.d.). *Ambient Water Quality Criteria for Bacteria. Washington, DC: USEPA.*

Uusitalo, L. (2007a). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3–4), 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Uusitalo, L. (2007b). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3–4), 312–318.

Vasquez, J., Maier, H., Lence, B., Tolson, B., & Foschi, R. (2000). Achieving Water Quality System Reliability Using Genetic Algorithms. *Journal of Environmental Engineering-Asce - J ENVIRON ENG-ASCE*, *126*. https://doi.org/10.1061/(ASCE)0733-9372(2000)126:10(954)

Wallis, P., Bounsombath, N., Brost, S., Appelbee, A., & Clark, B. (2003). Outbreak of Waterborne Cryptosporidiosis at North Battleford, SK, Canada. *Cryptosporidium: From Molecules to Disease*, *3*, 341–344. https://doi.org/10.1016/B978-044451351-9/50047-1

Water Quality Assesssment of Peace River Above Alces River. (2003). BWP Consulting.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J., & Roderick, P. J. (1999). General practice study of infectious intestinal disease in England: And reported to national surveillance. *Bmj*, *318*(April), 1046–1050.

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., Neumann, N., Ruecker, N., Topp, E., & Lapen, D. R. (2009). Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. *Water Research*, *43*(8), 2209–2223. https://doi.org/10.1016/j.watres.2009.01.033

Wohlsen, T., Bates, J., Gray, B., Aldridge, P., Stewart, S., Williams, M., & Katouli, M. (2006a). The occurrence of Cryptosporidium and Giardia in the lake Baroon catchment, Queensland, Australia. *Journal of Water Supply: Research and Technology—AQUA*, *55*(5), 357–366.

Wohlsen, T., Bates, J., Gray, B., Aldridge, P., Stewart, S., Williams, M., & Katouli, M. (2006b). The occurrence of Cryptosporidium and Giardia in the Lake Baroon catchment, Queensland, Australia. *Journal of Water Supply: Research and Technology - AQUA*, *55*(5), 357–366. https://doi.org/10.2166/aqua.2006.044

Woolf, B. P. (2009). Chapter 7—Machine Learning. In B. P. Woolf (Ed.), *Building Intelligent Interactive Tutors* (pp. 221–297). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-373594-2.00007-1

Wu, J., Rees, P., Storrer, S., Alderisio, K., & Dorner, S. (2009). Fate and Transport Modeling of Potential Pathogens: The Contribution From Sediments 1. *JAWRA Journal of the American Water Resources Association*, *45*(1), 35–44.

Xiang, Y., & Jiang, L. (2009). *Water quality prediction using LS-SVM and particle swarm optimization*. 900–904.

Xiao, S., Hu, S., Zhang, Y., Zhao, X., & Pan, W. (2018). Influence of sewage treatment plant effluent discharge into multipurpose river on its water quality: A quantitative health risk assessment of Cryptosporidium and Giardia. *Environmental Pollution*, *233*, 797–805. https://doi.org/10.1016/j.envpol.2017.11.010

Xiao, S., Yin, P., Zhang, Y., Zhao, X., Sun, L., Yuan, H., Lu, J., & Sike, hu. (2018). Occurrence, genotyping, and health risk of Cryptosporidium and Giardia in recreational lakes in Tianjin, China. *Water Research*, *141*. https://doi.org/10.1016/j.watres.2018.05.016

Xu, L., & Liu, S. (2013). Study of short-term water quality prediction model based on wavelet neural network. *Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012*, *58*(3), 807–813. https://doi.org/10.1016/j.mcm.2012.12.023

Xu, T., Coco, G., & Neale, M. (2020). A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Research*, *177*, 115788. https://doi.org/10.1016/j.watres.2020.115788

Yan, L. J., & Cercone, N. (2010). Bayesian network modeling for evolutionary genetic structures. *Computers & Mathematics with Applications*, *59*(8), 2541–2551. https://doi.org/10.1016/j.camwa.2009.12.039

Yang, X., Liu, Q., Luo, X., & Zheng, Z. (2017). Spatial Regression and Prediction of Water Quality in a Watershed with Complex Pollution Sources. *Scientific Reports*, *7*(1), 8318–8318. PubMed. https://doi.org/10.1038/s41598-017-08254-w

Young, I., Smith, B. A., & Fazil, A. (2015). A systematic review and meta-analysis of the effects of extreme weather events and other weather-related variables on Cryptosporidium and Giardia in fresh surface waters. *Journal of Water and Health*, *13*(1), 1–17. https://doi.org/10.2166/wh.2014.079

Yu, R., & Zhang, C. (2021). Early warning of water quality degradation: A copula-based Bayesian network model for highly efficient water quality risk assessment. *Journal of Environmental Management*, *292*, 112749. https://doi.org/10.1016/j.jenvman.2021.112749

