ARTIFICIAL NEURAL NETWORK BASED NOWCASTING MODEL FOR BEACH WATER QUALITY

By

Jainy Mavani

Bachelor of Engineering in Civil Engineering,

Gujarat University, India, 2005

A thesis presented to Ryerson University

In partial fulfillment of the requirements of

Master of Applied Science

In the program of

Civil Engineering

Toronto, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

ARTIFICIAL NEURAL NETWORK BASED NOWCASTING MODEL FOR BEACH WATER QUALITY

By

Jainy Mavani

Master of Applied Science 2014

Department of Civil Engineering, Ryerson University

Abstract

Recreational water users may be exposed to elevated pathogen levels that originate from various point and non-point sources. Current daily notifications practice depends on microbial analysis of indicator organisms such as Escherichia coli (E. coli) that require 18-24 hours to provide sufficient response. This research evaluated the use of Artificial Neural Networks (ANNs) for real time prediction of E. coli concentration in water at Toronto beaches (Ontario, Canada). The nowcasting models were developed in combination with readily available real-time environmental and hydro-meteorological data during the bathing season (June-August) of 2008 to 2012. The results of the developed ANN models were compared with historic data and found that the predictions of E. coli concentrations generated by ANN models slightly outperforms than currently used persistence model with better accuracy. The best performing ANN models for each beach are able to predict approximately 74% to 82% of the E. coli concentrations.

Acknowledgements

This thesis could not have succeeded without the love and support of my dearest family and friends. I would like to express my deepest gratitude to my supervisors Dr. Songnian Li and Dr. Darko Joksimovic for their guidance, support and patience during the development of this thesis.

I would like to thank Dr. Darko Joksimovic for his willingness to always help along with patience, kind words, inspiration and belief in me, which pulled me through this thesis. His countless proof readings, corrections and support made this document possible.

I would like to acknowledge special thanks to Mr. Mahesh Patel from Toronto Public Health for his collaboration in model development which was crucial to the completion of this thesis.

Finally, I would like to express my special thanks to my family for their endless love and patience throughout my study.

Dedications

To my husband Maulin, for his understating and encouragement he has given

To my parents, for the lifetime of support

And

To my son Dev, for his sacrifices

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Networks
CFU	Colony Forming Units
CSO	Combined Sewer Overflow
E. coli	Escherichia Coli
GTA	Greater Toronto Area
GUI	Graphical User Interface
MLR	Multiple Linear Regression
MOE	Ministry of the Environment
NNTOOL	Neural Network Toolbox (MATLAB Mathworks Inc.)
PWQO	Provisional Water Quality Objective
RMSE	Root Mean Squared Error
TPH	Toronto Public Health
TRCA	Toronto and Region Conservation Authority
USEPA	U.S. Environmental Protection Agency

1 INTRODUCTION

1.1 Background

Beaches are treasured natural resources that provide significant value, including recreational benefits in summer time. However, beach waters can hold various pathogenic micro-organisms which are a potential threat to human health and can cause beach closures for particular period. Due to the same reason, there has been an increasing interest for the last couple of decades to develop the models for fast assessment of beach water quality.

Beach activities are affected if Escherichia coli (E. coli) concentrations in beach water is found to be higher than the required standard. E. coli is found in the faeces of both humans and animals and is used as an indicator of contamination. E. coli is often used to be a sign of the presence of fecal wastes and other harmful bacteria in lakes and streams (MOE, 1994). E. coli enters waterways via a variety of sources including sewer systems, septic systems, wildlife, livestock, pets, waterfowl and organic fertilizers. Beaches are subjected to certain sources of E. coli contamination, such as geographic location, extent of enclosure, or presence of a stromwater or creek outfall. Numerous factors may explain fluctuations in E. coli concentrations, including rainfall, wind speed and direction, wave height, turbidity, direction of flow and biological factors (Nevers et al., 2009).

In Ontario, the microbiological quality of beach water is considered by measuring the geometric mean concentrations of the E. coli during the swimming season, which is typically during June to August of each year. Geometric mean concentration of E. coli is the average of logarithmic values of a data set, which is converted back to a base 10 number (Costa, 2013). It is measured at representative beach locations and has been the base that establishes whether water

quality meets/exceeds the safety level for recreational purposes. However, due to the time required to obtain culture-based results (18-24 hours), E. coli counts are typically not available until the day next the day of the actual sample collection. The delay, attached with the temporal and spatial variability connected with E. coli, sometimes results in unnecessary beach closures or the lack of a swimming advisory when E. coli counts are, in fact, elevated and a public health risk exists (Amaral, 2010).

Various activities have been developed to address this time lag problem, including attempts to shorten analysis time for water quality monitoring, use of quicker predictive models and communicating beach water quality information to the public on a timely (e.g., near-daily) basis so more informed decisions can be made by the public regarding recreational water use. The government agencies have been dealing with an ongoing problem on how to provide for a seasonal assessment of a beach site versus guidance for day-to-day management (Ashbolt et al., 2010). Extensive efforts have been made to develop predictive models for nowcasting and forecasting the concentrations of E. coli around the world for beach condition. Nowcasting refers the current situation and what changes to expect over the next 2-6 hours. Nowcast systems operate continuously with little user intervention, which is appropriate for the time scales of the phenomena of interest. In contrast, forecasting is a periodic process, typically done two or four times per day, with significant interpretation and product development by a trained forecaster (Frick et al., 2008). Nowcast system does not reduce the importance of the forecaster rather it allows the forecaster to concentrate on the larger scale and the impact on the longer-term planning cycle.

Despite the growing use of Artificial Neural Networks (ANN) which are computational models inspired by animal central nervous system that are capable of machine learning and

pattern recognition, in water resources applications (Wikipedia, 2013, Varma and Vijayan, 2009, Motamarri and Boccelli, 2012), little work involving the use of ANNs for the prediction of indicator organisms in freshwater beaches is reported in the literature. Such research could provide a useful predictive tool, but also has the potential to offer insight into the processes controlling the generation, fate and transport of E. coli contaminants (Mass and Ahlfeld, 2007). These are issues of interest to those involved in beach water management as well as the protection of public health, source and recreational water protection.

ANN can be represented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. Each artificial neuron forms a computational node in the larger ANN. Like their biological counterparts, artificial neurons receive input from another neuron or an external stimulus, process the input signal using an activation or transfer function, and generate a transformed output signal, which may be the input to another node or the final output from the network (Basheer and Hajmeer, 2000). During the process of ANN model training, which is analogous to the calibration of process-based models, the weighted connections between the artificial neurons are updated to minimize model error. ANN models are available in several commercial software packages e.g., MATLAB, Neural Ware as well as various shareware and freeware programs that can be run on desktop personal computers.

1.2 Statement of Problem

Swimmable beaches are often used as an indicator of Toronto's environmental performance and quality of life (City of Toronto, 2009). According to current practice for assessment of beach water quality in Greater Toronto Area (GTA) it takes 18-24 hours to get the results by using the

"persistence model - using yesterday's E. coli to predict today's conditions", which is inadequate for communicating same-day water quality risks to the public to prevent exposure (USEPA, 2010). In the persistence model (laboratory test based model), samples are collected daily at swimmable beaches by the Toronto Public Health (TPH) department, City of Toronto staff and sent for analyses. TPH calculates the geometric mean by substituting the value of the detection limit for samples reported as having an E. coli concentration less than the detection limit. For example, if the results from five sampling stations at a beach were <10, 20, 15, <10 and 20, the geometric mean would be calculated using the following values: 10, 20, 15, 10 and 20. Bacteria analyses in the laboratory take about 24 hours to get the complete results. When E. coli is found in water samples at concentrations greater than 100 E. coli per 100 milliliters of water, the beaches are posted with advisory signs because swimming could lead to health effects such as skin rashes or gastrointestinal illnesses (Amaral, 2010). TPH department determines the public health implications of the bacteria data, posts the result on their website (http://app.toronto.ca/tpha/beaches.html) and conveys this information to the jurisdiction that manages a particular beach; in most cases city's parks department. Based on the data, the beach manager might post signs advising the public of the increased risk of getting sick from swimming at a particular beach (MOE, 1994). The information used to issue closure notification on any day is often based on sample data from the previous day. Conversely, elevated indicator densities detected in the current day's sample might no longer be present when the analytical results are received, resulting in unnecessary closure and an unjustified adverse economic effect, as E. coli numbers in water sources can change significantly over much shorter time periods (Zhang et al., 2012).

The Ontario Ministry of the Environment (MOE) has made attempts to improve the process of posting beach status and to reduce unnecessary and incorrect postings by developing statistical models. Using criteria for beach selection which was based on analysis of measured E. coli sampling results and dates of beach posting provided by the MOE, 3-5 Toronto beaches were selected for development of Multiple Linear Regression (MLR) models.

1.3 Objective of Research

The objective of this thesis is to develop an ANN based predictive model to nowcast beach status using the readily available explanatory variables data. In order to accomplish this objective, the following issues need to be addressed:

- Exploring the available predictive models and software tools,
- Exploring the correlations between indicator organism concentrations and other water quality and meteorological variables,
- Developing ANN models to forecast E. coli concentration for Toronto's three beaches during summer season,
- Investigating the influence of different input parameter selection methods on ANN model development and performance,
- Investigating the influence of variable transformation on ANN model performance,
- Assessing the use of varied performance criteria for ANN model development for target parameters with significant variability, such as indicator organism concentrations,
- Comparing ANN model performance to that achieved by the current practice.

1.4 Significance of Research

The delay, coupled with the temporal and spatial variability associated with E. coli, sometimes results in unwarranted beach closures or the lack of a swimming advisory. Awareness and concern about the limitations of assuming prior day E. coli concentrations to accurately reflect current day conditions, has resulted in growing interest and research on predictive modelling to estimate concentrations of this widely used fecal indicator organism based on meteorological, hydrologic and other environmental explanatory variables.

The developed multilayer feedforward ANN model for prediction of E. coli concentration for different beaches of Toronto, Ontario is very inexpensive, simple and could potentially be easily used on daily base or at any specific time necessitated by adverse weather conditions that could affect the water quality at Toronto beaches.

1.5 Thesis Outline

Chapter 1 describes the objectives and the scope of the research with brief introduction to the research topics.

Chapter 2 presents an overview of health concerns and beach water quality monitoring and discusses factors to consider when designing a predictive model. The chapter also addresses several topics related to E. coli concentration modelling and ANN.

Chapter 3 provides the knowledge regarding the ANN's basics, its functionality, types of ANNs, and limitations.

Chapter 4 describes the required data, sources and the analysis of for ANN models development.

Chapter 5 explains the basic theory behind the methodology using MATLAB based ANN tool to develop nowcast models for E. coli concentration at Toronto beaches.

Chapter 6 presents the results and discussion of the performance of developed models.

Chapter 7 provides the thesis summary and conclusions reached in present study. It also outlines possible implementations and recommendations for further research work on this topic.

2 LITERATURE REVIEW

2.1 Introduction

Fecal contaminations of water have always been a large area of concern. Locating the origin of the source of the pollutants of water is a very difficult task. Bacteria associated gastrointestinal illness is the most widely studied diseases caused by unsafe recreational water. Although, since 1990s, viral and protozoan pathogens have gained attention as areas of potential concern (MOE, 1994). Contamination due to fecal matters is a threat to human health and is a problem world-wide. E. coli is a large and diverse group of bacteria that are commonly found in the intestines of warm blooded animals (He and He, 2008).

Clearly indentifying the goals is the first step in designing a time-relevant beach water quality and public notification model. The ultimate goals are to protect public health from potential health risks associated with use of contaminated beach waters and to notify members of the public who use these waters of any potential risks with the help of new predictive model. This literature review first presents a brief summary of health concerns and beach water quality monitoring, then discusses factors to be considered while designing a predictive model. Several topics related to E. coli concentration modelling and ANN is addressed at the end of the chapter.

2.2 Public Health and Beach Water Quality

The organisms such as bacteria, viruses and protozoa which can cause disease to human are generally referred as pathogens. These pathogens normally come from feces of human and warm blooded animals. Beach water users can be exposed to pathogens mainly through ingestion of contaminated water or through skin contact (Enns et al., 2012). If taken into the body, pathogens

can cause various illnesses and, on rare occasions, even death. Waterborne illnesses include diseases resulting from bacterial infection (such as cholera, salmonellosis, and gastroenteritis), viral infection (such as infectious hepatitis, gastroenteritis, and intestinal diseases), and protozoan infections (such as amoebic dysentery and giardiasis) (Cabral, 2010). Conventional beaches and recreational water quality monitoring often relies on the "indicator organisms" to measure the likelihood of the presence or absence of pathogens (Reicherts and Emerson, 2010).

The Canadian Walkerton Inquiry highlights the dangers of waterborne transmission of pathogens such as E. coli O157:H7. Significant morbidity and seven fatalities occurred when Walkerton's municipal water supply became contaminated with Campylobacter and E. coli O157:H7. It was presumed that the contamination arose from farm animal run-off into a shallow well, from which the water supply was taken (Kinzelman and Mcphail, 2012).

2.2.1 Indicator Organisms

It is merely impossible to test every pathogen, such as Salmonella, Shigella, Yersinia, S.aureus and C.botulinus as most are difficult and time-consuming to detect and culture and provides the rationale to use indicator organisms (Nevers and Boehm., 2011). A group of organisms known as the coliform bacteria have been popularly used to indicate fecal contamination of water. This is because they inhabit the intestinal tract in high numbers. They also generally live longer than disease-causing bacteria, so an absence of coliform bacteria can indicate that the water is safe. The coliform bacteria are defined as facultativaly anaerobic, gram negative, non-spore-forming, rod-shaped bacteria that produce gas upon lactose fermentation within 48 hours at 35°C (Kinzelman and Mcphail, 2012). The coliform group, as shown in Figure 2-1, includes organisms of fecal and non-fecal origin; therefore a more restrictive definition is needed to refine the group to fecal origin. This group is known as the fecal coliform,

and they are different from the total coliforms in that they ferment lactose and produce acid and gas at 44.5°C within 24 hours (Madigan et al., 2012). The total coliforms are a broader range of bacteria that can be found in nature and are usually used to test for drinking water to ensure safety. The fecal coliforms are more fecal specific in origin and are used abundantly in ensuring the safety of recreational waters (USEPA, 2012).

Major bacteria primarily found in feces and fall under the term 'fecal coliform' include Bactericides, Bifid bacterium, Clostridium perfringens, Escherichia coli and Enterococci (Nevers and Boehm., 2011). Bactericides, Bifid bacterium and Clostridium perfringens are obligate anaerobes which are difficult to culture. E. coli and Enterococci are facultative anaerobes which make them easy to culture. The U.S. Environmental Protection Agency (USEPA) reviewed many cases of gastrointestinal illnesses and found that E. coli was a far more reliable fecal indicator in freshwater than Enterococci (USEPA, 2012).

2.2.2 Escherichia coli (E. coli)

E. coli is a gram negative, rod shaped, facultative anaerobic bacterium that is usually found in the gastrointestinal tract. It can be classified into 3 groups which include commensal, diarrheagenic and extra intestinal. The commensal E. coli is the most common type which normally lives in the gastrointestinal tract of warm blooded animals. Most strains of E. coli, like the commensal groups are harmless but there are some virulent types. The diarrheagenic strains can cause diseases such as diarrhea, hemorrhagic colitis, hemolytic uremic syndrome, inflammatory colitis and dysentery. The extra intestinal strains can cause urinary tract infections, septicemia and neonatal meningitis (Nevers and Boehm., 2011, USEPA, 2010).

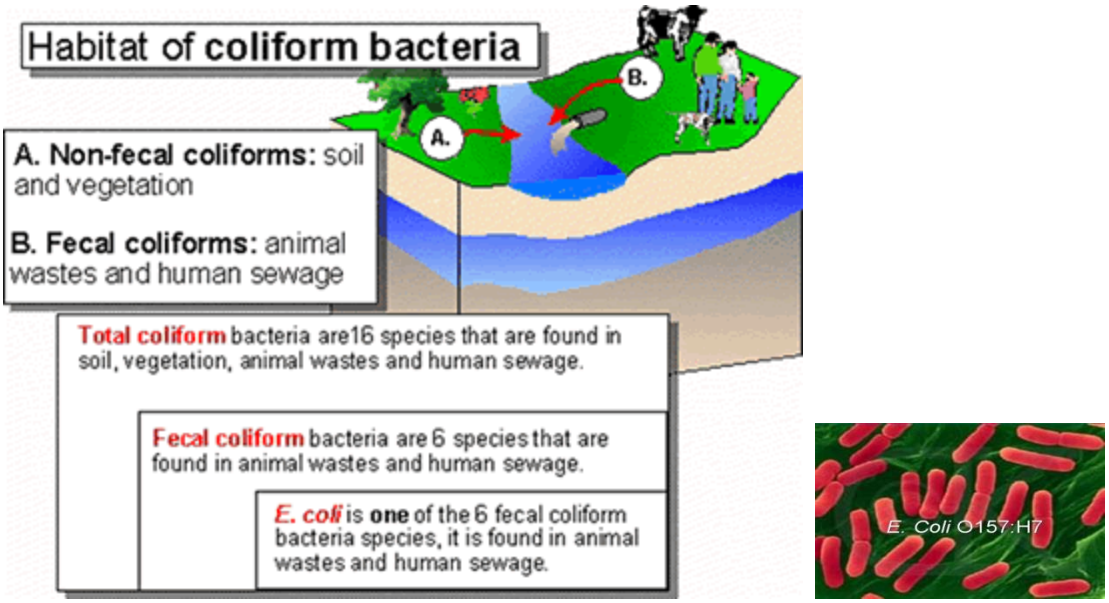


Figure 2-1 Indicator organisms in water

Source: http://dl.clackamas.cc.or.us/wqt111/unit-8-coliformhabitat.htm

E. coli can live from 4 – 12 weeks in water, depending on environmental conditions. E. coli can be shown to live through various stresses and it is well documented that it can survive through temperatures well below freezing (Cabral, 2010, USEPA, 2010). The ability for E. coli to survive through all different types of environmental stress could be due to the fact it has a high genetic diversity, as a higher genetic diversity tends to increase adaptability as well as resistance.