Zarei, A., & Mohammadzadeh Asl, B. (2020). Performance evaluation of the spectral autocorrelation function and autoregressive models for automated sleep apnea detection using single-lead ECG signal. *Computer Methods and Programs in Biomedicine*, *195*, 105626. https://doi.org/10.1016/j.cmpb.2020.105626

Zhan, S., Zhou, B., Li, Z., Li, Z., & Zhang, P. (2021). Evaluation of source water quality and the influencing factors: A case study of Macao. *Physics and Chemistry of the Earth, Parts A/B/C*, *123*, 103006. https://doi.org/10.1016/j.pce.2021.103006

Zhu, Z., Broersma, K., & Mazumder, A. (2012). Impacts of Land Use, Fertilizer and Manure Application on the Stream Nutrient Loadings in the Salmon River Watershed, South-Central British Columbia, Canada. *Journal of Environmental Protection*, *03*, 809–822. https://doi.org/10.4236/jep.2012.328096

.

Appendices

Appendix A : Supplementary Data for Chapter 3

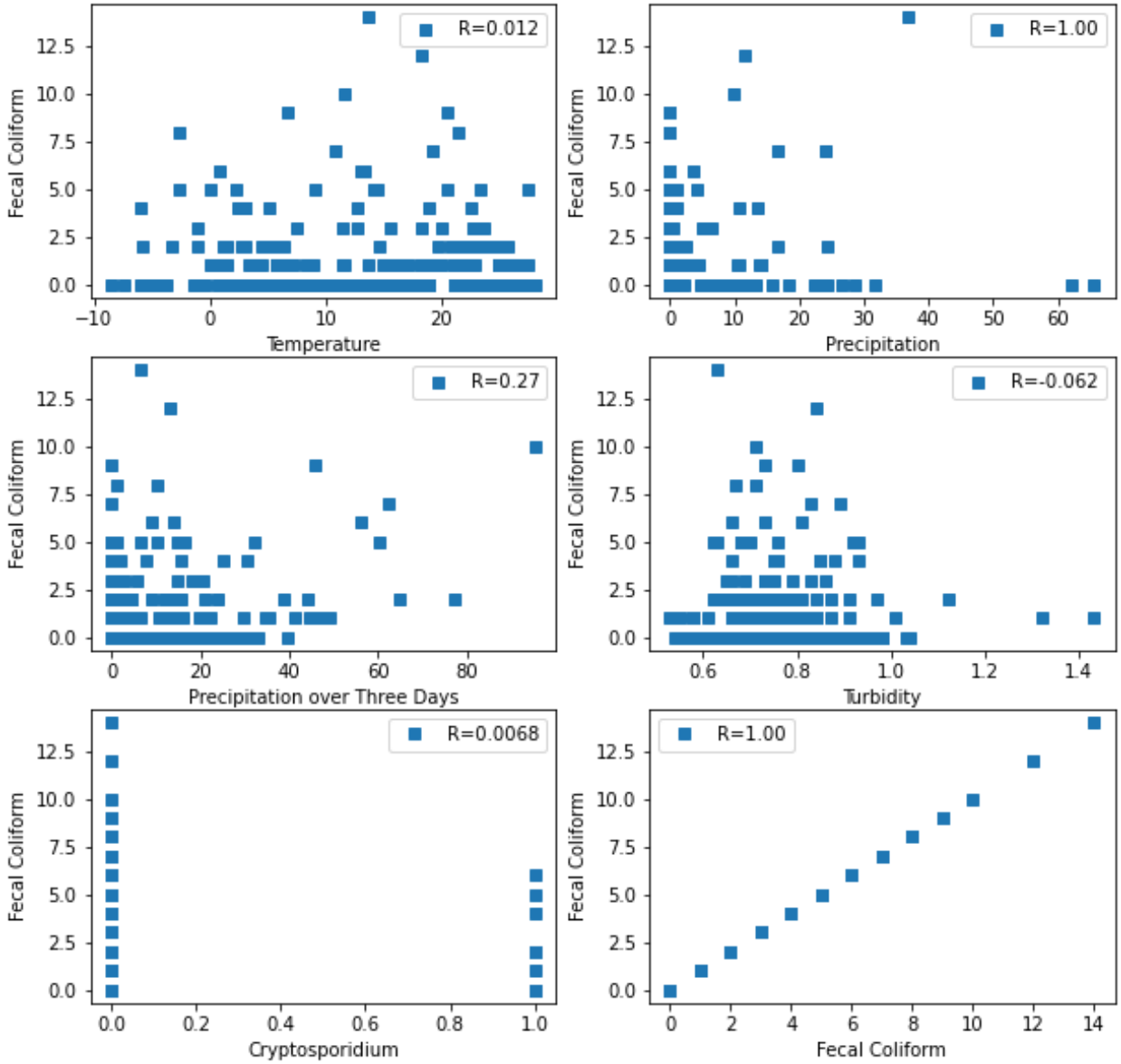


Figure S1. Scatterplot of fecal coliform on sampling day with other parameters of Kensico Reservoirs.

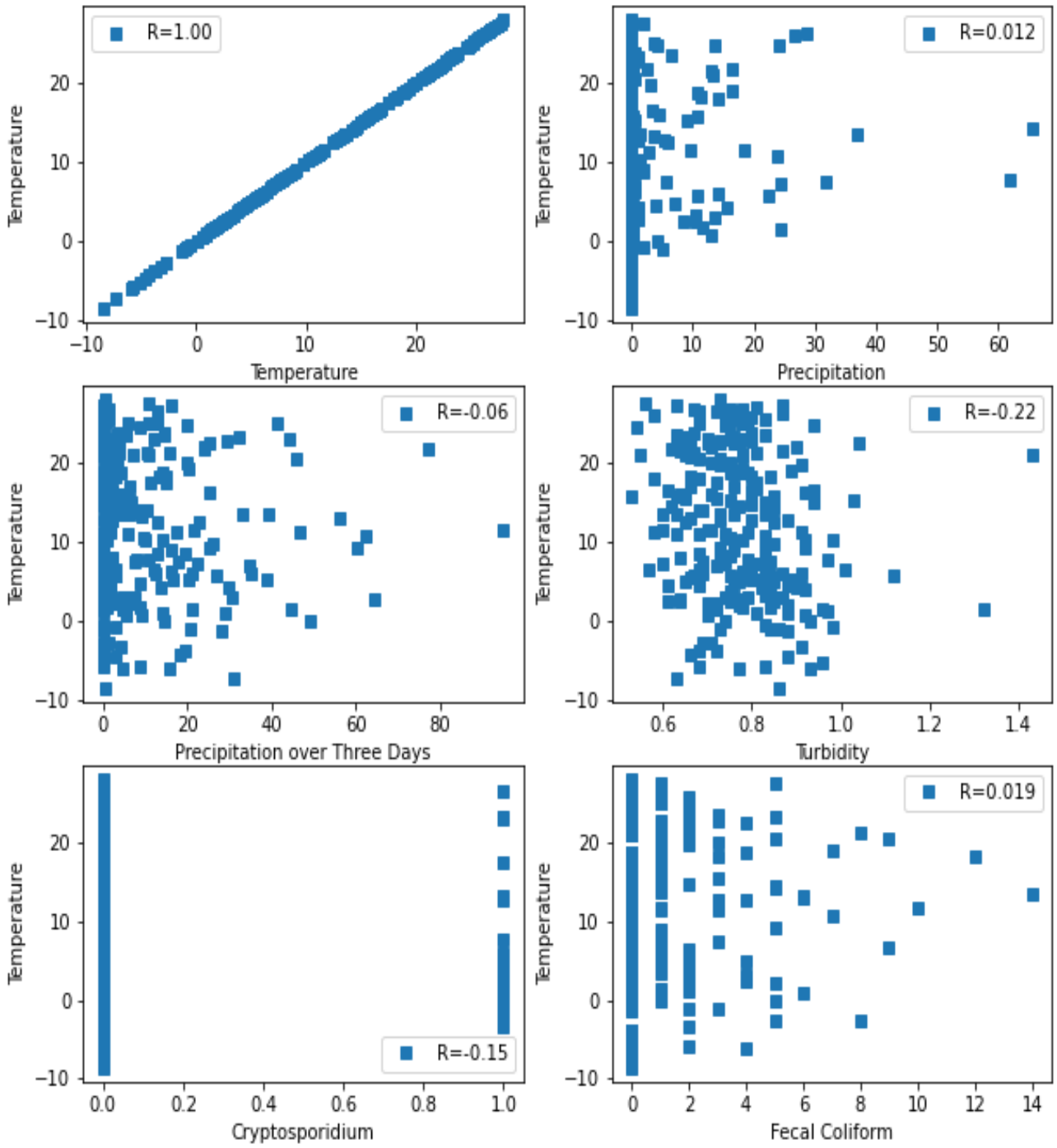


Figure S2. Scatterplot of temperature with other parameters of Kensico Reservoirs.