E. coli is a good fecal indicator as it is possible to differentiate E. coli from other bacteria using inexpensive and simplistic methods. Due to this reason, E. coli is being utilized as a fecal indicator not only in North America but worldwide. In Ontario the standard for beach water quality is set by the MOE at 100 E. coli count / 100 milliliters of water (MOE, 1994, City of Toronto, 2009).

2.2.3 Sources of E. coli Contamination

While point sources of bacteria are typically associated with sewage discharges, nonpoint sources can be divided into three general categories: human, animal and plant (Cabral, 2010). Human sources include failing septic systems, municipal landfills and land application of wastewater sludge. E. coli can also come from a variety of animal sources, including domestic pets, wildlife, livestock, land application of manure, pasture areas and feedlots (Hamelin et al., 2006). Research also indicates that natural organic substrates, such as bark and brush can be an additional source of bacteria.

Rainfall runoff are the primary source for E. coli to enter beach waters, which picks up other E. coli as it moves through the environment sources (such as leaking sewers, failing septic systems, wastes from wildlife such as birds or domestic animals) or through point source discharges (i.e., sewage from a pipe or other specific source). Other natural events such as heavy rainfall ("wet weather") can also elevate pathogen levels in beach waters as rainfall can flush pathogens into a water body from other areas of the watershed (Zhang et al., 2013). During high rainfall events, excess discharge of rainwater containing sanitary waste water is occasionally discharged in to recreational water body by the combined sewer pipes. This type of combined sewer overflow (CSO)s could be partially treated or non treated before they are released in to the recreational water. Sanitary sewer overflow (SSO)s are also prospective source of pathogen in recreational water, which are occasional and inadvertent discharge of raw sewage (Gironimo et al., 2009). Other point sources of potential water contamination include stromwater runoff from properties bordering the water bodies. Nonpoint-source discharges from poorly maintained or failing septic systems or other source of ground contamination can also prove to be a source of bacterial contamination of beach water.

2.2.4 Critical Environmental Factors

Most reported models predict the concentration of E. coli as a function of environmental factors. To demonstrate that models adopting these environmental factors are potentially robust enough for prediction, it is essential to review and discuss the physical basis behind the association of a number of environmental factors with microbial beach water quality. The variation in E. coli concentration has been reported to be affected by the environmental parameters discussed below.

- ***Rainfall***

Significant rainfall produces stromwater runoff and other surface water runoff (i.e. streams and rivers), which are the primary pathways for indicator bacteria and pathogens to reach beaches. Depending on the land use in the beach watershed, the stromwater runoff can contain animal feces and other bacterial sources that would have deposited on land between storm events. It has consistently been found that beach water quality declined after rainfall. Rainfall and microbial bacterial concentrations showed a positive correlation in a study by (Zhang et al., 2013).

- ***Streamflow***

Increases in streamflow are typically associated with rain events and runoff, and they could be indicative of high pollutant loads. Higher river flows are typically correlated with higher indicator bacteria levels at beaches located near a river outlet (Mass and Ahlfeld, 2007, USEPA, 2010).

- ***Solar radiation***

Mortality of bacteria increases with the intensity of ultraviolet radiation. Hence, an E. coli concentration is negatively correlated with the intensity of global solar radiation of the previous day. As a rule of thumb, water quality is usually good after a prolonged period of sunlight (Mas and Ahlfeld, 2006, Thoe et al., 2012).

- ***Turbidity***

Turbidity can be increased by stromwater input or streamflow, wind speed and direction, wave activity, swimmer activity and other factors. Some of these factors might be associated with input of pollution sources such as stromwater, streamflow and swimming (Nevers and Whitman, 2005).

- ***Water temperature***

Water temperature may indicate favorable or unfavorable conditions for indicator bacterial persistence in the environment, since some are intolerant to extreme high or low temperatures. Additionally, large changes in water temperature can indicate stromwater or streamflow inputs that may discharge indicator bacterial loads (Chan et al., 2013, Nevers et al., 2009).

- ***Wind Speed and Direction***

Wind strongly influences the transport of pollutant. Smith et al. (1998) studied the effect of wind speed and direction on the distribution concentration near an outfall. It was observed that bacterial concentration was significantly higher downwind of the outfall. Wind can be a good predictor of water quality since wind influences wave formation. Wind also affects the vertical mixing and suspends any accumulated bacteria on beach bottom (Smith et al., 1998).

- ***Lake level***

Beaches of the same lake typically have minor difference in water level caused by tides. Incoming tides are associated with onshore currents, which tend to prevent pollutants from flowing seaward. Tidal flushing of an embayment might occur, moving pollutants out from beach areas. For these reasons, tidal activity or changes in lake level has the potential to affect ambient water quality conditions either by increasing or decreasing indicator bacteria levels. In estuaries, the dominant mixing and transport processes are influenced by tidal flows (Chan et al., 2013).

- ***Wave Height***

The three main characteristics of waves are their height, wavelength and the direction from which they approach. Wave action can cause polluted stromwater runoff to remain in the near-shore zone or indicator bacteria in bottom sediments or sand to be re-suspended by wave action. Wave height has been consistently found to be statistically associated with microbial water quality by a number of studies on routine beach water quality monitoring records at a range of beach types (Nevers and Whitman, 2005, Chan et al., 2013).

The above environmental factors have been consistently found to be statistically associated with microbial water quality by a number of studies on routine beach water quality monitoring records at a range of beach types (Thoe et al., 2012, Zhang et al., 2012, Francy et al., 2013).

- ***Past E. coli Concentrations***

Unless there is heavy rainfall or sudden discharge of contaminated water, it is suspected that a beach with a good water quality "history" will likely be clean in the 'future' (Chan et al.,

2013). Reproduction rate of E. coli will be lower if the past E. coli concentration is low for a particular beach.

2.3 Study Area

Beaches are a key feature of Toronto's waterfront parks which contribute significantly to the quality of life in the city. Toronto's lakefront spans 157 kilometers of shoreline, with 24.4 kilometers made up of sand and cobble beach. Over 97% of the beach is owned or operated by the City and the TRCA. 18.9 kilometers of "wild" beach are not supervised by lifeguards or monitored for beach water quality, have few facilities and limited access and are typically used for walking and bird watching. The remaining 5.5 kilometers of supervised beach, designated for swimming at 11 locations (Figure 2-2), which are open during summer time. Consecutively, swimmable beaches are often used as an indicator of Toronto's environmental performance and quality of life.

Figure 2-2 Location of different beaches in Toronto (City of Toronto, 2009)

The state of the City's beaches has improved noticeably over the past few years and eight beaches have received Blue Flag. The blue flag program is internationally administered by the Foundation for Environmental Education in Denmark (www.blueflagorg) and by Environmental Defence (www.blueflag.ca) in Canada. Blue flag, a benchmark for high standards in water quality, cleanliness, safety and services, has been so far awarded to over 3,200 beaches and marinas across 37 countries in Europe, Africa, the Caribbean, New Zealand and Canada.

Toronto's 11 swimming beaches can be grouped into two categories according to their beach water quality (Figure 2-2). Eight beaches (Woodbine, Cherry, Ward's Island, Centre Island, Kew-Balmy, Bluffer's Park, Gibraltar Point and Hanlan's Point) fly the Blue flag, which requires that individual beaches have water quality which enables them to be open for at least 80% of the swimming season. As per the action plan for 2009-2010 set by the Toronto city council, the remaining three beaches (Sunnyside, Rouge and Marie Curtis Park East) are located near the mouth of major river systems and are with the poorest beach water quality and are regularly posted against swimming. As per this action plan the city council is most concerned about the assessment of water quality at these three city beaches in order to take required action (2009).

2.3.1 Factors Affecting Beach Water Quality in Study Area

E. coli concentrations in the GTA's streams are lowest in the headwaters and increase downstream toward the stream outlets. The Don Watershed and older urbanized portions of the Humber, Etobicoke and Mimico watersheds often receive untreated stromwater and some areas also have CSO as shown in Figure 2-3 below.

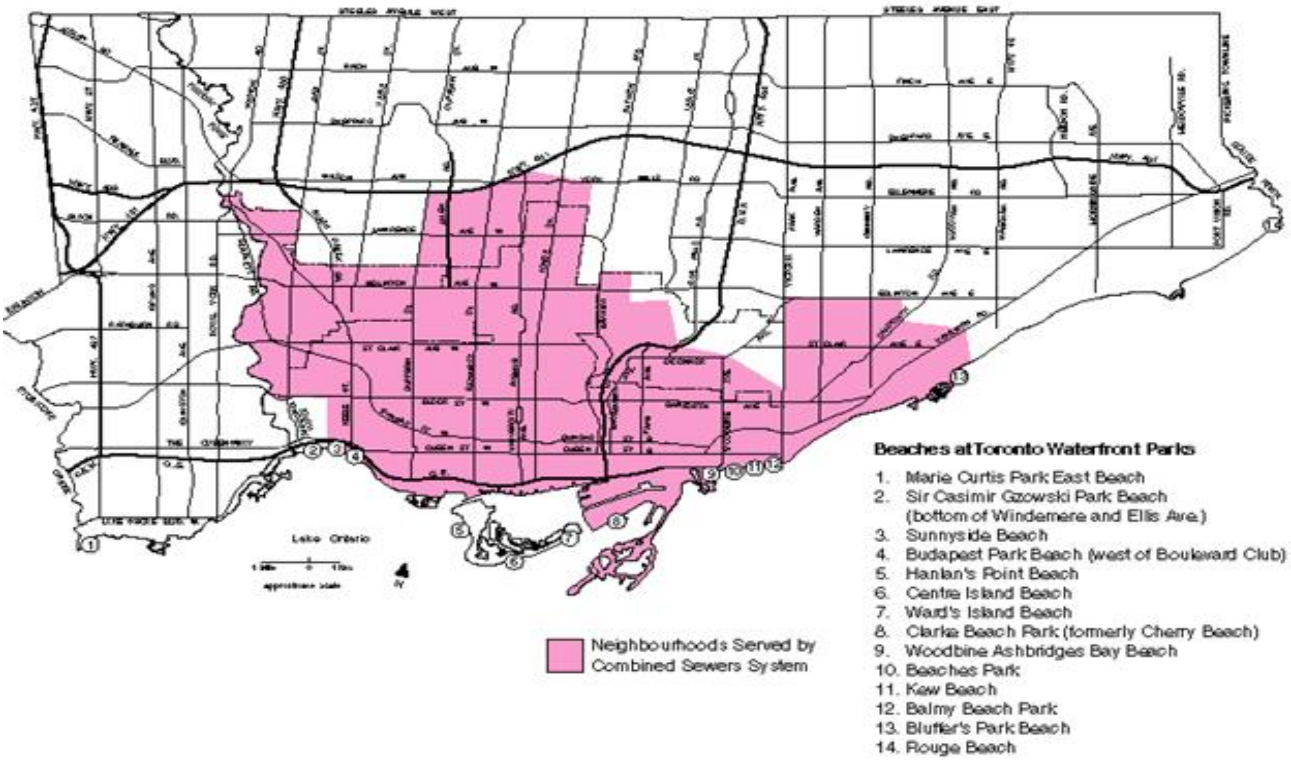


Figure 2-3 Toronto's combined sewers system (Amaral, 2010)

CSOs provide relief for combined systems in period of peak flow. This means that during periods of peak flow, runoff passes into the discharges runoff combined with raw sanitary sewage into local waterways and finally into the lake which contain high concentrations of E. coli (Amaral, 2010). Figure 2-4 represents the impact of high E. coli contamination in GTA. Rainfall is an important factor for beach closure, where sudden increases in flow throughout the watershed carry high concentrations of E. coli directly to the beaches.

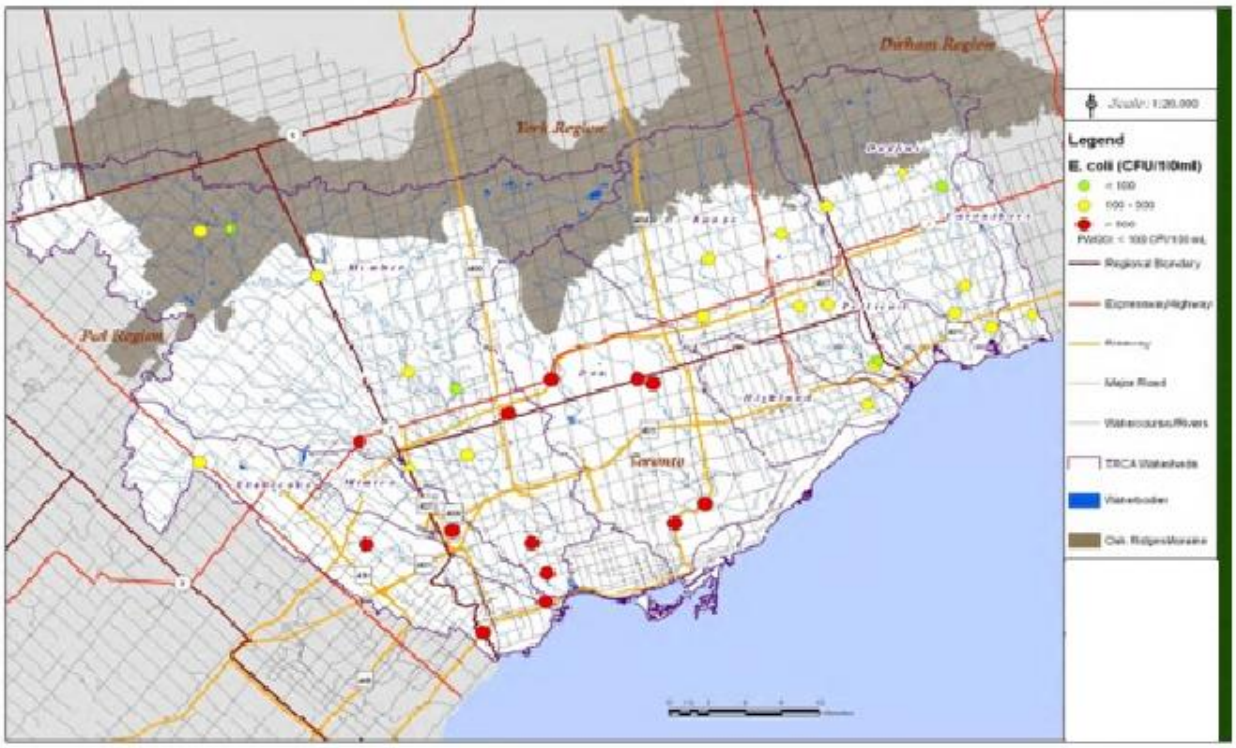
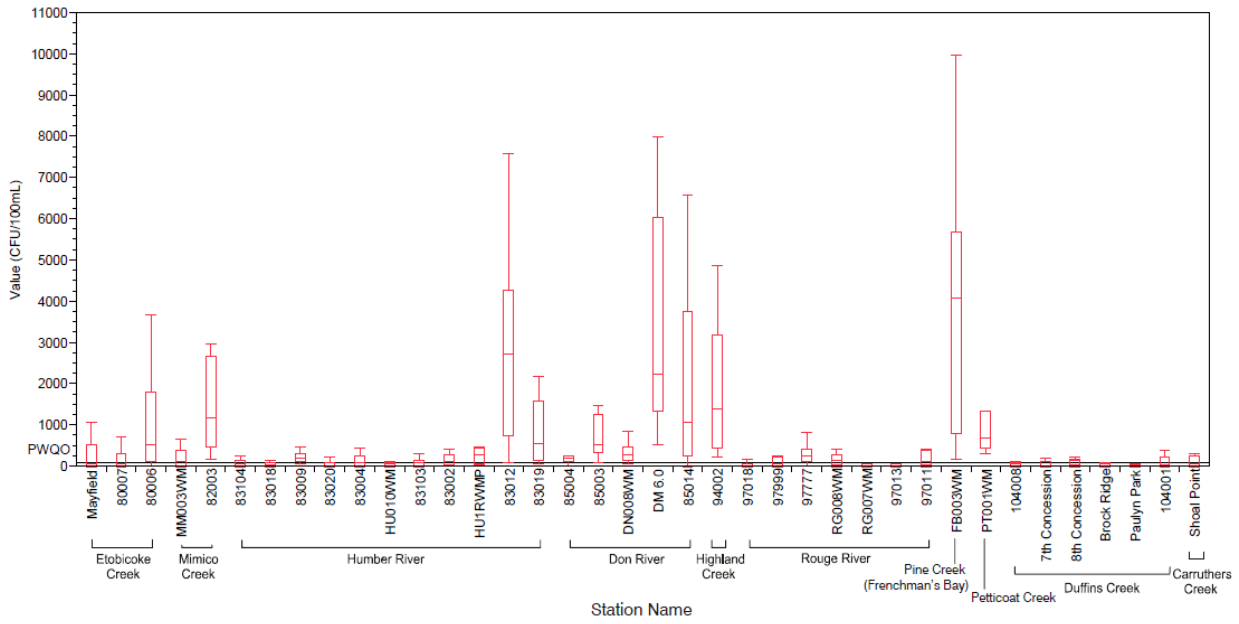


Figure 2-4 E. coli concentrations in GTA's area (Amaral, 2010)

Figure 2-5 shown below is one example of E. coli concentration in watershed around GTA from 2003 to 2007. The median values for 10 stations were above 500 CFU / 100 ml and 6 of those were above 1000 CFU / 100 ml (TRCA, 2009).



***Escherichia coli* concentrations (CFU/100mL) at 38 sites within TRCA jurisdiction (PWQO: 100 CFU/100 mL)**

Figure 2-5 E. coli concentrations within TRCA jurisdiction (TRCA, 2009)

Median E. coli concentrations at 89% of the sites were monitored above the PWQO of 100 CFU / 100 ml. Areas of concern include Etobicoke Creek, lower Mimico Creek, lower Humber River, Don River, Highland Creek and a mid-section of the Rouge River (TRCA, 2009).