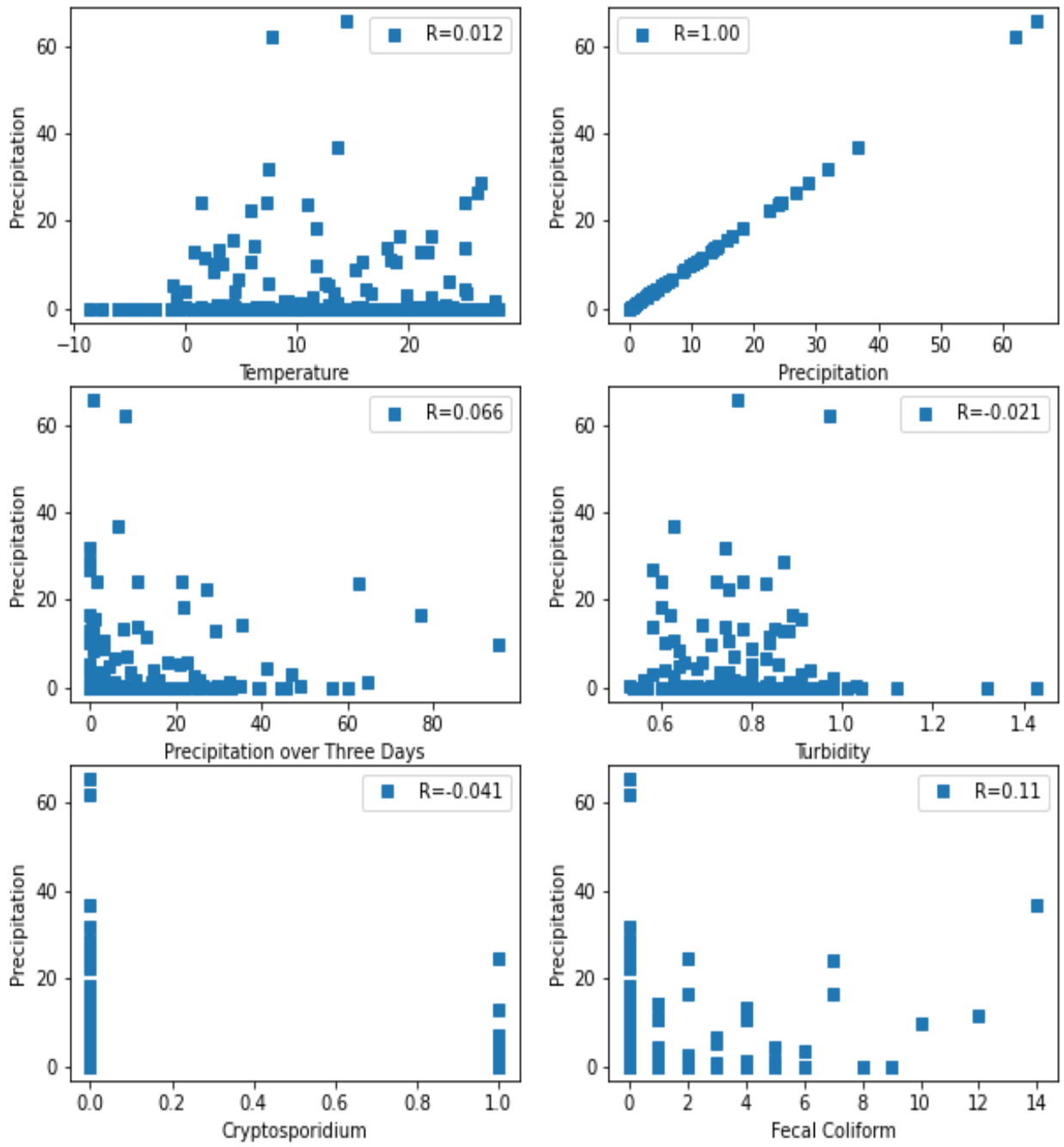


Figure S3. Scatterplot of precipitation with other parameters of Kensico Reservoirs.

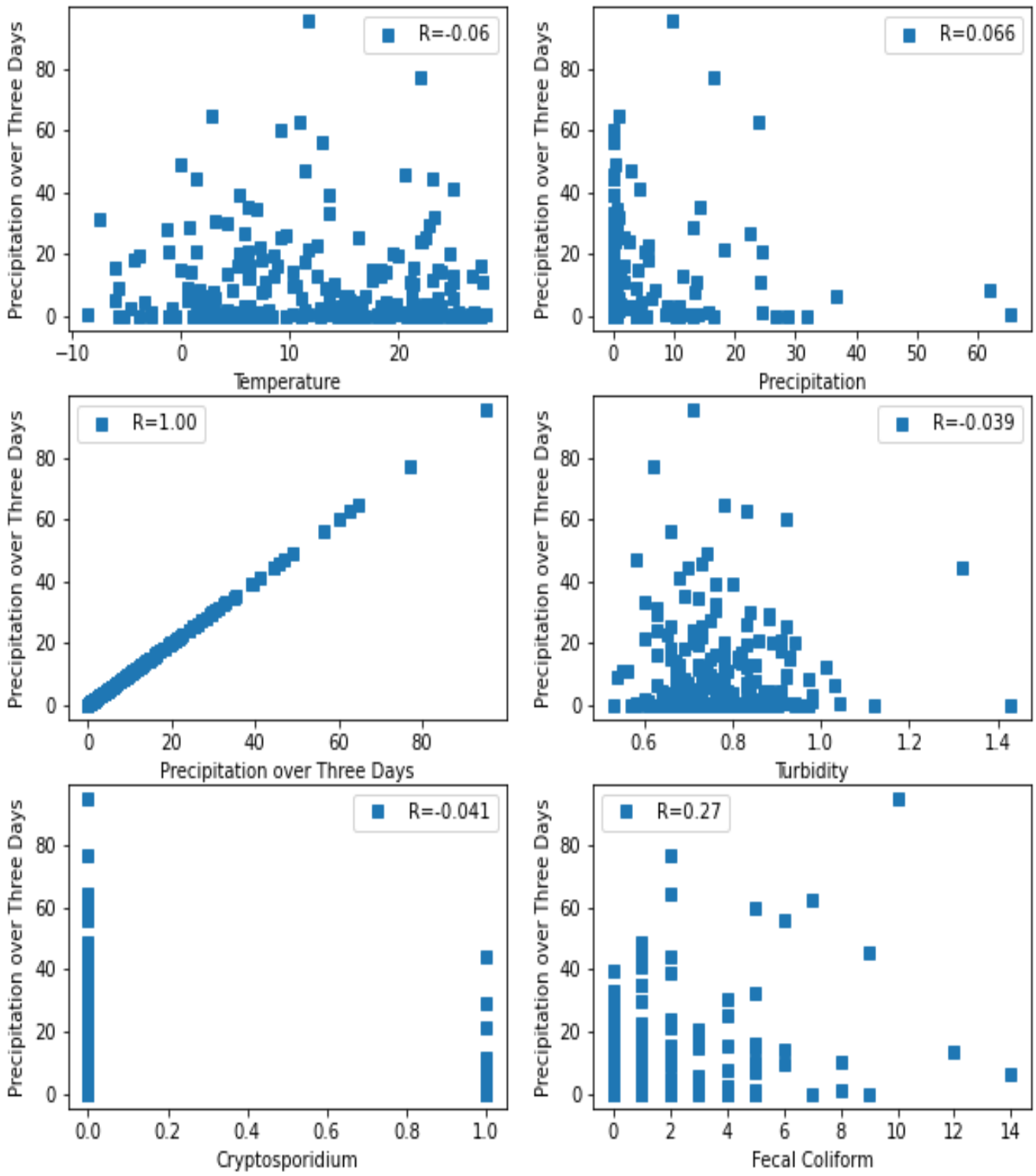


Figure S4. Scatterplot of precipitation over three days with other parameters of Kensico Reservoirs.

Table S.1. Prediction accuracy of *Cryptosporidium* with simple Model (Logistic Regression)

	(a) Test data used in balancing			(b) Test data not used in balancing		
Model	**Overall Accuracy**	**Prediction Accuracy of Absence**	**Prediction Accuracy of Presence**	**Overall Accuracy**	**Prediction Accuracy of Absence**	**Prediction Accuracy of Presence**
ADA_ Linear Regressio	60%	55%	72%	78%	89%	10%
SMOT-Linear Regression	60%	55%	72%	62%	67%	33%