2.4 Currently Used Predictive Models

According to the current practice based on traditional analysis method, 18-24 hours of time is required before the E. coli concentration can be reported. The persistence model i.e. using last available value to manage beaches is therefore unsatisfactory because of its lag period. To address this time relevance of water quality assessment issue numbers of strategies have been proposed. As a result of limitations associated with laboratory quantification of microbial water quality and the need for beach managers to balance access to water recreation with protection of public health, researchers have worked to develop real-time or near real-time predictive tools to

aid in beach management decisions. Rapid analytical techniques, deterministic models, regression models and artificial neural network based models are being some of them.

2.4.1 Rapid Analytical Techniques

Rapid analytical techniques of indicator organism quantification, such as Quantitative polymerase chain reaction (QPCR), currently take 4-6 hours to complete, exclusive of sample transport and laboratory preparation time, so the lag time between sampling and quantitative indicator organism information is several hours at best.

During the cross comparison studies in Racine, USA for the detection and quantification of E. coli and Enterococci positive correlation was observed between QPCR against agar based (US EPA Method 1600) and defined substrate methods in all surface water samples (river or freshwater bathing beach) except in direct stromwater discharge. The correlation even improved by introducing site specific corrective factors. This model predicted correct result by 98% of the time for the bathing beaches (Lavender and Kinzelman, 2009).

2.4.2 Deterministic Models

Deterministic or process based models are those that use mathematical equations to describe specific fate and transport mechanism for E. coli or other indicator organisms. There are three types of deterministic models based on type of mathematical equation being used: Gaussian, Lagrangian and Eulerian. USEPA provides a compendium of models that fall into this category; they include CORMIX, EFDC, HSPF and TPM to name a few. These models tend to be computationally intensive and require substantial data in order to be properly calibrated. As a result, they are not commonly used for beach predictive modelling (USEPA, 2010).

Deterministic models not only use complex meteorological parameters (such as water temperature, wind speed and direction) which may affect the time lags but depending on the reaction time and direct correlation test (using data obtained at the same point in time) may not necessarily produce meaningful results (Nevers and Boehm., 2011). Due to the complexity of the physicochemical processes involved this physically based models cannot reproduce accurately all results for E. coli contamination. Also, these models are time consuming to run and require a large amount of data for their application.

2.4.3 Statistical Models

Statistical model is a general term for any type of statistical modelling approach to predicting particular entity for a variety of applications. Linear regression models assume a linear relationship between factors or combinations of factors and indicator organisms (Nevers and Whitman, 2005, Olyphant and Whitman, 2004). The most highly developed statistical model approach is a MLR relationship between indicator organism and several independent variables. Typical, easy-to-measure environmental and water-quality variables include the following: meteorological conditions (solar radiation, air temperature, precipitation, dew point, wind speed and direction), water quality (turbidity, pH, conductivity/salinity and UV/visible spectra), hydrodynamic conditions (flows of nearby tributaries, magnitude and direction of water currents, wave height and tidal stage) and other factors such as number of birds or bathers. The most common model outputs are estimated levels of indicator organisms or the probability of exceedance of the state water quality standard (USEPA, 2010).

Francy et al. (2003) developed regression models to predict E. coli concentrations at five bathing beaches in Ohio, USA. Multiple models, using different explanatory variables were developed for each site. The model explanatory variables included rainfall, the number of

antecedent dry days, date, wind direction and speed, turbidity, streamflow in a nearby river, number of birds on the beach at time of sampling, lake-current direction and wave height. Models were evaluated using R-squared values as well as percentage of correct exceedance and non exceedance of the 235 count / l00 ml water quality standard for E. coli, and false positives and false negatives relative to that standard. R-squared values ranged from 17-58% for the models, while the correct classifications ranged from 73.2% to 80.9%. Percentages of false positives were reported as 4-15% and false negatives were 4-20%.

Forecast models of beach water quality have appeared in the literature, with the pioneering work concentrating on the freshwater beaches in the Great Lakes in the United States. Statistical models were developed to predict E. coli at 63rd Street Beach, Chicago, Illinois, based on the hydro-meteorological conditions and bacterial concentration measured on the site. The explanatory variables of the best-fit model included wind vector, rainfall, sunlight, lake stage, water temperature and turbidity. The model could predict 13 out of the 14 high bacterial level events. They remarked that no single environmental variable could successfully predict E. coli concentrations even at the relatively simple beach (enclosed embayment, no sewage outfalls) that they studied (Olyphant and Whitman, 2004).

The use of ordinary least squares (OLS) regression and logistic regression (LR) models were investigated to predict fecal coliform levels in the Charles River watershed in Massachusetts, USA. The models used a combination of meteorological and hydrological variables as well as previous day fecal coliform concentrations. Values of adjusted R-squared ranged from 46.11% to 60.40% for the OLS regression models. The model with the best overall fit used lag-l fecal coliform concentrations, a term describing the interaction between lag-1 bacteria concentrations and the amount of rainfall in the prior 24 hours, the time since

rainfall greater than 0.10 inches and average daily wind speed (Eleria and Vogel, 2005).

At effluent dominated beaches in Lake Michigan, the researchers developed predictive models for E. coli, which included wave height, lake chlorophyll and turbidity as explanatory variables. The models could explain about 60% of the variation in bacterial concentration and predict 6 out of the 11 exceedance events. The models were used to assist beach management on an experimental basis in summer 2005 (Nevers and Whitman, 2005).

This study developed Multiplicative Autoregressive Integrated Moving Average (ARIMA) models using the Box-Jenkins time series approach and compared them with Autoregressive Moving Average (ARMA) models developed for the short term forecasts of stream water quality. Monthly DO and BOD water quality records for 10 years from the River Ganges, India, were considered. The result showed that the Multiplicative ARIMA models produced closer forecasts to the observed values and yielded minimum RMSE values which were 0.077, 0.207 and 0.059 for DO data and 0.432, 0.560 and 0.271 for BOD data compared to traditional approach. Based on the overall performance of the Multiplicative ARIMA models, the authors recommended these types of models for water quality management in short-term forecasting (Shamshad et al., 2006).

2.4.4 Artificial Neural Network Models

An ANN is a construct of software that partially mimics the workings of a biological neural network. ANNs are often applied as nonlinear statistical data modelling tools. They can be used to model relationships between inputs and outputs or to find patterns. The technique is often useful when relationships between inputs and outputs are complex and not clearly understood. An ANN learns relationships between inputs and outputs using a learning algorithm (USEPA, 2010).

Mass and Ahlfeld (2007) observed that ANNs performed better than OLS and binary LR methods for predicting surface water fecal coliform concentrations in a mixed land use watershed.

He and He (2008) successfully used ANNs to predict indicator organism at marine recreational beaches receiving watershed base-flow and stromwater runoff in Southern California. The input variables of the network included water temperature, conductivity, turbidity, time lapse from last rain, rainfall amount, tide height, wave height, pH and flow rate of an incoming river. These more sophistical models are usually marginally better than the simpler MLR models and can better capture the extreme values.

Varma and Vijayan (2009) carried out the research work to predict fecal coliform concentration in surface water of the Achancovil River in Kerala, India. Different inductive models were developed using ANN and compared with statistical model, developed with SPSS tool and using same parameters. ANN models were developed using 5 readily available environment variables such as temperature, pH, turbidity, flow value and D.O. Out of the collected data for consecutive five years from 1996 to 2000, first four year data was used for ANN model development, means training and testing, and last year-2000's data was used for simulation. Out of all different combinations of input parameters for model development, highest correlation was obtained when pH, turbidity, flow value and D.O values were used as inputs. The correlation coefficient of 0.911 was achieved with ANN whereas with SPSS the same was 0.874. It was clear that ANN models outperformed the statistical models with the same type and number of variables used.

Zhang et al. (2012) compared ANN model developed in MATLAB toolbox for nowcasting and forecasting E. coli concentrations with other two models developed using US

EPA Virtual Beach (VB) Program at Gulf Coast beaches in Louisiana, USA. The ANN model included 15 readily available environmental variables such as salinity, water temperature, wind speed and direction, tide level and type, weather type and various combinations of antecedent rainfalls. The ANN model was trained, validated and tested using data sets (collected in 2007, 2008 and 2009) with an average linear correlation coefficient (LCC) of 0.857 and a Root Mean Square Error (RMSE) of 0.336. The two VB models, including a linear transformation-based model and a nonlinear transformation-based model, were constructed using the same data sets. The linear VB model with 6 input variables achieved an LCC of 0.230 and an RMSE of 1.302 while the nonlinear VB model with 5 input variables produced an LCC of 0.337 and an RMSE of 1.205. The results indicated that the ANN model with 15 parameters performs better than the VB models with 6 or 5 parameters in terms of RMSE.

A comprehensive study of beach water quality prediction had been carried out for four representative beaches in Hong Kong. The data analysis showed strong correlation of E. coli with seven hydro-environmental variables: rainfall, solar radiation, wind speed, tide level, salinity, water temperature and past E. coli concentration. The relative importance of the parameters was beach-specific and depends on the local geographical and hydrographical characteristics as well as location of nearby pollution sources. MLR and ANN models were developed from the regularly monitoring data (2002-2006) to predict the next-day E. coli concentration using the key hydro-environmental variables as input parameters. The models were validated against daily monitoring data in the bathing seasons of 2007 and 2008. The models were able to track the dynamic changes in E. coli concentration and predict WQO compliance with an overall accuracy of 70-96%. The MLR and ANN models had similar performances; ANN model tended to be better in predicting the high-end concentrations (Thoe et al., 2012).

The learning vector quantization (LVQ), MLR and ANN approaches were used in Charles River Basin, Massachusetts to provide a quick prediction of microbial concentrations for classification purposes using meteorological, hydrologic and microbial explanatory variables. With respect to classification, all three models adequately represented the non-violated samples (>90%). The MLR approach had the highest false negative rates associated with classifying violated samples (41-62% vs. 13-43% (ANN) and <16% (LVQ)) when using five or more explanatory variables. The ANN performance was more similar to LVQ when a larger number of explanatory variables were utilized (Motamarri and Boccelli, 2012).

2.4.5 Summary of Models Based on Predicted Fecal Indicator Organism (FIO)

Table 2-1 summarizes the different types of models adopted in the past for the prediction of fecal indicator organism.

Table 2-1 Summary of models based on predicted fecal indicator organism (FIO)

Study	FIO[a]	Explanatory variables	Modelling approach[b]	R^2	TN/TP[c] (%)	Standard (cfu / 100 ml)
Eleria and Vogel (2005)	FC	23 different variables	LR	0.54-0.69	97/63	1000
			LogR	0.46-0.56	97/63	
Francy et al. (2003)	EC	9 different variables	LR	0.35-0.44	TN: 53-99	235
					TP: 26-93	
Heberger et al. (2008)	Ent	Precipitation; intra-event time; discharge	LR	0.42-0.82	TN: 88, 84	61/305
					TP: 89, 100	
Hellweger (2007)	EC	Discharge; CSO; wind speed/direction	LR	0.60	80/98	235
			Mechanistic		93/70	
			Ensemble (50/50)		97/77	
			Ensemble (max)		74/99	
Chandramouli et al. (2007)	FC	7 different variables	ANN	0.63-0.94	97/61	200
Mass and Ahlfeld (2007)	FC	7 different variables	LR		TP: 51/38	20/200
			LogR		TP: 58-75/46	
			ANN		TP:61-81/46-62	
Tufail et al. (2008)	EC	Discharge; turbidity	LR	0.66-0.69		Three classes
			ANN	0.58-0.73	Overall 84-88	

			FFSGA	0.70		

a FIO - fecal indicator organism: FC - fecal coliform; EC - E. coli; Ent - Enterococci.
b LR - linear regression; LogR - logistic regression; ANN - artificial neural network; FFSGA - fixed functional set genetic algorithms.
c TN - true negative; TP - true positive.

2.5 Concluding Remarks

Concentration of bacteria such as E. coli is used the most frequently as criteria for beach water quality assessment. However since the bacterial concentrations vary dynamically with meteorological factors, assessing the beach solely based on past water samples may not be sufficient to protect the public health. Similarly to the analytical procedure, there is an increasing trend in using predictive models to assist in the beach monitoring.

While many deterministic, statistical and empirical models exist for beach water quality prediction, ANN models are increasingly being used for forecasting of water resources variables because ANNs are often capable of modelling complex systems for which behavioral rules or underlying physical processes are either unknown or difficult to simulate. Relatively few applications of ANNs involving indicator organism modelling for fresh beach water are reported in the literature (Mass and Ahlfeld, 2007). The primary difference between statistical models and ANN models is that in the former, a specific functional form is imposed on the data. For example, in MLR it is assumed that the output or dependent variable is a function of the linear combination of the input or dependent variables. If the assumption is incorrect, there will be an error in the prediction. In ANN models, although a functional form is imposed, it contains many more parameters that are determined through the training/learning process. As a result, the function form is more flexible and therefore may be better at approximating the "underlying rules" governing a relationship between input and output data.

Based on this literature review, it is clear that there has been research done to predict E. coli concentrations at bathing beaches using different types of data driven predictive models. Awareness about the limitations of assuming prior day E. coli concentrations to accurately reflect current day conditions for Toronto beaches has resulted in a strong requirement of research to create forecasting models to predict E. coli concentrations based on meteorological, hydrologic, and other potential environmental explanatory variables. ANNs have proved to be useful tool for modelling in a multitude of applications because of their ability to be trained using historical data and better accuracy compared with statistical models. Moreover, the networks possess the ability to learn non-linear relationships with limited prior knowledge about the process structure and can be applied to multivariable systems.

Due to above mentioned reasons, this research will focus on developing ANN model for prediction of E. coli concentration for most concerned beaches in Toronto, Ontario, i.e. Sunnyside, Rouge and Marie Curtis Park East Beach. These models would provide very inexpensive and simple way to predict E. coli concentration. Simultaneously, it may be used on a daily base or even as an emergency response system when it is not possible to collect water sample during bad weather conditions, which is one of the very crucial conditions for Toronto beaches.

3 ARTIFICIAL NEURAL NETWORK

This chapter reviews the theoretical background of ANN, including its learning algorithms, its paradigms, limitations, explains the mathematical foundations and biological inspirations behind ANN.

3.1 Motivation

ANNs were inspired by the need to develop artificially intelligent systems that can execute sophisticated computations similar to what a human brain consistently performs. ANNs acquire knowledge and learns through examples, similarly to how the human brain does, but they still have not managed to reach the stage where they can simulate even a trivial brain function (Basheer and Hajmeer, 2000). Nevertheless, ANNs provide an approach that has great potential in computationally solving complex problems. ANNs exhibit many characteristics which make them attractive and appropriate for nowcasting/forecasting. Zhang et al. (1998) provide key features of ANNs that can be listed as:

- Unlike traditional statistical and model-based approaches, ANNs are self-driven data adaptive methods, with the ability to learn (i.e. through examples), while refining their structure without the need of any predefined rules. In other terms, in the presence of correct data, ANNs can be viewed as experts of the domain who can analyse data effectively.
- Secondly, after learning from the presented data, ANNs can often correctly generalize the unseen data even if the training data contained noise. In principal, ANN seem to be

an ideal choice as nowcasting is usually carried out while predicting future trends based on historical behaviour.

- Thirdly, ANNs can be viewed as a non-parametric statistical approach, which determines the complex dependencies based on the observed data, without the presence of any functional framework. This means ANNs, as opposed to statistical methods, are universal functional approximators that can estimate any underlying continuous function with specified accuracy.
- Lastly, ANNs have shown clear potential in improving the timeliness and quality of econometric predictions, particularly in datasets that exhibit non-linearity between factors. ANNs can carry out nonlinear modelling without prior knowledge regarding the dependencies between input and output variables, which make them a general and flexible modelling tool for nowcasting.

These very facts and distinguished characteristics of ANNs serve as the main motivation for using ANNs over other statistical methods for this research work.

3.2 Introduction

An ANN is a statistical modeling tool, which can model or find non-linear correlations between input and output. ANN models have been founds useful and efficient in a wide variety of tasks which are hard to solve using ordinary rule-based programming.(Basheer and Hajmeer, 2000).

ANNs are based on the structure and function of biological neural networks such as the central nervous systems of humans and animals. The brain contains approximately 100 billion nerve cells or neurons, which are specialized to carry information and to action commands from

the brain via electrochemical processes. These processes take approximately 10^{-3} seconds to complete. While neural events are five to six orders of magnitude slower than the silicon logic gates that function inside a computer, the number of neurons and number of connections between them make the brain extremely efficient and capable of performing certain tasks such as pattern recognition much faster than digital computers (Haykin, 1994).

The brain learns to perform many tasks through experiences. A significant portion of this learning occurs in humans in the first few years of life when the synaptic connections which mediate the transmission of signals between neurons are formed. The brain is said to be very "plastic" at this time and this plasticity lasts to some degree throughout a human lifetime because the brain has the ability to adapt from its environment by modifying existing synaptic connections or creating new ones (Jain et al., 1996). ANNs seek to capitalize on the plasticity and efficiency of biological nervous systems for problem solving. ANNs are massively parallel distributed processors, each node in an ANN receives input, generates an output and distributes that output either to another node as input or as the final output of the network (Haykin, 1994). There are similarities to the human brain because, like the brain an ANN acquires knowledge through a learning or training process and the interconnection strengths between the basic processing units of an ANN are based upon the concept of synaptic weights between neurons in the brain (Fausett, 1994).

McCulloch and Pitts (1943) designed systems that are generally regarded as the first ANNs. The weights on a McCulloch-Pitts neuron are set so that the neuron performs a particularly simple logic function (Fausett, 1994). The perceptron which are large classes of ANNs, was first introduced by Rosenblatt (1962).