Appendix B : Supplementary Data for Chapter 4

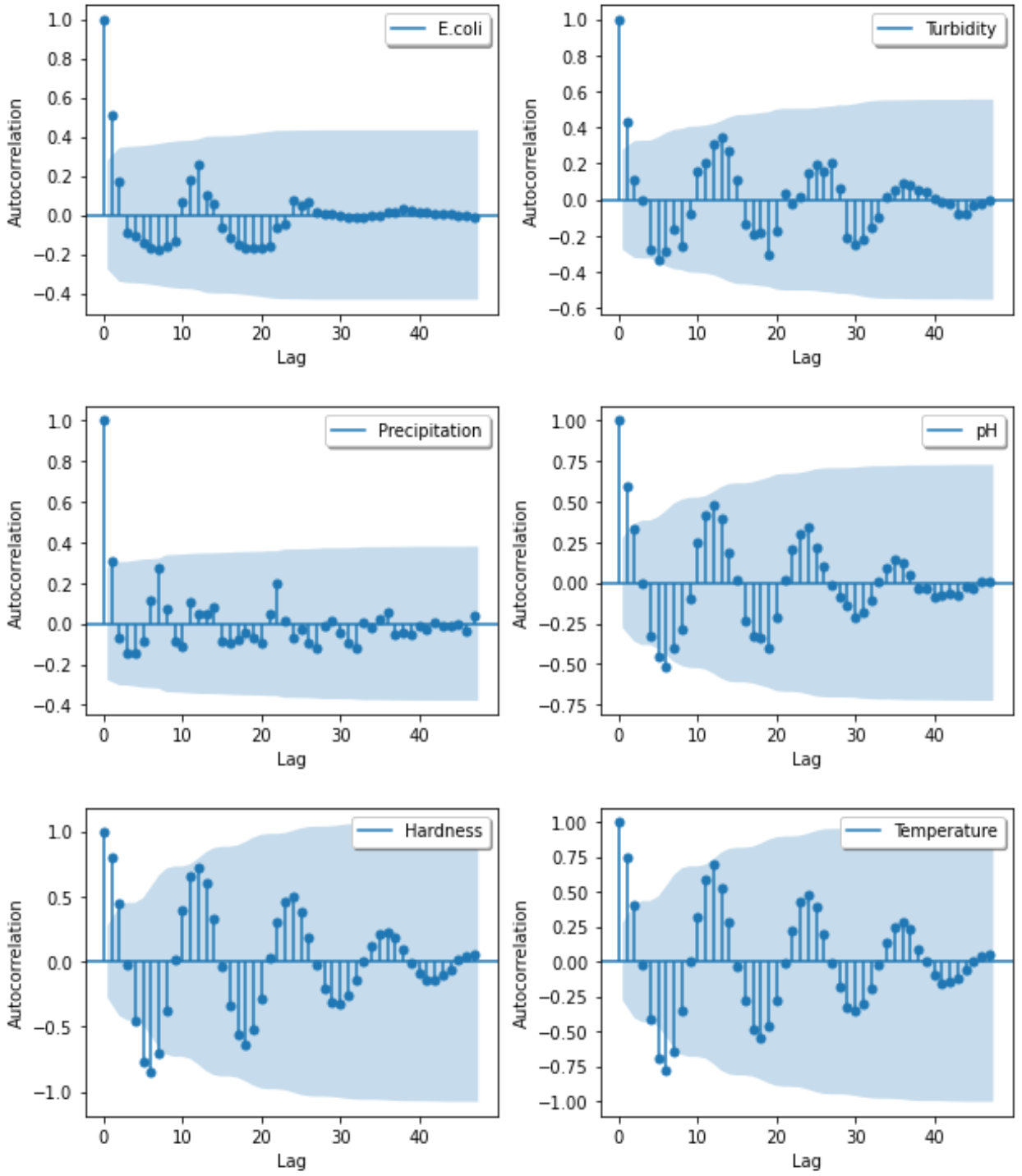


Figure S5. Autocorrelation of all Cheakamus River parameters based on one month time lag and confidence interval of 95%.