An ANN is an information processing system that has certain performance characteristics in common with biological neural networks. An ANN has four common assumptions:

- Information processing occurs at many simple elements called neurons,
- Signals are passed between neurons over connection links,
- Each connection link has an associated weight, which in a typical neural net, multiplies the signal transmitted,
- Each neuron applies an transfer function (usually non-linear) to its net input (sum of weighed input signals) to determine its output signal (Fausett, 1994).

3.2.1 Functionality

"At the most abstract level, a neural network can be thought of as a black box, where data is fed in on one side, processed by the neural network which then produces an output according to the supplied input" (Sarle, 1994). In general, a neural network is capable of computing any kind of data, e.g. qualitative or quantitative information. To enable faster training and optimized results, the input data of the neural network should be preprocessed (e.g. filtered, transformed). As a matter of fact the selection, preprocessing and coding of information is one of the most crucial tasks to consider while working with neural networks.

3.2.2 Artificial Neuron

A neuron is the fundamental processing element of a neural network which collects information from all preceding neurons relative to the flow of the information and propagates its output to the neurons in the following layer. This building block of human awareness encompasses a few general capabilities. A biological neuron accept inputs from preceding sources, process the inputs by combining them in some way, performs generally a nonlinear

operation on the result and then outputs the final result. Figure 3-1 shows the relationship of these four parts.

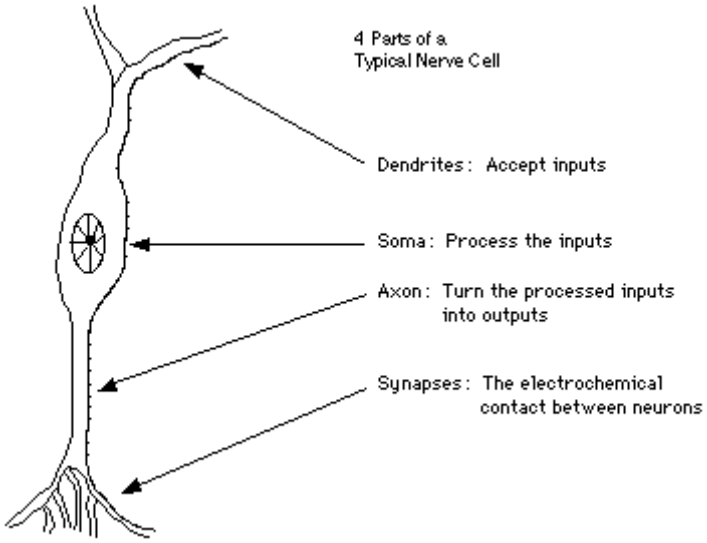


Figure 3-1 Basic components of neuron (Haykin, 1994)

Natural neurons receive signals through *synapses* located on the dendrites or membrane of the neuron. Dendrites are hair-like extensions of the soma which act like input channels. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse and might activate other neurons. Similarly to perform the same task, the artificial neurons, the basic unit of neural networks, simulates the four basic functions of natural neurons. Figure 3-2 shows a fundamental representation of an artificial neuron.

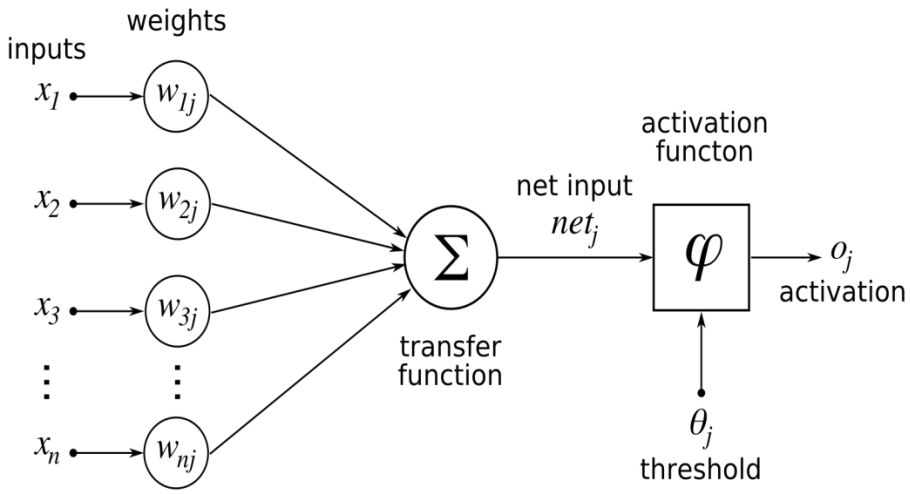


Figure 3-2 Model of artificial neuron (Haykin, 1994)

A network consists of a number of elements or nodes, denoted as x_i. Each node receives signals from other nodes, processes and forwards them to other nodes. Here at every moment, an activity in each node denoted as x_i. Nodes are connected by directed connections, denoted w, which has a weight or strength. Node x_n is connected with node x_j with connection w_{nj} (Haykin, 1994).

Signal dynamics of a network can be modeled either as continuous or discrete. The discrete is easier to explain. Input or other elements' activities are transformed into signals and proportionally strengthened by the weights. When inside the node, all signals are summarized. The transfer function takes the summarized input as argument and the output value of this function is the nodes' resulting activity or output. The activating function is denoted with φ and the resulting activity with o_j. Depending upon the application's requirements, the most appropriate transfer function is chosen (Fausett, 1994).

An ANN is characterized by,

- Its pattern of connections between the neurons (called its architecture),
- Its method of determining the weights on the connections (called its training or learning algorithm) and
- Its transfer function (Haykin, 1994).

3.2.3 Connections

The paths between neurons are called Connections. Very often the neurons of two succeeding layers are fully interconnected and all the information flows through these connections. There might exists additional connections going to further layers or even missing connections between certain neurons (Zhang et al., 1998). Basic function of connection links is to send input from one node to another in an ANN.

3.2.4 Weights and Biases

Each connection is equipped with an individual weight and bias that modifies the signal flow on the respective connection. The weight works as a conceptual connection strength between neurons and is adjusted during learning algorithm. The bias neuron works as a fine tuning which lies in one layer and connected to all the neurons in the next layer. By using bias neuron, product of weight and output from the preceding layer is added to successive layer. As the information is stored and distributed through weights and bias neurons in a neural network, even a negligible destruction of the same, will result in to a large effect on the recall of the learned function.

Depending upon the influence of the input, value of weight of artificial neuron could be higher or lower. Sometimes, weights can be negative, which means that the signal is suppressed by the negative weight. Desired output can be obtained for the particular set of

inputs by manually adjusting the weight of a neuron. With ANN of numerous neurons, it would be quite complicated to find manually all necessary weights. However, through the process of learning or training, which uses algorithms to adjust weights of ANN, can be used to obtain desired results (Gershenson, 2003).

3.2.5 Transfer functions

The transfer function, which is sometimes called the squashing function, is applied to the net input received by a node. The function controls when the neuron should be active, depending on whether a given threshold is reached or not. Transfer functions are the processing units of a neuron and they can be linear or non-linear. The output or range of the transfer function is usually 0 to 1 or -1 to 1. Some useful transfer functions are pure linear, log-sigmoid and tan-hyperbolic functions, depicted in Figure 3-3 (Cybenkot, 1989):

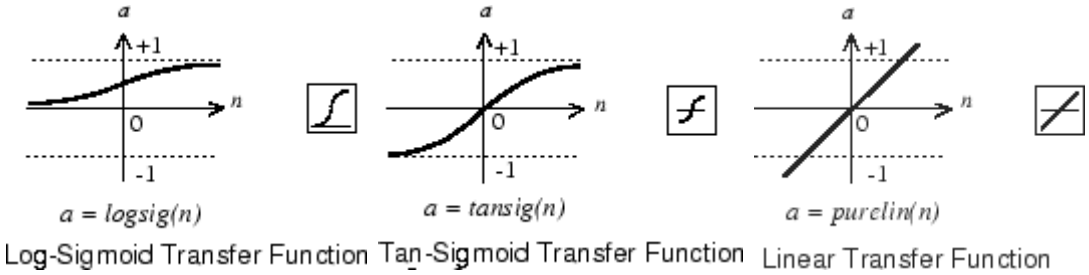


Figure 3-3 MATLAB built-in transfer functions (Beale et al., 2013)

The mathematical formulation of the above functions is given as follows:

$$\text{Log-sigmoid function: } f(n) = \frac{1}{1+e^{-n}} \qquad (3\text{-}1)$$

$$\text{Tan-sigmoid function: } f(n) = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}} \qquad (3\text{-}2)$$

Linear function: $f(n) = n$ *(3-3)*

When choosing the transfer function some important factors need to be considered, as they affect the performance of an ANN. Using linear functions in multilayer network is pointless because the biological correspondence is nonlinear. For non-linear datasets, most commonly used functions are log sigmoid and hyperbolic tangent.

3.2.6 Learning in ANNs

As the name signifies, during the training phase a neural network learn itself a sample pattern upon presenting set of the input data. While learning, the weights and biases of the neural network are adjusted. The learning procedures of ANNs can be classified into supervised and unsupervised learning. In both cases, however, every training starts with a recall where the input is propagated through the neural network and all its neurons change their activity accordingly. Supervised learning requires an external source to control the learning and incorporate the global information. The source may be a training data set or an observer who grades the performance. Examples of supervised learning algorithms are the least mean square (LMS) algorithm and radial basis function network (Fausett, 1994). In supervised learning, the ANN is trained to have the optimal agreement between the ANN output and the training data set. In environmental modelling applications, the training data set can be composed of environmental quality observations. In training of the ANN, the value of the weights in the connections between the neurons is modified according to the input/output samples. In the case of unsupervised learning, the system organizes itself by internal criteria and local information designed into the network. Most important aspect of unsupervised learning is to decide the point where to terminate training, as sometimes it is possible to over-train a neural network. “Namely, at some point the

neural network starts to memorize exactly the training examples with their inherent noise and later on it will not be able to generalize from the trained examples to new patterns presented during recall" (Cybenkot, 1989). As the name suggests, during unsupervised classification, the neural network classifies the data by itself.

3.2.7 Types of ANNs

Based on the way the neurons are interconnected in a model, neural networks can be broadly classified into two types namely feedforward and feedback networks. In feedforward ANNs the data moves in a forward direction, i.e. from the input layer towards the output layer, where as in feedback ANNs data is sent back to layers as the 'feedback'.

3.2.7.1 Feedback Networks

Particular network is also referred as recurrent network and unlike feedforward ANNs, feedback networks contain at least one feedback loop (Figure 3-4).

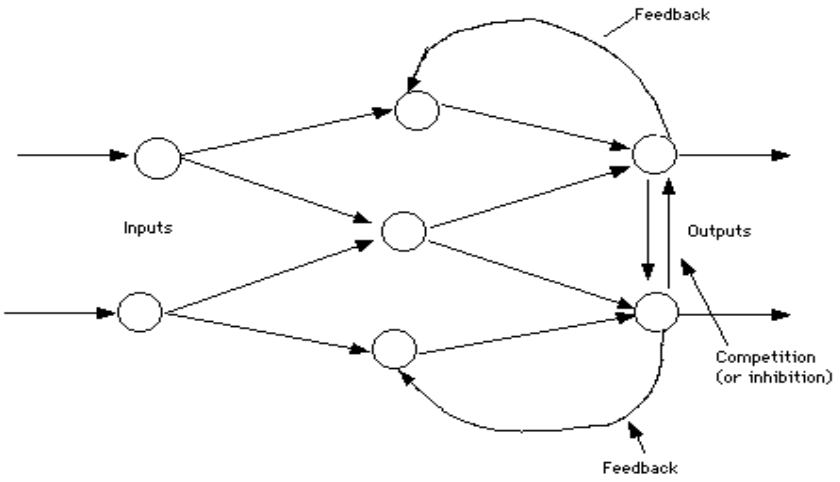


Figure 3-4 Simple feedback networks (Wikipedia, 2013)

The presence of feedback loops result in nonlinear dynamic behavior in the ANN (Haykin, 1994). The training associated with feedback models is typically more complex than for feedforward networks.

3.2.7.2 Feedforward Networks

A single-layer feedforward network is the simplest form of this type of network. As the name suggests, such a network consists of only an input layer of source nodes that feed information to a layer of computational nodes that are also the output layer (Haykin, 1994). Because of the presence of only one computational layer, this network architecture is called single-layer, as shown in Figure 3-5.

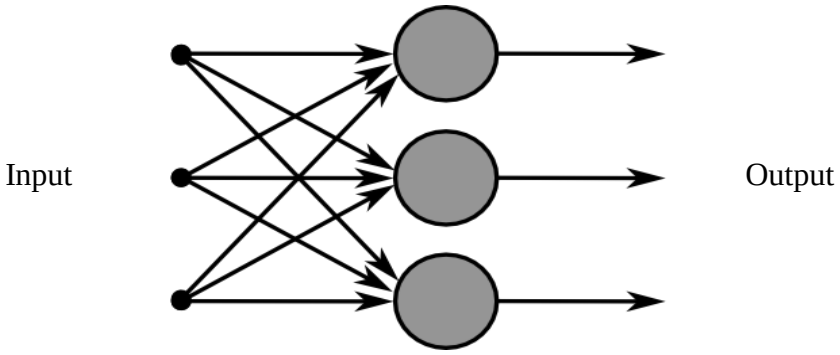


Figure 3-5 Single layer network (Wikipedia, 2013)

Multilayer feedforward networks consist of one or more layers that are located between the input layer of nodes and the output layer (Figure 3-6). These layers, called hidden layers, receive input from the preceding layer. Neurons in the hidden layers perform the type of computations described above and pass the resultant output to the next layer in the network. The output of the final layer is the overall response of the ANN to the network input. A multilayer network in which each node in one layer is connected to every node in the next forward layer is called a

fully connected network. If some connections between nodes in adjacent layers are absent, the ANN is partially connected. Feedforward ANNs are further divided as linear and non linear

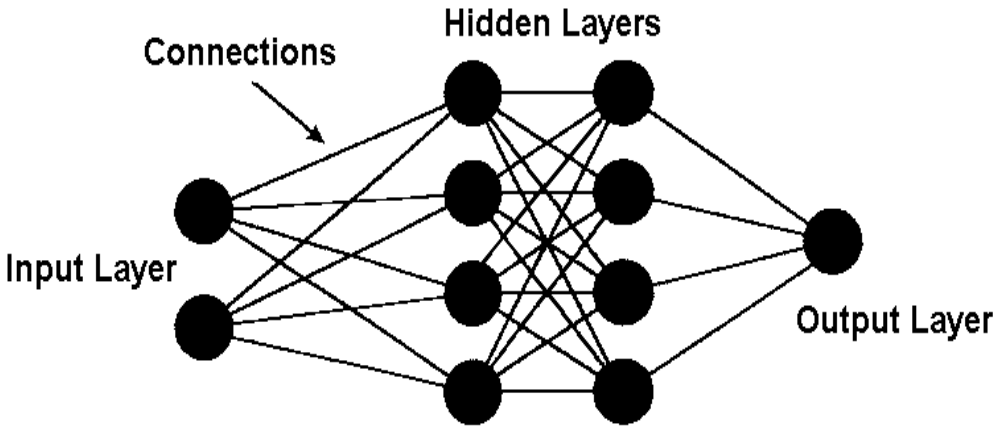


Figure 3-6 Typical Model of a Feedforward Neural Network (Wikipedia, 2013)

models. Non-linear models are used particularly where datasets exhibit nonlinear dependencies and are often used under the supervision of user or without any supervision. In this research work, supervised approach is adopted as datasets are provided to ANN models.

Whether the particular case is of feedforward or feedback network, the network can be either single layer or multilayer perceptron. The examples for these types of models are discussed here.

- **Adaptive Linear Element (ADALINE)** is a single layer perceptron and feedforward network that accepts several inputs and produces one output.
- **Multiple Adaptive Linear Element (MADALINE)** is a multilayer perceptron and feedforward network, composed of more than one adaptive linear element.
- **Back-propagation network** is a multi layer perceptron and feedforward network that employs the back-propagation algorithm which uses the gradient descent technique with the error propagated backwards.

- **Hopfield, Kohonen and Adaptive Resonance Networks** are all feedback and multilayer perceptron models, each of which can be used in various disciplines.
- **Hybrid networks** are those composed of certain networks each performing its own function. These models could both be feedforward or feedback, depending on the type of the network configured. Examples for these can be parallel network models and differentiation models.

3.3 Learning Algorithms in Neural Network

The basic concept behind the successful application of neural networks in any field is to determine the weights to achieve the desired target and this process is called learning or training. Learning processes consist of supervised, unsupervised and reinforcement learning and its success is typically measured by some performance matrix. Simulation is the testing of the model with input data that was not used to train the model in order to assess its ability to generalize the relationship between input and output data (Basheer and Hajmeer, 2000, Kisia and Uncuoglub, 2005). Because of the commonness of supervised learning in the prediction of time-series data such as water quality data, this section will focus on methods of supervised learning.

In supervised learning there is an output or target specified for every input used in the training process. During the training, input-output pairs are used. The input consists of a vector of real numbers with each element of the vector corresponding to an explanatory variable. Each input is propagated through the ANN and the model output is compared with the target data. The target data is also a vector of real numbers that gives the values of the variables being modeled by the ANN.

The goal of the training process is to optimize the ANN to minimize the differences between ANN output and target data values by adjusting the weights between nodes. The following discusses three common methods used during supervised training of multilayer feedforward networks: back-propagation, Conjugate gradients and the Levenberg-Marquardt algorithm.

3.3.1 Back-propagation algorithm

Back-propagation is the most commonly used training algorithm for feedforward ANNs and is a gradient descent method (Haykin, 1994). This algorithm is based on minimizing the error of the neural network output compared to the required output. "A learning cycle starts with applying an input vector to the network, which is propagated in a forward propagation mode which ends with an output vector. Next, the network evaluates the errors between the desired output vector and the actual output vector. It uses these errors to shift the connection weights and biases according to a learning rule that tends to minimize the error. These steps are repeated until the error is either small or time is exhausted. The adjusted weights and biases are then used to start a new cycle" (Zhang et al., 1998).