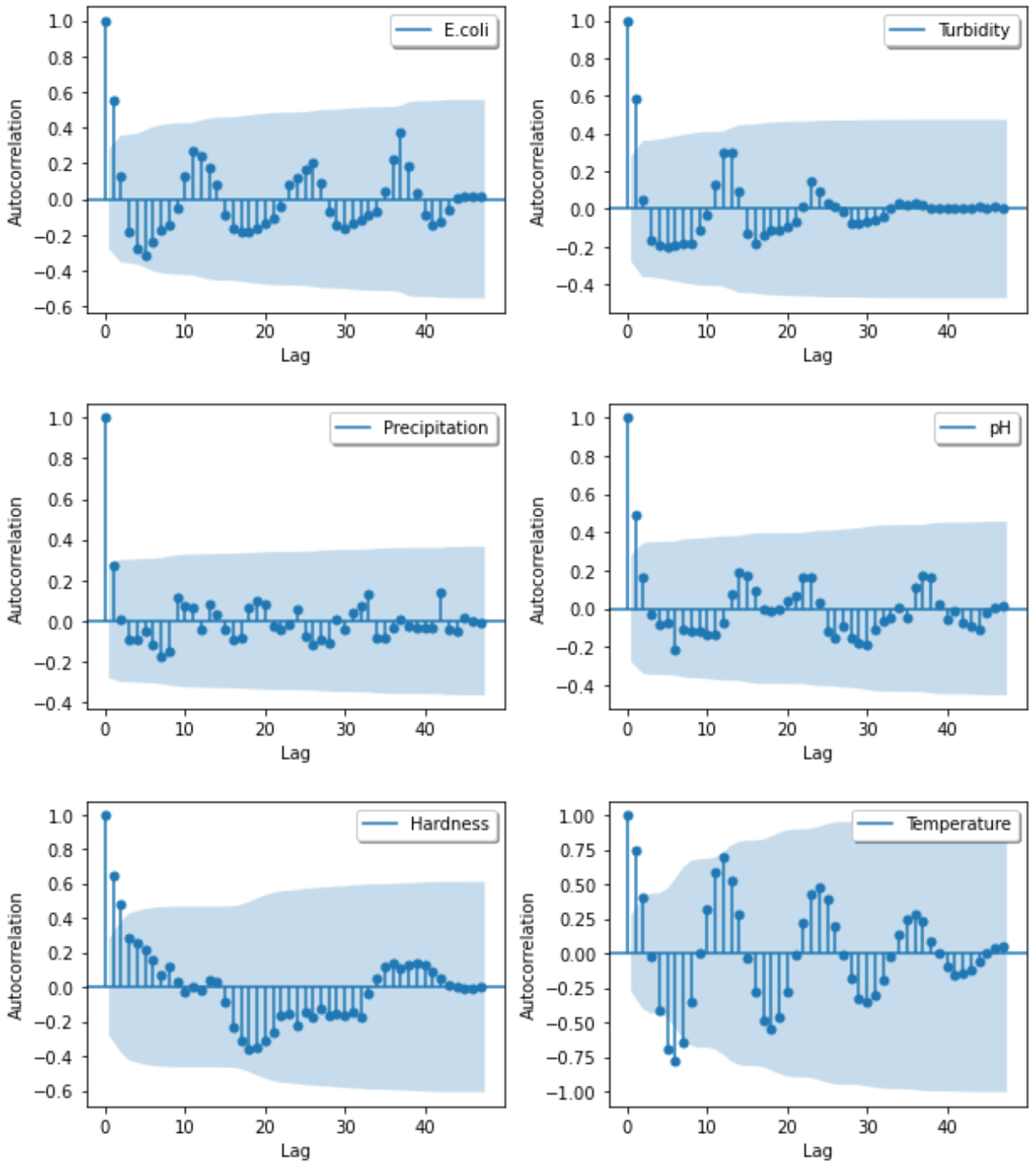


Figure S6. Autocorrelation of all Peace River parameters based on one month time lag and confidence interval of 95%.