The error of network is relative to the training set which is defined as the sum of the partial errors of network E_k relative to the individual training patterns and depends on network configuration:

$$E = \sum_{k=1}^{p} E_k = \sum_{j} \sum_{i} \left(d_{ji} - y_{ji}\right)^2 \qquad (3\text{-}4)$$

Where E_k is partial network error *j* is the number of patterns or inputs in the training set, *i* is the number of output nodes and d_{ji} and y_{ji} are the target and actual output values for the i^{th} node on the j^{th} pattern.

This total squared error should be below a certain specified value. If E is not below the tolerance value the network must go through another training epoch. The error decreases until the goal or the maximum number of epochs is reached.

3.3.2 Conjugate Gradient algorithm

Conjugate gradient algorithms provide an alternative to the back-propagation technique described above, but may still incorporate a gradient descent method like back-propagation. The basic conjugate gradient algorithm adjusts the weights in the steepest descent direction (the most negative of the gradients) (Kisia and Uncuoglub, 2005). This is the direction in which the performance function is decreasing most rapidly.

3.3.3 Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm is based on two algorithms, namely steepest descent algorithm and Newton's method. Out of these two optimization methods, the primary is based on first order Taylor series and following is on second order Taylor series. Newton's method can be defined by,

$$\Delta w_{ij}(n) = -A_n^{-1}\left(\frac{\partial E}{\partial w_{ij}}\right) + \Delta w_{ij}(n-1) \qquad (3\text{-}5)$$

Which also can be written as,

$$\Delta w_{ij}(n) = -A_n^{-1} g_n + \Delta w_{ij}(n-1) \qquad (3\text{-}6)$$

where A_n is the Hessian matrix of the performance function at the current (n^{th} value) of weights and biases and g_n is the current gradient of the performance function (Beale et al., 2013). When the performance function is the sum of squares, the Hessian matrix can be approximated as,

$$H = J^T J \tag{3-7}$$

and the gradient given by,

$$g = J^T e \tag{3-8}$$

where $\boldsymbol{J}$ is the Jacobian matrix consisting of the first derivatives of the weights and biases of the network and e is the vector of network errors (Beale et al., 2013). The Levenberg-Marquardt algorithm uses the approximation in the following form,

$$\Delta w_{ij}(n) = -[J^T J + \mu I]^{-1} J^T e + \Delta w_{ij}(n-1) \tag{3-9}$$

where I is the unit matrix (Kisia and Uncuoglub, 2005). When μ is set to zero, Equation (3-9) becomes Newton's method described above. When μ is large, Equation (3-9) becomes gradient decent method.

In the MATLAB implementation of the Levenberg-Marquardt algorithm used for this research, the value of μ decreases with every training set that reduces model error and is increased only when a tentative step would increase the error. This algorithm is used throughout the research described here because of its speed and efficient implementation in MATLAB (Beale et al., 2013). In addition, generally good performance is reported on moderately sized neural networks and the Levenberg-Marquardt algorithm has been shown to converge to an optimum solution for problems when conjugate gradient and standard back-propagation with

variable learning rate algorithms failed to converge (Kisia and Uncuoglub, 2005, Motamarri and Boccelli, 2012).

3.4 Simulation and Generalization

It is necessary to test the performance of a trained neural network model before applying the model. In the simulation process, an ANN is subject to input data never used in the training process and the ability of the model to match the target output values is measured. Training of an ANN can continue until either,

- the network reaches a minimum error as specified by the modeler,
- a maximum runtime/number of epochs specified by the modeler is reached

For the simulation purpose, two types of model performance assessment measures are commonly used

- R-squared or adjusted R-squared: These performance statistics can be used to describe the variability in observed outputs by the model. The higher the R-squared value, the better the model performance in terms of explaining variability in observed output.
- RMSE: It is another method to quantify model bias and precision. It is a way to aggregate the model residuals (i.e. the difference between predicted and observed values of output) into a single value. The lower the value of the RMSE, the smaller the differences between observed and predicted values (Helsel and Hirsch, 2002).

From the discussion of these performance statistics, it is possible to see how a few or large errors could dominate the calculation of the statistics. For that reason, multiple measures of model performance are recommended.

The ability of an ANN to correctly approximate target values for given inputs that are not part of the training set is called generalization. Good generalization ability typically requires the following,

- Inputs which contain enough information about the target that it is possible for the ANN to develop a functional relationship between inputs and outputs with an adequate degree of accuracy;
- The function which model is trying to learn is at least somewhat smooth, i.e., a small change in inputs produces a small change in outputs;
- The training cases are sufficiently large and representative of the subset or sample of the larger population of data that the model is required to be able to generalize.

If a model shows poor generalization ability it is commonly because, the training set was not representative of the larger population to be modeled or the model was over fit. Overfitting occurs when the model learns too many specific input-output relationships. It essentially memorizes the training data and is unable to correctly interpolate or extrapolate the functional form for the relationship (Haykin, 1994).

If a beach management authority is concerned with the question of whether or not the model prediction is above or below a water quality standard value, assessing the false positives and false negatives is useful. False positive results when the model predicts an E. coli value greater than the beach water quality guideline and the observed E. coli concentration is less than the guideline value or a posting or notification when one is not warranted. False negative results when the model predicts an E. coli concentration less than the beach water quality guideline and the observed E. coli concentration is greater than the guideline value or the beach is not posted when water quality conditions are such that it should be. The percentages of false positives, false

negatives, true positives and true negatives can be calculated and can also be shown visually on a scatter plot of observed and predicted values of E. coli with lines indicating the values of the water quality standard of interest.

In addition to the calculation of statistics, plotting the time series of observed and predicted values can be advantageous and help to diagnose strengths or weaknesses in a model. Evaluation of the persistence model and comparison of those results with the performance of ANN models should also be performed in order to assess if ANN models are providing any advantage over currently used methods for beach notification posting.

3.5 Summary

This chapter reviewed different types of ANNs, algorithms, functionality and terminology related to the development and testing of ANNs. Based on the discussion presented above, it is clear that due to simplicity, the existence of a well-defined learning algorithm and ability to predict non-linear dependencies feedforward networks are best suited for the nowcast modelling. Due to the same reason, this research work has been performed using feedforward network. According to Khanna (1996) following are the in general steps to be followed for creating a neural network application:

- Analysis of the problem and collection of all available data
- Analysis of the collected data
- Choice of the neural network type that is capable of solving the problem
- Selection of the important features that will be used
- Coding of the information, using the result of the data analysis
- Separation of data basis into training and test set

- Design of the appropriate neural network topology, choice of the neurons' functions and basic decision about the amount of neurons to be used in each layer
- Training of the neural network and monitoring its performance on the test set
- Optimization of the neural network by changing the topology, the amount of neurons and the neurons' functions.
- Simulation of the network on the data set which are not introduced before to the network.

4 DATA DESCRIPTION AND ANALYSIS

4.1 Introduction

This chapter discusses data gathering and exploratory data analysis for the beach water quality of selected Sunnyside beach, Rouge beach and Marie Curtis Park East beach in Toronto, the first critical steps in the development of a predictive model. As discussed in Section 2.3 - Study Area- because of poorest beach water quality, city council is most concerned about the assessment of water quality of these three beaches. Due to those reasons and suggested by Toronto Public Health authorities these three beaches were selected out of Toronto's 11 swimmable beaches to study the typical variation of the E. coli concentrations and hydro-meteorological factors.

4.2 Data Sources

Sufficient quantity and quality of data is at the core of predictive model development. Under the regular beach monitoring programme conducted by the Toronto Public Health Department, City of Toronto, at each beach, a number of hydro-meteorological and water quality parameters are typically measured on each sampling trip. The data are sampled every day during the bathing season (i.e. June to August). Water temperature and wave action are readily measured onsite during sampling time, while E. coli concentrations and turbidity are obtained through laboratory analysis of the water samples collected from the beach, using the standard methods.

Period of record and frequency of indicator organism monitoring are among the most important considerations in predictive model development. There are a wide range of environmental variables that may be used in the development of predictive models for beach water quality. As discussed in Chapter 2, the most commonly used variables are those that both have some relationship to beach water quality and are typically readily available: rainfall, streamflow, solar radiation, lake level, wind speed and direction, turbidity, wave height and past E. coli concentrations.

Out of all these variables, water quality and other meteorological data for model development were provided by the different government authorities either through direct delivery of electronic files or via links to downloadable publicly available data. A summary of the data sources and information about online availability is provided in Table 4-1. Data for hydro-meteorological parameters were considered only up to previous day's midnight taking into account maximum lag time out of all input parameters; this is to maintain the nowcasting ability if they are used as inputs to the water quality predictive models. The data taken for all beaches are also listed in Table 4-1.

Table 4-1 Data Sources for Toronto Beach Modelling

Data	Source	Historical	Real-time	Website	Notes on Availability
E. coli	TPH	X	X	N/A	Available after 18-24 hours using the persistence model
Precipitation	TRCA	X	X	N/A	Available at 5-minute interval on a daily base.
	Toronto Water	X	X	N/A	Available at 5-minute interval on a daily base.
Solar Radiation	TRCA	X	X	N/A	Available at 15-minute interval on daily base.
Streamflow	Environment Canada	X	X	http://www.wateroffice.ec.gc.ca/text_search/search_e.html?search_by=p&region=ON	Typically, 15-minute discharge data is published with an approximately 9 hour lag time.
Lake Level	Fisheries and Oceans Canada	X		http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/twl-mne/inventory-inventaire/interval-intervalle-eng.asp?user=isdm-gdsi&region=CA&tst=1&no=13320	
	Fisheries and Oceans Canada		X	http://www.tides.gc.ca/C&A/wldata/torthis.htm	Hourly lake levels are available with an approximate 8 hour lag time for online publication.
Wave height, Wind speed & direction	NOAA National Data Buoy Center	X		http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/waves-vagues/search-recherche/list-liste/data-donnees-eng.asp?medsid=C45139	
	NOAA National Data Buoy Center		X	http://www.ndbc.noaa.gov/station_page.php?station=45139	Buoy data are available with an approximate 1 hour lag time. Buoys typically decommissioned during the winter months.

Figure 4-1 shows the locations of the all rain gauges, buoy station, lake level station and stream gauge stations.

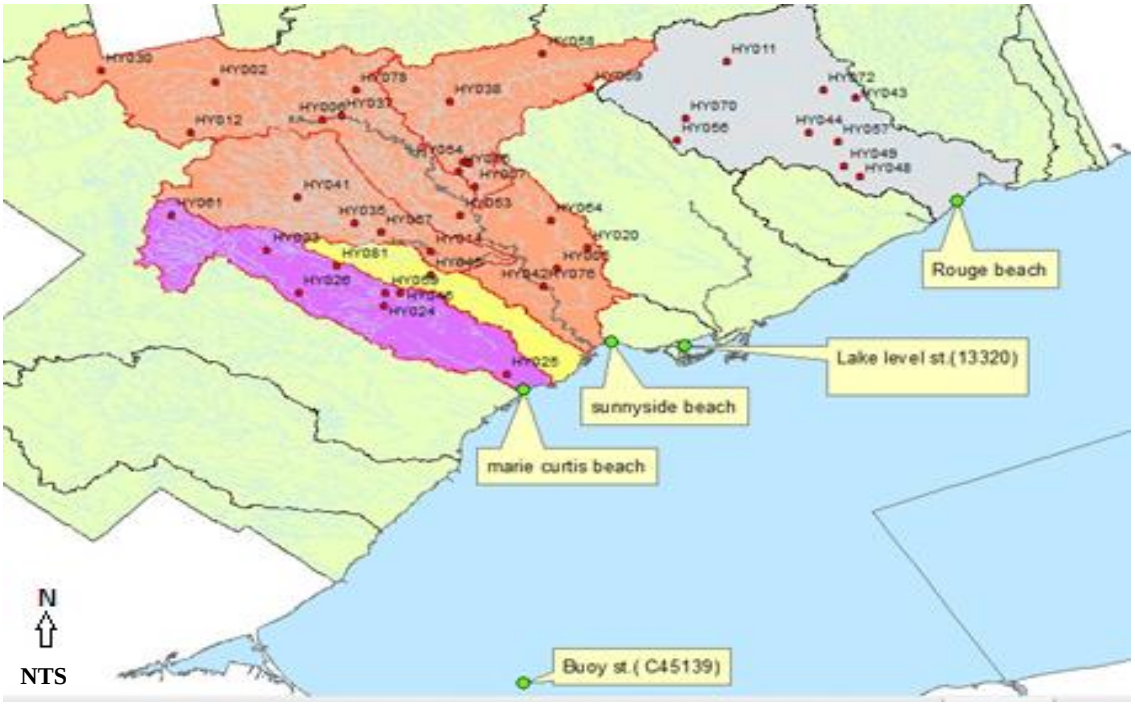


Figure 4-1Locations of the beach, buoy station and lake level station

4.3 Characteristic of Explanatory Data

Taxonomy of required data is one of the most important, time consuming and decisive task to perform prior to model development. The observed bacterial variability at each beach can be related to the distribution of the governing environmental factors. A range of explanatory variables are used in the statistical analysis to detect associations with E. coli and potential utility for predictive model development. The following are the brief characteristics of explanatory variables and the procedure adopted to sort out required data.

Precipitation: Data for numerous rain gauges were received from TRCA and Toronto Water for the period of 2008 to 2012 bathing season (i.e. July to August). Unlike the MLR model which used only Toronto International Airport (YYZ) rain gauge station for all beaches, for this research work the precipitation data were obtained from the rain gauge stations located inside the

watershed area for a particular beach which reflects more realistic situation. Rain gauge station selection was performed by following the steps mentioned below:

- Categorization of the required stations was done based on the criteria if the station is,
    - located inside or outside of the related watershed area of the beach. Stations located outside watershed area were omitted.
    - missing any major part of the data for bathing season.
    - there is higher value of correlation between precipitation and E. coli data. The same is explained in more detail in Section-4.5.
- ArcGIS 10.1 was employed to locate rain gauges, stream gauges and weather stations (measuring air temperature, solar radiation, wind speed and direction), provided by TRCA and Toronto Water as shown in Figure 4-2. To locate all these stations with their Easting-Northing, the required GIS maps were collected from Ryerson University Geospatial Map and Data Centre.

Daily precipitation values from the selected rain gauge stations were calculated based on 5 min-interval data using Microsoft Access and Pivot Table function. Hourly data was calculated based on this data set. 24-hour precipitation was calculated for a day based on the previous day's midnight to midnight precipitation data. The 48-hour precipitation is accumulated precipitation from the past 48 hours ending at previous day's midnight and so on for 72-hour precipitation data. Previous day's midnight was chosen as a standard time, as one has to consider the lag time till data availability.

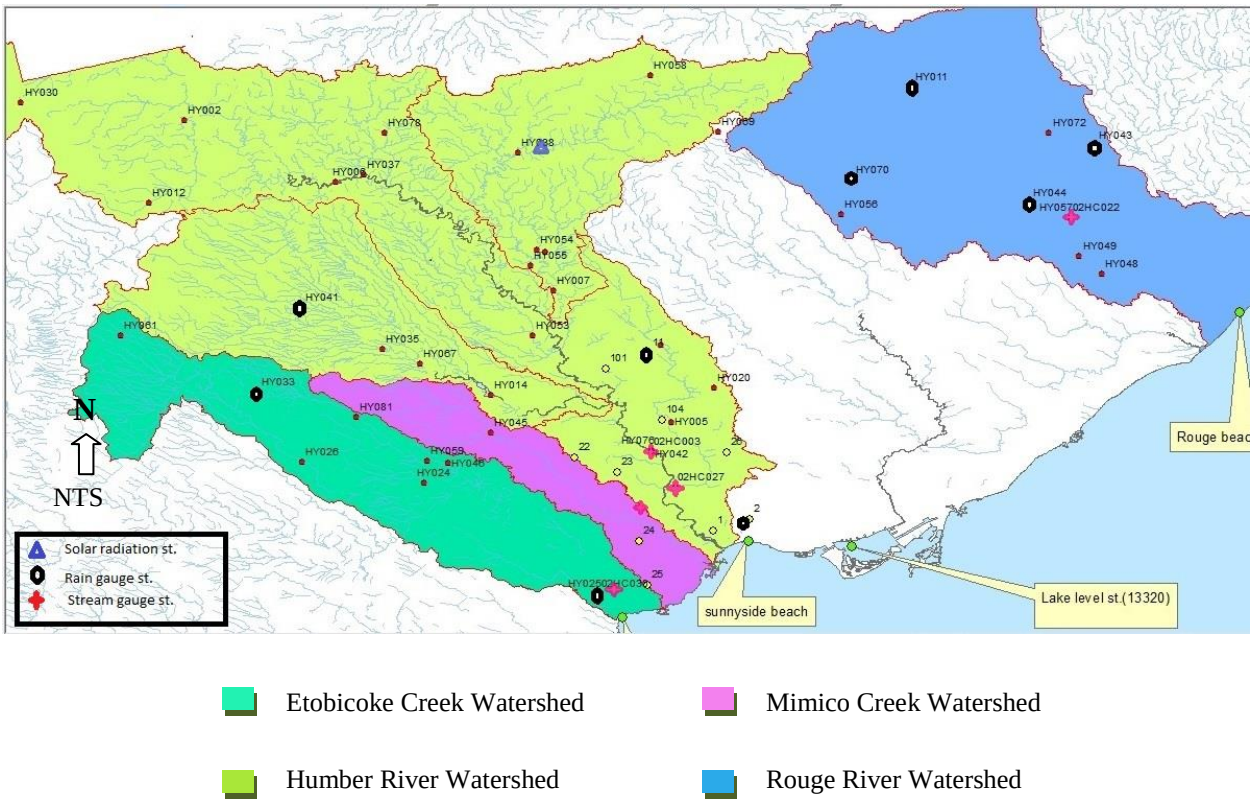


Figure 4-2 Location of TRCA and Toronto Water Stations, solar station and stream gauge location

Solar Radiation: Real-time and historical solar radiation data were obtained from the TRCA's HY039 station located in the Humber river watershed as shown in Figure 4-2, for 2008 to 2012 time period without any missing data. For all beaches, the previous day's average solar radiation ending at midnight was considered which were available at 15 minute interval.

Streamflow: Historical and real-time streamflow data were downloaded through the Environment Canada website for 2008 to 2012 time period. The streamflow data include river stage and a calculated river discharge. Real-time streamflow data was typically recorded at 15 minute intervals and become available after 8 to 9 hours of collection for online access. Therefore the mean daily streamflow can only be calculated for the previous day, midnight to midnight. It is also worth noting that for the Sunnyside beach discharge from the Humber River,

Mimico Creek and Black Creek were incorporated unlike considering only Humber River discharge as per the current MLR model. The streamflow data for all three beaches was collected from the stations are as follows and shown graphically in Figure 4-2.

Sunnyside Beach: (All the stations had missing data for year 2011)

- Humber River at Weston (ON) (02HC003)
- Mimico Creek at Islington (ON) (02HC033)
- Black Creek near Weston (ON) (02HC027)

Marie Curtis Park East Beach:

- Etobicoke Creek below Queen Elizabeth Way (ON) (02HC030)

Rouge Beach:

- Rouge River near Markham (ON) (02HC022)

Wave Height, Wind Direction and Wind Speed*:* Buoy data in Canada is provided by the National Oceanic and Atmospheric Administration's National Data Buoy Center, which provides data from buoy stations operated by Environment Canada. Hourly and historical wind and wave height data for all beaches were obtained from the Buoy station C45139 as shown in Figure 4-1, located on West of Lake Ontario for 2008 to 2012 period. Average significant wave height (meters) (labeled as WVHT on the webpage), average wind direction (the direction the wind is coming from in degrees clockwise from the true North) (labeled as WDIR on the webpage) and average wind speed (m/s) (labeled as WSPD on the webpage) from the previous day midnight to midnight were used in ANNs model development. There were no missing data points in the daily wind and wave data set from 2008 through 2012.

Lake Level: Real-time and historical lake level data for the Great Lakes region is available from Fisheries and Oceans Canada. Hourly lake levels (meters) were available for Lake Ontario from Toronto 13320 station shown in Figure 4-1. For all three beaches daily lake levels for Lake Ontario were calculated by determining the average hourly lake level in the previous 24 hour period, ending 12 pm prior day, similar to the data analysis for the other hourly data sets.

Past E. coli Data: Beach water quality data were obtained from Toronto Public Health for 2008 to 2012 time period. Due to a labor strike by city workers in late June 2009, water quality data were not consistently collected at all beaches for an approximately 4 week period.

Other Variables: Several other parameters were not considered mainly due to not being readily available on a daily basis or being difficult to be predicted or simply because of missing data for considerable period of time. Thus they are not suitable for real-time prediction of beach water quality. These parameters are briefly described as follows:

The turbidity, wind direction and speed, waterfowl counts, wave height category (low, moderate, high) and water temperature are field measurements collected at the time the of E. coli samples were collected at each beach and 2008-2012 time period data was provided by TPH for model development. Turbidity sample was taken by Toronto Water field sampling crews at monitoring locations of particular beach and measured in laboratory. Turbidity data were not collected on regular basis; only 10% to 25% samples for entire summer season were collected, depending on the beach. Compared to turbidity, waterfowl count observations were estimated more frequently at each beach and data were collected for the 60% to 85% of the entire summer season while collecting samples. There is a significant amount of missing data from 2008 to 2010 beach seasons for all this data.

Table 4-2 summarise the data availability for each explanatory variables by the year. Occasionally, E. coli data were missing for a particular day. The probable reason could have been bad weather condition due to which sampling of water at the beach would have not been possible. The approach adopted for solving missing data is discussed in detail in the Section-5.1 Preprocessing Data of the next chapter.

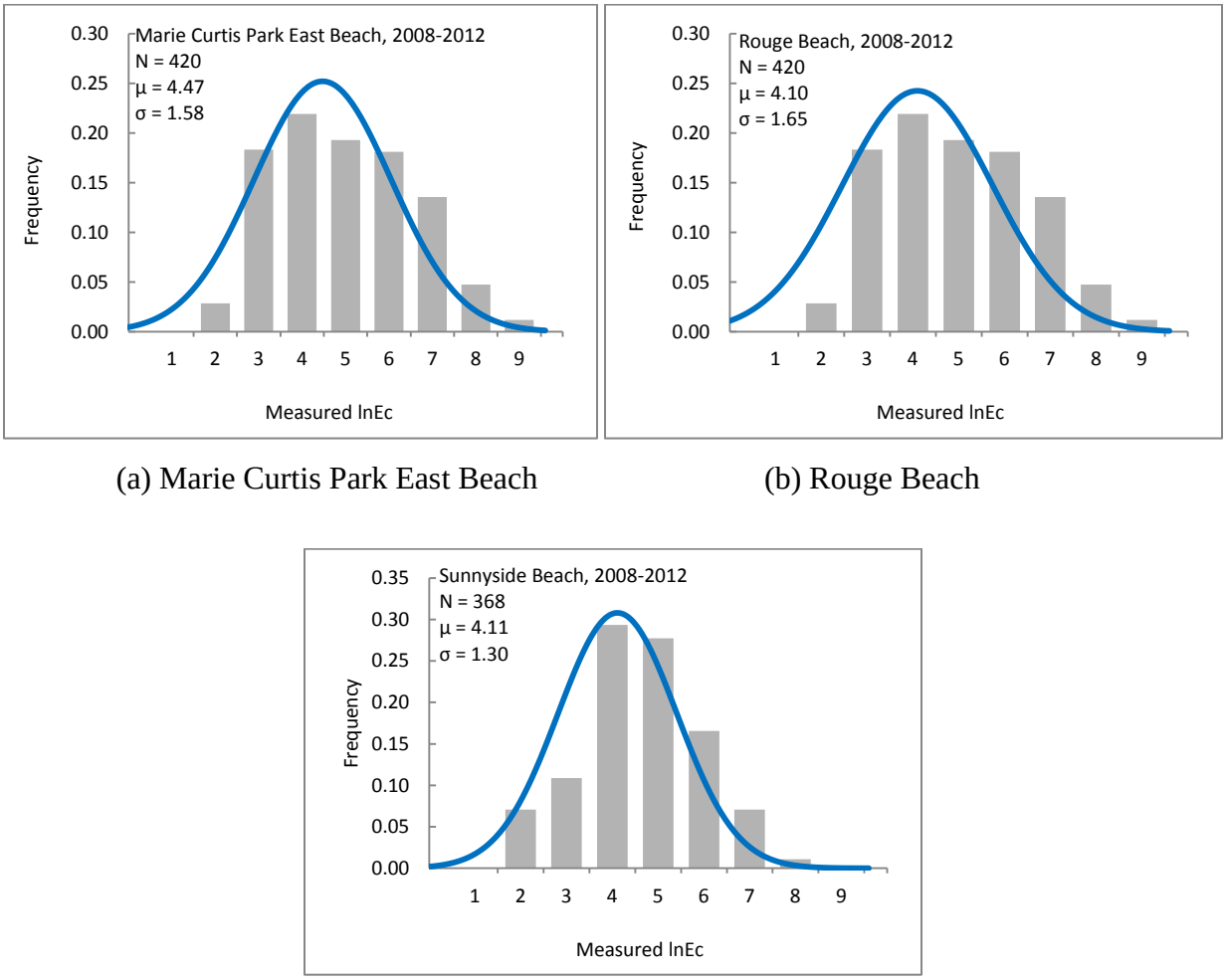
Table 4-2 Data availability of explanatory variables by year

Explanatory Variables	**Applicable Beach**	**Data Availability by the Year:**				
		2008	2009	2010	2011	2012
Previous lnEC	All Three Beach	☑	☑	☑	☑	☑
Flow of Humber River(m^3/s)	Sunny Side Beach	☑	☑	☑	☒	☑
Flow of Mimico Creek(m^3/s)	Sunny Side Beach	☑	☑	☑	☒	☑
Flow of Black Creek(m^3/s)	Sunny Side Beach	☑	☑	☑	☒	☑
Flow of Rouge River(m^3/s)	Rouge Beach	☑	☑	☑	☑	☑
Flow of Etobicoke Creek(m^3/s)	Marie Curtis Park East Beach	☑	☑	☑	☑	☑
Lake level(m)	All Three Beach	☑	☑	☑	☑	☑
Wave ht. (m)	All Three Beach	☑	☑	☑	☑	☑
Wind Direction(deg)	All Three Beach	☑	☑	☑	☑	☑
Wind speed(m/s)	All Three Beach	☑	☑	☑	☑	☑
Solar radiation(MJ/m^2)	All Three Beach	☑	☑	☑	☑	☑
Rainfall from respective rain gauge stations (mm)						
HY041	Sunny Side Beach	☑	☑	☑	☑	☑
TW2	Sunny Side Beach	☑	☑	☑	☑	☑
TW11	Sunny Side Beach	☑	☑	☑	☑	☑
HY044	Rouge Beach	☑	☑	☑	☑	☑
HY070	Rouge Beach	☑	☑	☑	☑	☑
HY025	Marie Curtis Park East Beach	☑	☑	☑	☑	☑
HY033	Marie Curtis Park East Beach	☑	☑	☑	☑	☑

Precious lnEC- Previous day geometric mean of ln E. coli
Streamflow – Previous day's mean streamflow ending at midnight
Lake level – Average hourly lake level of previous day ending at midnight
Buoy Wave Height – Previous 24 hours average wave height ending at midnight
Buoy Wind Speed & Direction – Previous 24 hours average wind speed and direction ending at midnight
Solar radiation – Previous day's average solar radiation ending at midnight
Precipitation – Total precipitation for previous 24 hours ending at midnight, the 48-hour precipitation is accumulated precipitation for previous 48 hours ending at previous day's midnight and so on for 72-hour precipitation data

4.4 Characteristic of Indicator Data

The distributions of the E. coli concentration data at beaches were tested and it was found that the E. coli concentration data of Sunnyside Beach, Rouge Beach and Marie Curtis Park East Beach could be well approximated by lognormal distribution. Figure 4-3, represents the histograms of E. coli concentration in natural logarithm (lnEC) at three beaches,



(a) Marie Curtis Park East Beach (b) Rouge Beach

(c) Sunnyside Beach

Figure 4-3 Histograms of lnEC at (a) Marie Curtis Park East Beach (b) Rouge Beach and (c) Sunnyside Beach

for each beach, the number of data sample (N), the mean (μ) and the standard deviation (σ) of lnEC are displayed, because of the large variability in E. coli concentrations, the natural log of the data is plotted. For this research, E. coli concentration data is transformed to natural logarithm (lnEC) before it is correlated with or predicted from different explanatory parameters.

4.5 Analysis of data relationships

Once data for potential explanatory variables, listed in Table 4-1 is gathered, the data should be reviewed for any obvious errors (e.g negative values, values orders of magnitude outside the suspected range, etc.). Then scatter plots and correlation analysis can be used to detect potential relationships between variables. Scatter plots are obtained to visually investigate the relationship between lnEC and the other environmental variables. Correlation analysis is carried out between lnEC and different hydro-meteorological factors to identify the critical factors that can affect beach water quality. A high correlation does not necessarily occupy a contributory relationship but it does indicate that two parameters are covariant. As the parameters that are covariant, are good candidates for explanatory variables in predictive modelling. This section describes correlation representation performed to guide the selection of input parameters for the ANN models.

Pearson's r is the most commonly used linear correlation coefficient. The Pearson's correlation coefficient between two parameters, x and y is defined as:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad (4\text{-}1)$$

The Pearson's correlation between the natural logarithm transformed E. coli concentration (lnEC) and the parameters is calculated for all beaches. Table 4-3 shows the correlation

coefficients of lnEC with different environmental variables for Sunnyside, Rouge and Marie Curtis Park East Beach which range from 0 (no correlation) to 1 (perfectly correlated). The higher the absolute value of the correlation coefficient, the stronger the relationship between the two variables where negative sign signifies that they are inversely correlated with each other (Helsel and Hirsch, 2002). As shown in Table 4-3, the inclusion/exclusion is based on the results of correlation analysis and practical consideration in real-time implementation of the models. During this selection it should be noted that even though turbidity results showing higher correlation, were omitted due to missing data for a sizeable period of time. Usually a default value of 0.05 is used, but a higher or lower value may be specified to determine statistical significance of the correlation (Helsel and Hirsch, 2002). Generally speaking, variables having correlation of more than 0.1500 were selected but there were some exception made based on the general tendency of all the parameters. For instance, for the Marie Curtis Park East Beach, all the variables revealed less correlation value compared to other two beaches. Due to the same reason, lake level and wave height parameters were chosen as one of the inputs for the modelling.

Table 4-3 Pearson's r correlation between lnEC and explanatory variables for 2008-2012

Sunnyside Beach		**Rouge Beach**		**Marie Curtis Park East Beach**	
N	368	N	420	N	420
Previous lnEC	**0.496**	Previous lnEC	**0.272**	Previous lnEC	**0.232**
Beach Turbidity(NTU)	0.1023	Beach Turbidity(NTU)	0.3781	Beach Turbidity(NTU)	0.1622
Flow of Humber River(m^3/s)	**0.2185**	Flow of Rouge River(m^3/s)	**0.3502**	Flow of Etobicoke Creek(m^3/s)	**0.3145**
Flow of Mimico Creek(m^3/s)	**0.3224**	Lake level(m)	-0.0469	Lake level(m)	**-0.0960**
Flow of Black Creek(m^3/s)	**0.2707**	Wave ht.(m)	**0.3106**	Wave ht.(m)	**0.0936**
Lake level(m)	**-0.1050**	Wind Direction(deg)	**-0.2495**	Wind Direction(deg)	-0.0700
Wave ht. (m)	0.0763	Wind speed(m/s)	**0.1514**	Wind speed(m/s)	0.0666
Wind Direction(deg)	-0.0738	solar radiation(MJ/m^2)	**-0.3022**	solar radiation(MJ/m^2)	**-0.1820**
Wind speed(m/s)	**0.1683**	HY044 Station		HY025 Station	
solar radiation(MJ/m^2)	**-0.2488**	1 day rain(mm)	0.3299	1 day rain(mm)	0.1725
HY041 Station		2 day rain(mm)	**0.3943**	2 day rain(mm)	**0.3280**
1 day rain(mm)	0.2099	3 day rain(mm)	0.3451	3 day rain(mm)	0.2764
2 day rain(mm)	**0.3607**	HY070 Station		HY033 Station	
3 day rain(mm)	0.3250	1 day rain(mm)	0.2782	1 day rain(mm)	0.1965
TW2 Station		2 day rain(mm)	**0.3294**	2 day rain(mm)	**0.3198**
1 day rain(mm)	0.2000	3 day rain(mm)	0.2903	3 day rain(mm)	0.2589
2 day rain(mm)	**0.4091**	-		-	
3 day rain(mm)	0.2586	-	-	-	-
TW11 Station		-	-	-	-
1 day rain(mm)	0.1298	-	-	-	-
2 day rain(mm)	**0.2704**	-		-	-
3 day rain(mm)	0.1634	-	-	-	-

*Pearson's r correlation values that are significant (p>0.05) are in **bold italics***

4.5.1 Sunnyside Beach

Sunnyside Beach is located along the Toronto shoreline of Lake Ontario and includes approximately 1.3 kilometers of beach. The beach area is protected by a system of break walls

that are located between 50 to 175 meters away from the shoreline. Total of seven Toronto Water monitoring stations are staggered along the beach. Three storm sewer outfalls are located along the beach. These three outfalls only discharge water in large storm events, approximately once per year, since the stromwater is being intercepted by the Western Beaches Tunnel, located east of the Sunnyside Beach.

The Pearson's r correlation coefficients were computed for the natural log transformed E. coli concentration and the suite of potential explanatory variables shown in Table 4-3. Strong and statistically significant correlations with streamflow in the Humber River, Mimico Creek and Black Creek, cumulative of last 2-day rain, wind speed and previous day E. coli counts were observed for all years in the period of record. Inverse correlations were observed with lake level and solar radiation. Graphical representation in terms of scatter plots of E. coli and several potential explanatory variables suggest relationships between natural log transformed E. coli and last 2-day rain, wind direction, previous day E. coli concentrations, wind speed, wave height, streamflow, solar radiation as shown in Figure 4-4 and Figure 4-5.