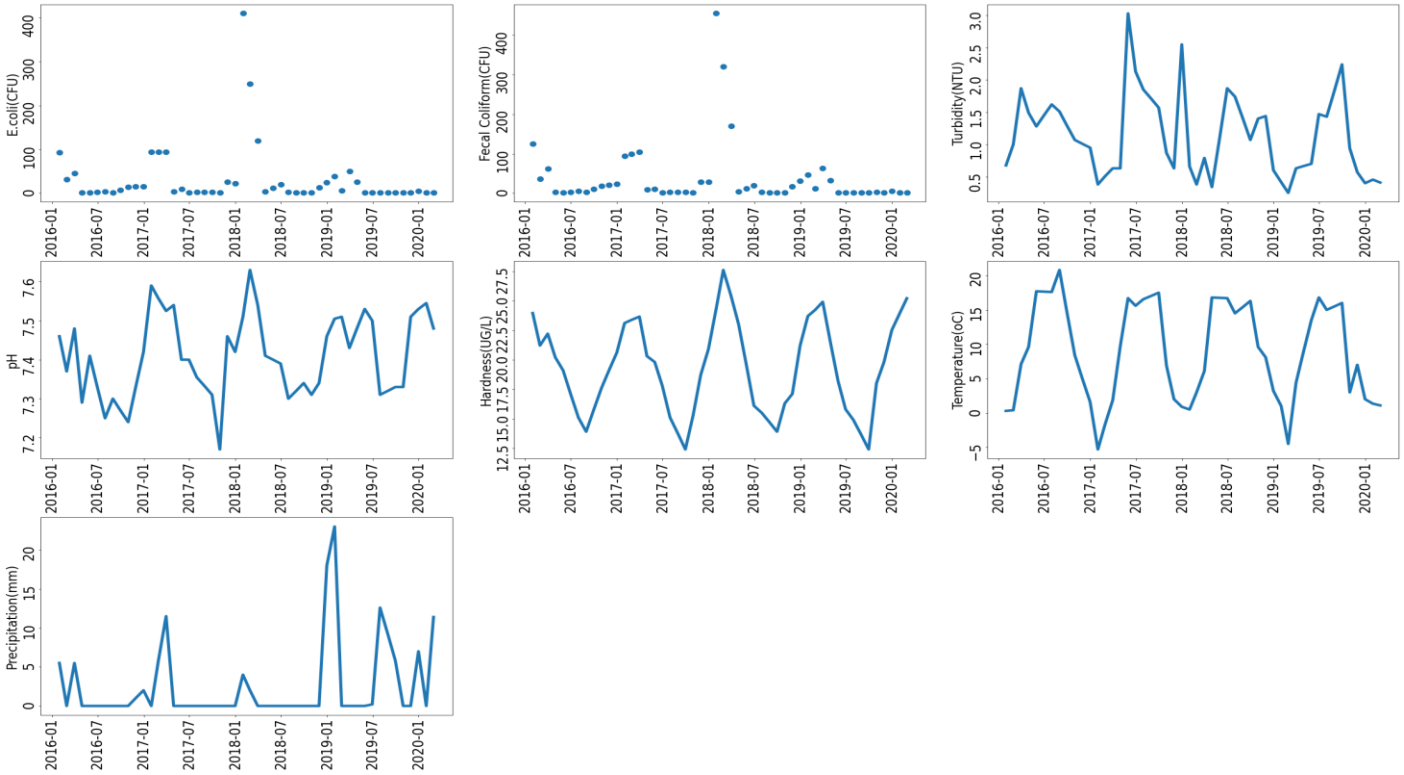


Figure S7. Monthly time series of all water quality and weather parameters of Cheakamus River. The time intervals are equal to one month.

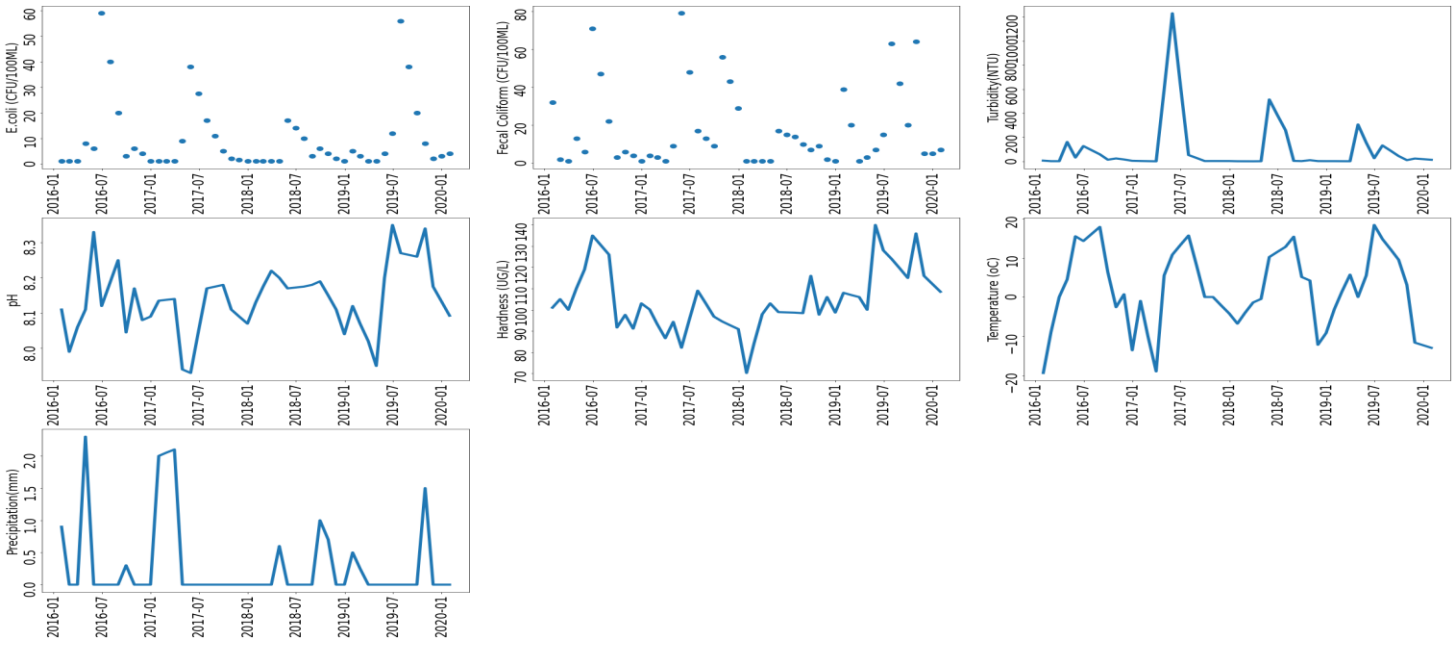


Figure S8. Monthly time series of all water quality and weather parameters of Peace River. The time intervals are equal to one month.

Table S.2. Prediction accuracy of *E. coli* with simple Model (Logistic Regression)

Model	Location	Overall Accuracy	Prediction Accuracy of *E. coli*<20 CFU/100	Prediction Accuracy of *E. coli*>20 CFU/100
Logistic Regression	Cheakamus River	71%	21%	97%
	Peace River	88%	92%	50%
	Salmon River	72%	2%	99%