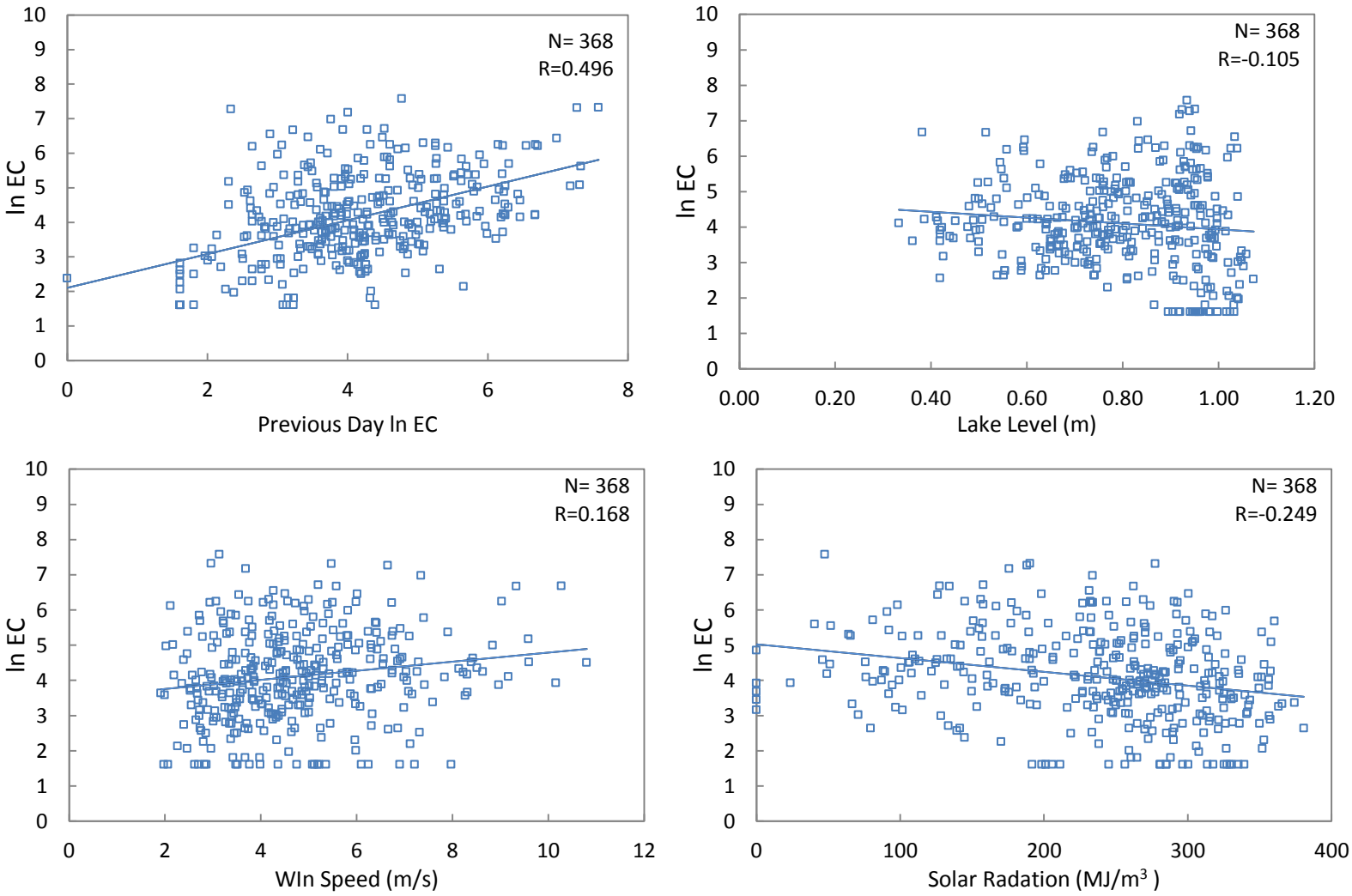


Figure 4-4 Scatter plots of lnEC with (a) previous day E. coli, (b) lake level, (c) wind speed and (d) solar radiation, at Sunnyside Beach, 2008-2012

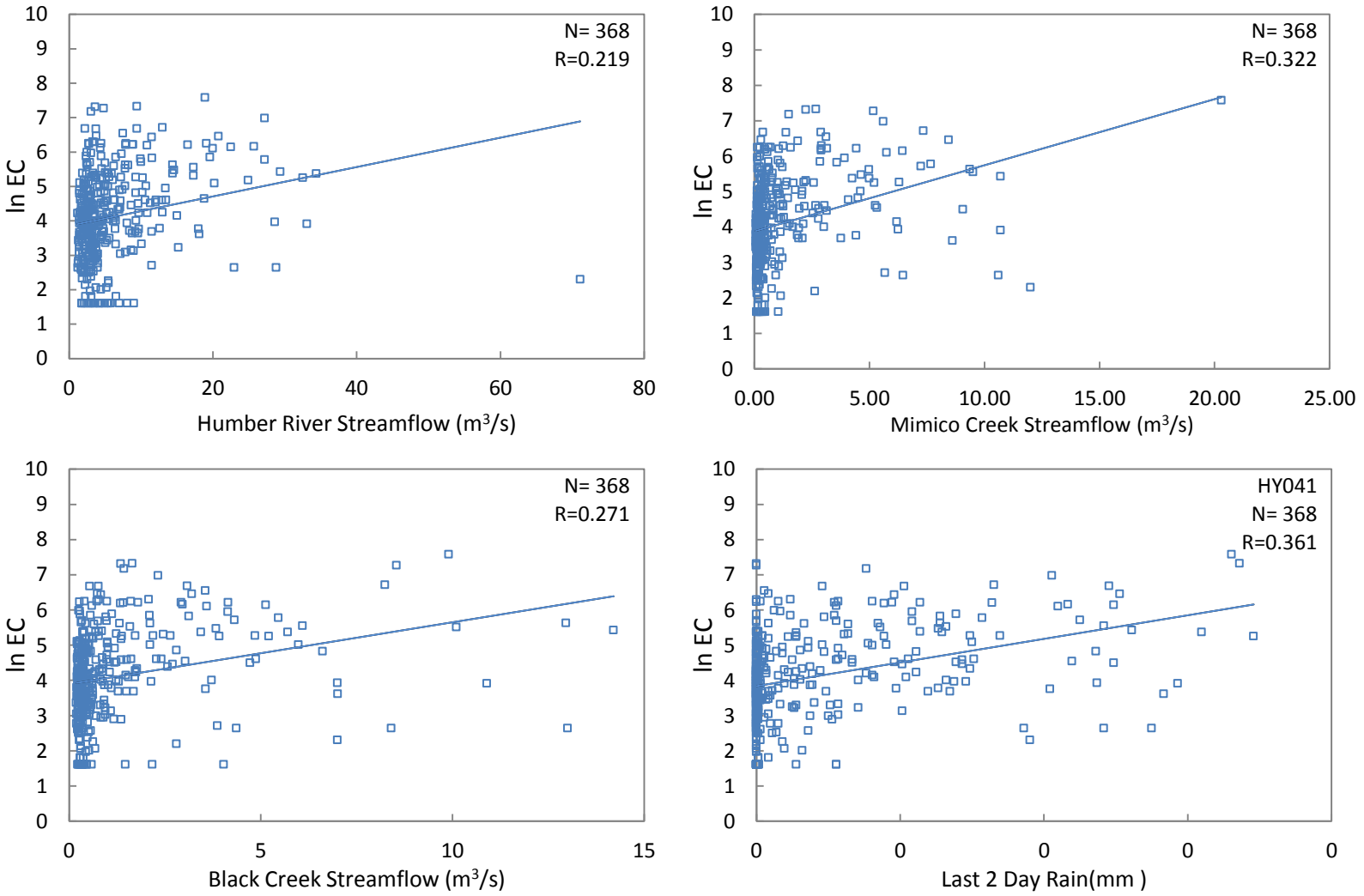


Figure 4-5 Scatter plots of lnEC with (a) Humber river streamflow, (b) Mimico creek streamflow, (c) Black creek streamflow and (d) HY041 station last 2 day rain, at Sunnyside Beach, 2008-2012

Based on graphical analysis as shown in Figure 4-4 and Figure 4-5 and the results of the correlation analysis, it appears that several parameters are potential explanatory variables for ANN predictive model development, especially streamflow and Last 2-day rainfall, solar radiation, wind speed and past day E. coli concentrations.

4.5.2 Rouge Beach

Rouge Beach is the easternmost beach in Toronto and is located on just southwest of the mouth of the Rouge River. The beach is approximately 200 meters long with five sampling locations staggered along the shoreline. There is no stromwater outfalls located near the beach.

The Pearson's r correlation coefficient was computed for the natural log transformed E. coli concentration and the suite of potential explanatory variables shown in Table 4-3. Moderate and statistically significant correlations with previous day E. coli, last 2-day rain, wave height and Rouge river streamflow were observed for all years in the period of record. As indicated in the table, low correlations were calculated between E. coli concentrations and lake level, last day and last 3-day rain. High, inverse correlations were observed with wind direction and solar radiation. Scatter plots of E. coli and several potential explanatory variables suggest relationships between natural log transformed E. coli and last 2-day rain, wind direction, previous day E. coli concentrations, wind speed, wave height, streamflow, solar radiation as shown in Figure 4-6 and Figure 4-7.

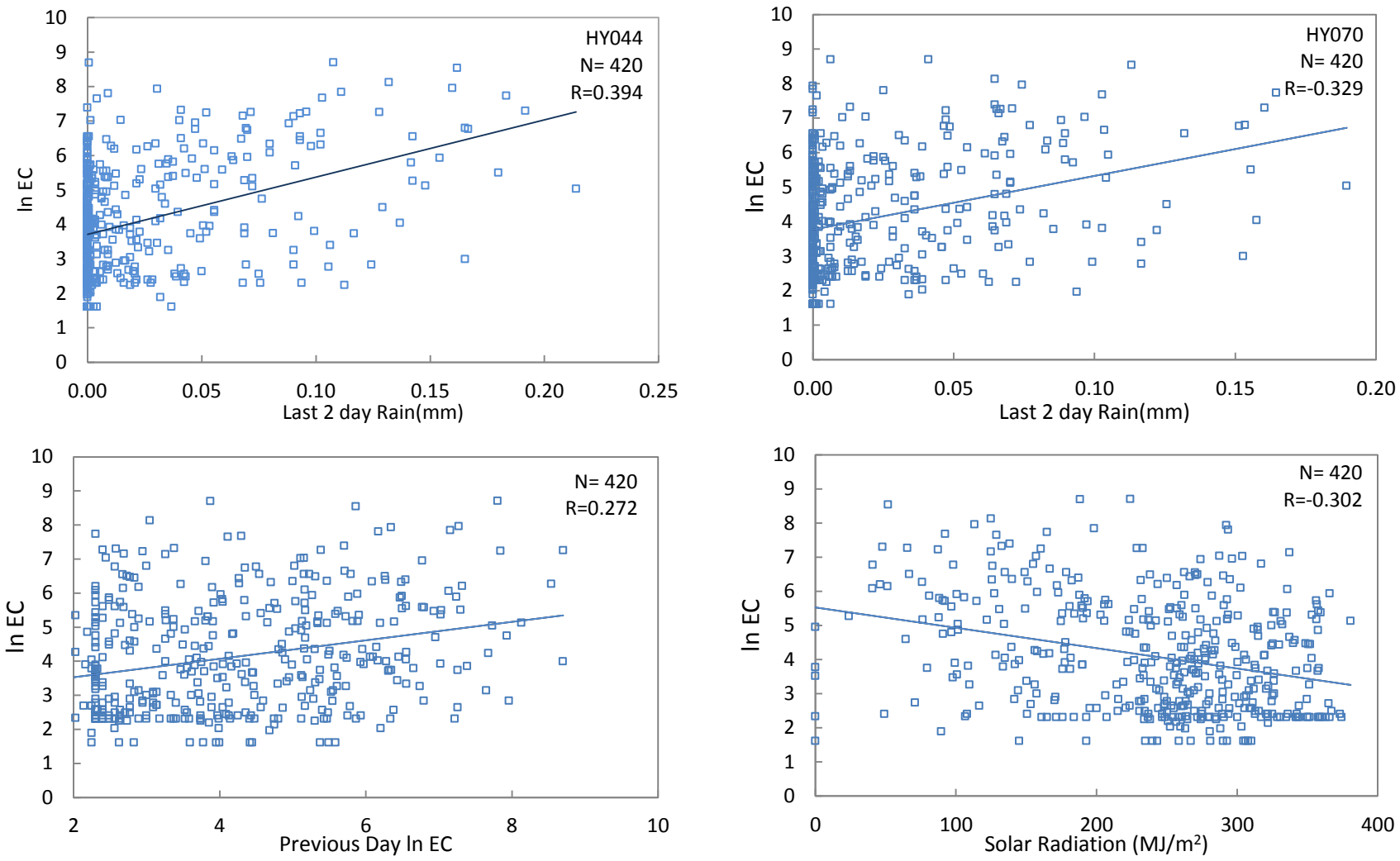


Figure 4-6 Scatter plots of lnEC with (a) last48 Hours rain of HY044, (b) last48 Hours rain of HY070, (c) Previous day E. coli and (d) solar radiation, at Rouge Beach, 2008-2012

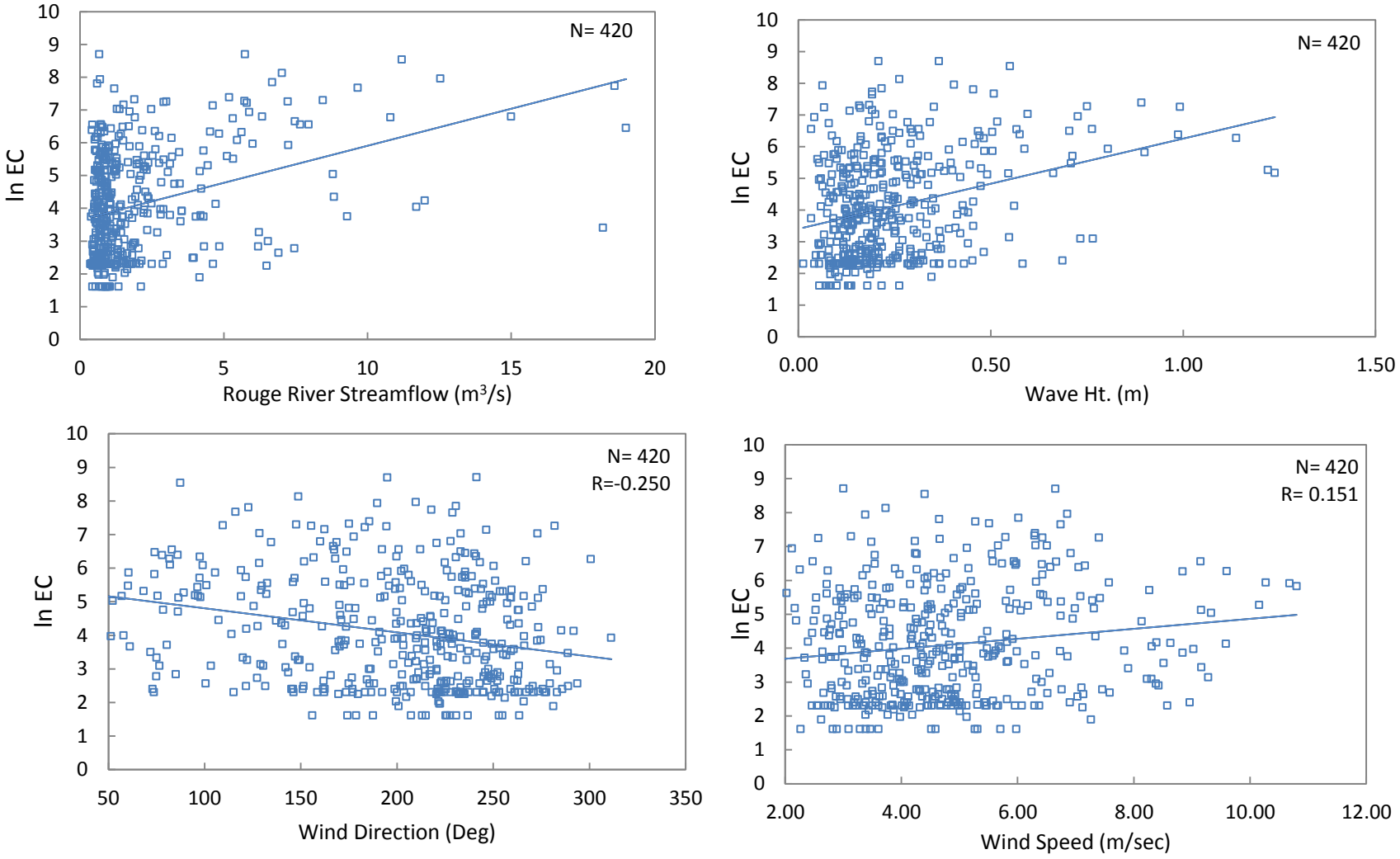


Figure 4-7 Scatter plots of lnEC with (a) streamflow, (b) wave height,

(c) wind direction and (d)wind speed, at Rouge Beach, 2008-2012

Based on the correlation coefficient calculations and graphical analysis, last 2-day rain, wave height, wind direction, solar radiation, past day E. coli concentrations and streamflow appear to be potential explanatory variables for predictive model development.

4.5.3 Marie Curtis Park East Beach

Marie Curtis Park East Beach is the westernmost public beach in Toronto on the Lake Ontario shoreline. The beach extends approximately 150 meters, from the outlet of Etobicoke Creek to the edge of Marie Curtis Park. Toronto Water monitors five sampling locations on this beach. One stromwater outlet is located in Etobicoke Creek approximately 0.5 kilometer upstream of the discharge into Lake Ontario. A second stromwater outlet is located at the end of 40^{th} street approximately 100 meters up the coast, northeast of Marie Curtis Park East Beach. In addition, the G. E. Booth (Lakeview) Wastewater Treatment Facility is located nearby and the final effluent discharges to Lake Ontario through a pipe reaching 1,250 meters offshore.

Table 4-3 shows the Pearson's r correlation coefficients among different variables for Marie Curtis Park east Beach for the time period of 2008-2012. As an example of the association of lnEC with environmental variables, Figure 4-8 and Figure 4-9 show the scatter plots of lnEC with previous day's rainfall, previous day's solar radiation, lake level and streamflow.

As indicated in the Table 4-3, for all years considered streamflow at Etobicoke Creek and past 2-day rain showed strong correlation with E. coli concentrations regardless of the time period considered. For the entire period of record, moderate positive correlations were observed with past day E. coli concentrations and wave height and moderate negative correlation with lake level and solar radiation. Based on the correlation and graphical analysis, it appears that several parameters are potential explanatory variables for predictive ANN model development,

especially streamflow, past 2-day rainfall, solar radiation, wave height and past day E. coli concentrations.

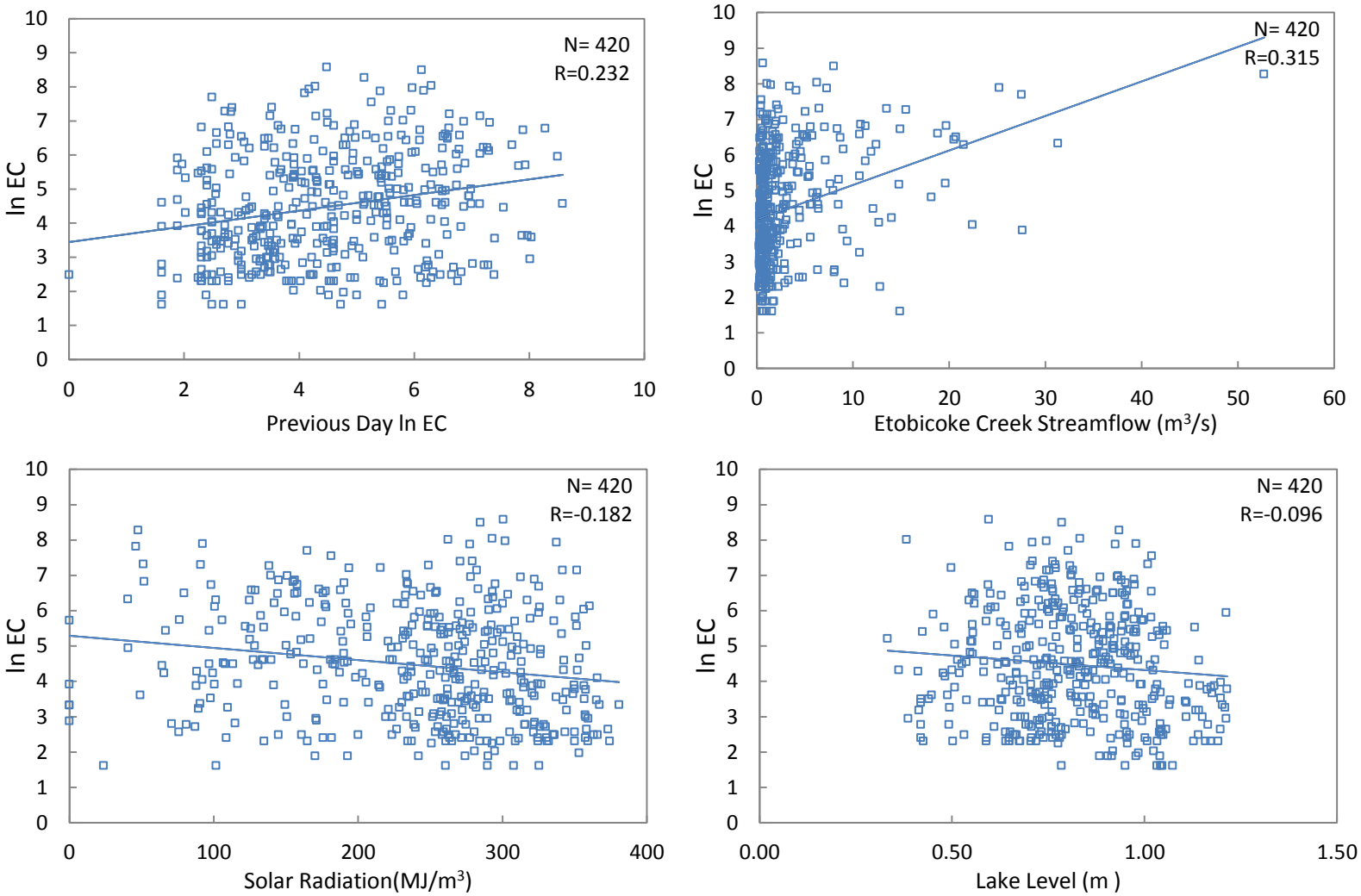


Figure 4-8 Scatter plots of lnEC with (a) previous day E. coli, (b) Etobicoke creek streamflow, (c) solar radiation and (d) lake level, at Marie Curtis Park East Beach, 2008-2012

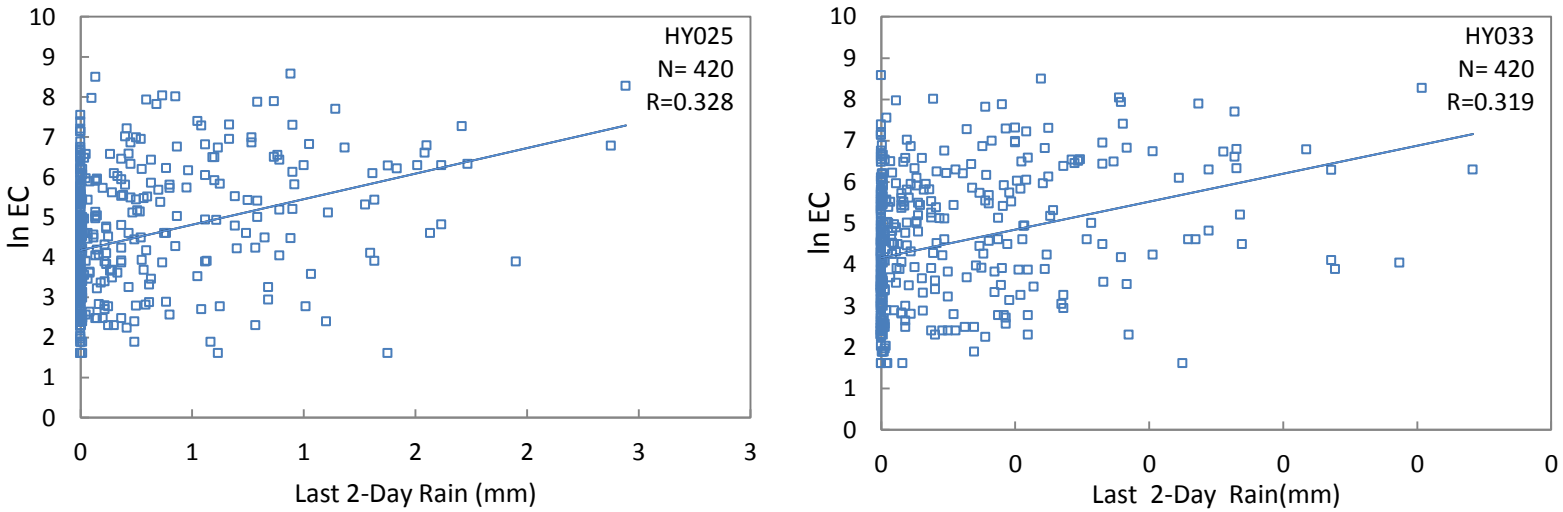


Figure 4-9 Scatter plots of lnEC with last 2 days rain for (a) HY025 and (b) HY033

4.6 Concluding Remarks

The main aim of this statistical analysis described in this chapter is to identify potential candidate variable for the model development. One goal of this statistical analysis described in this chapter is to identify potential candidate variables for predictive model development. It appeared that not a one set of explanatory variables will be suitable for all beaches in this study. For these three Toronto beaches in general, previous day E. coli concentrations, past 48 hrs rain, solar radiation and streamflow have consistently significant correlations with E. coli concentrations. Depending on the particular beach, different variables such as wave height, lake level, wind direction and wind speed may emerge as significant explanatory variables in the ANNs model development process. Despite, turbidity being one of the important explanatory variables, the same was neglected in model generation due to substantial amount of missing data. Some of the rain gauge stations were also omitted due to the same reason. Both transformed and untransformed data will be tested for analysis (i.e. lnEC and ln of other explanatory variables, E. coli and ln of other explanatory variables and lnEC and ln of selected explanatory variables) and found that transformation of the data (i.e., taking the natural log of the E. coli data) can improve the linearity of the relationship between variables. It was generally seen that the lnEC proportionately increases with the increase in rainfall and streamflow and decreases with the increase in solar radiation. Different beaches also have different characteristics, e.g. the correlation between lnEC and wind direction is higher at Rouge Beach but lower at Marie Curtis Park East Beach. The study of scatter plots reveal potential causative factors that affect the beach water quality, at the same time the great scatter in the data suggests that beach water quality forecast is possible but a challenging task. Some study also determined the use of analysis of

variance (ANOVA) methods for section of categorical variables, like wind direction, if it is there.

Table 4-4 summarizes the parameters that showed statistically significant linear correlations with E. coli at each of the beaches considered.

Table 4-4 Important explanatory variables at Toronto beaches based on graphical analysis and statistically significant Pearson's r correlations (p>0.05)

Explanatory Variables	**Nomenclature**	**Sunnyside Beach**	**Rouge Beach**	**Marie Curtis Park East Beach**
Previous day ln E. coli	*pr. lnEC*	☑	☑	☑
Flow of Humber River(m^3/s)	*st.fl.*	☑		
Flow of Mimico Creek(m^3/s)	*st.fl.*	☑		
Flow of Black Creek(m^3/s)	*st.fl.*	☑		
Flow of Rouge River(m^3/s)	*st.fl.*		☑	
Flow of Etobicoke Creek(m^3/s)	*st.fl.*			☑
Lake level(m)	*l.l.*	☑		☑
Wave ht. (m)	*w.ht*		☑	☑
Wind Direction(deg)	*w.dir*		☑	
Wind speed(m/s)	*w.spd*	☑	☑	
solar radiation(MJ/m^2)	*slr*	☑	☑	☑
Rainfall from respective rain gauge stations (mm)				
Cumulative last 24 rain(mm)	*past 24hrs rain*			
2 day rain(mm)	*past 48hrs rain*	☑	☑	☑
3 day rain(mm)	*past 72hrs rain*			

*st.fl-streamflow of Rouge River, pr.E. coli-previous day E. coli count, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, past 48hrs rain(3)**-past 48hrs rain of station HY041,T.W.2 and T.W.11,st.fl(3)*-streamflow of Humber river, Mimico creek and black creek, pr.E. coli-previous day Ln E. coli count, Ln-natural logarithm, w.spd-wind speed w.ht-wave height ,L.L-lake level, slr-solar radiation*

5 MODEL DEVELOPMENT AND IMPLEMENTATION

Designing an ANN model follows a number of systemic procedures. In general, there are five basic steps of modeling as shown in Figure 5-1.

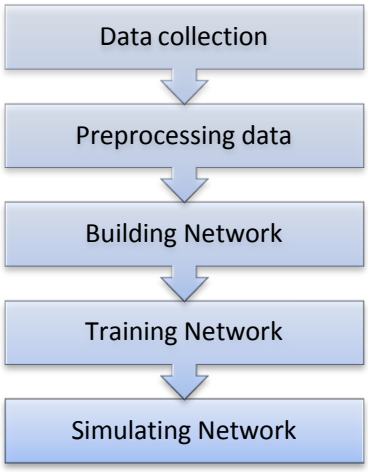


Figure 5-1 Basic flow for designing artificial neural network model

The ANN models were designed and implemented using "The MATLAB R2013a Neural Networks Toolbox" by the Mathworks Inc. The Neural Network Toolbox consists of Graphical User Interface-GUI as a replacement of writing down codes. The users have number of options to choose from variety of algorithms and function. Indeed this tool is very reliable and contains a variety of algorithms and functions to be tested.

5.1 Preprocessing data

After data collection from particular sources, preprocessing procedures are conducted to train the ANNs more efficiently. These procedures are: 1) solve the problem of missing data and 2) normalize data.

As mentioned in Section 4.3, occasionally E. coli data were missing for a particular day. These missing data were replaced by the average of neighbouring values (Mass and Ahlfeld, 2007, Shamisi et al., 2011).

Normalization procedure before presenting the input data to the network is generally a good practice, since mixing variables with large magnitudes and small magnitudes will confuse the learning algorithm on the importance of each variable and may force it to finally reject the variable with the smaller magnitude. In this way more reliable predictions can be made, hence all input data are linearly normalized into a particular range prior to applying transfer functions (Tufail et al., 2008). As the input data is preprocessed during model development, all new inputs applied thereafter to the trained network were preprocessed (normalized).

5.2 Building and Training the Network

In this research, 5-year data from 2008 to 2012 time period has been used for generating ANN models. The bias error decreases when the model size increases during training time. By modifying the design parameters, changing the size of input and target parameters from 5-year data during training and simulation periods and repeating the procedure will give the best results. From literature review it was found that, more data size during the training time helps to get better results for ANNs(Gershenson, 2003). On that basis for this research work data from jun-july-aug-2008 to jun-july-2012 were used to train models and aug-2012 data were kept on side for simulation of models.

Deciding possible group of explanatory variables is one of the most crucial steps for building the network. Using Correlation analysis result explained in Table 4-3 and graphical

analysis, possible explanatory variable combinations were formulated and has been listed in Table 5-1, Table 5-2 and Table 5-3 for each beach.

Table 5-1 Possible explanatory variable combinations for Sunnyside beach

No.	Explanatory variable combinations
1	ln st.fl(3)*,ln w.ht,ln w.spd,slr,past 48hrs rain(3)**
2	st.fl(3),w.ht,w.spd,slr,past 48hrs rain(3)
3	st.fl(3),w.spd,slr,rain 48hrs(3),pr. lnEC
4	st.fl(3),w.spd,slr,HY041 past 48hrs rain,pr. lnEC
5	Humber st.fl,w.spd,slr,HY041 past 48hrs rain,pr. lnEC
6	st.fl(3),w.spd,slr,past 48hrs rain(HY041,T.W.2),pr. lnEC,l.l..
7	st.fl(3),w.spd,slr,past 48hrs rain(3),pr. lnEC,l.l..
8	st.fl(3),w.spd,slr,past 48hrs rain(3),l.l..
9	st.fl(3),w.spd,past 48hrs rain (HY041,T.W.2),l.l..
10	st.fl(3),w.spd,slr,HY041 past 48hrs rain,pr. lnEC,l.l.
11	st.fl(3),w.spd,slr,HY041 past 48hrs rain,pr. lnEC
12	st.fl(Humber,Mimico),w.spd,slr,HY041 past 48hrs rain
13	st.fl(Humber,Mimico),w.spd,slr,HY041 past 48hrs rain,pr. lnEC
14	st.fl(3),slr,past 48hrs rain(HY041,T.W.2),pr. lnEC,l.l.,turbidity
15	st.fl(Humber,Mimico),slr,past 48hrs rain(HY041,T.W.2),pr. lnEC,turbidity
16	Humber st.fl,HY041 past 48hrs rain,pr. lnEC. (same combination as in MLR)

*past 48hrs rain(3)**-past 48hrs rain of station HY041,T.W.2 and T.W.11, st.fl(3)*-streamflow of Humber river, Mimico creek and black creek, pr. lnEC-previous day E. coli count, ln-natural logarithm, w.spd-wind speed w.ht-wave height ,l.l.-lake level, slr-solar radiation*

Table 5-2 Possible explanatory variable combinations for Rouge beach

No	Explanatory variable combinations
1	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC
2	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY044,HY070),pr. lnEC
3	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC
4	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044),pr. lnEC
5	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY070),pr. lnEC
6	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044,HY070)
7	st.fl,w.dir,slr,past 48hrs rain(HY044,HY070)
8	st.fl,slr,past 48hrs rain(HY044,HY070)
9	st.fl,w.dir,w.ht,past 48hrs rain(HY044,HY070)
10	w.dir,w.ht,slr,past 48hrs rain(HY044,HY070)
11	st.fl,w.dir,w.ht
12	st.fl,w.dir,pr. lnEC. (MLR)

st.fl-streamflow of Rouge River, pr. lnEC-previous day E. coli count, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, slr-solar radiation

Table 5-3 Possible explanatory variable combinations for Marie Curtis Park East beach

No	Explanatory variable combinations
1	ln st.fl,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
2	st.fl,w.ht,w.dir,slr,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
3	st.fl,w.ht,w.spd, w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
4	st.fl,w.ht,w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
5	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
6	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC
7	st.fl,past 48hrs rain (HY025,HY033),pr. lnEC

8	w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
9	st.fl,w.ht,slr,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
10	w.ht,slr,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.
11	st.fl,slr,past 48hrs rain (HY025,HY033),pr. lnEC
12	st.fl,w.ht,slr,pr. lnEC,l.l.
13	st.fl,w.ht,past 48hrs rain(HY025),pr. lnEC,l.l.
14	st.fl,w.ht,past 48hrs rain(HY033),pr. lnEC,l.l.
15	st.fl,past 48 hr rain(HY025),w.dir (MLR)

st.fl-streamflow of Etobicoke creek, pr. lnEC-previous day E. coli count, ln-natural logarithm, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, l.l.-lake level, slr-solar radiation

Each model generated using these combinations were evaluated with a multilayer feedforward back-propagation neural network with three layers (input, hidden and output) where the output layer has one neuron (the predicted value of ln E. coli concentration).

Out of several different back-propagation training algorithms, most commonly used algorithms namely *trainrp*, *traingd* and *trainlm* were investigated during the model development, as shown in Table 5-4. Based on trial and error most efficient algorithm was chosen among these three.

Table 5-4 Training algorithms trialed during ANN model development (Beale et al., 2013)

Training Functions	**Brief Explanation**
trainrp	Resilient Backpropagation
traingd	Gradient Descent Backpropagation
trainlm	Levenberg – Marquardt

Transfer functions in ANNs have limited ranges, i.e. (0, 1) for the logistic function and (-1, 1) for the hyperbolic tangent sigmoid function (Beale et al., 2013). Logistic (*logsig*) and tangent sigmoid (*tansig*) functions are used as the transfer function from input layer to hidden

layer; no transfer function is used from hidden layer to output layer. Linear scaling was used to transform the input data for this study to (0.1, 0.9) for use with the logistic function and (-0.9, 0.9) for use with the tangent sigmoid function using *mapminmax* script in MATLAB.

There are several guidelines or "rules of thumb" for the selection of hidden nodes. The number of hidden nodes is the square root of the product of the number of input parameters and number of target parameters. The final number of nodes in the hidden layer was adjusted by trial and error after the best set of inputs has been chosen (Mass and Ahlfeld, 2007).

In summary, the following combinations were possible to generate the best ANN architecture for a particular beach E. coli modelling,

1. Input parameter combinations = 12 to 16 combinations depending on the beach

2. Transfer function = 2 transfer function (*tansig* and *logsig*)

3. Back-propagation algorithms for training of ANNs = 3 training algorithms

4. Number of hidden neurons in the hidden layer = 5, 10, 20, 30, 50, 100

The same has been graphically represented in Figure 5-2

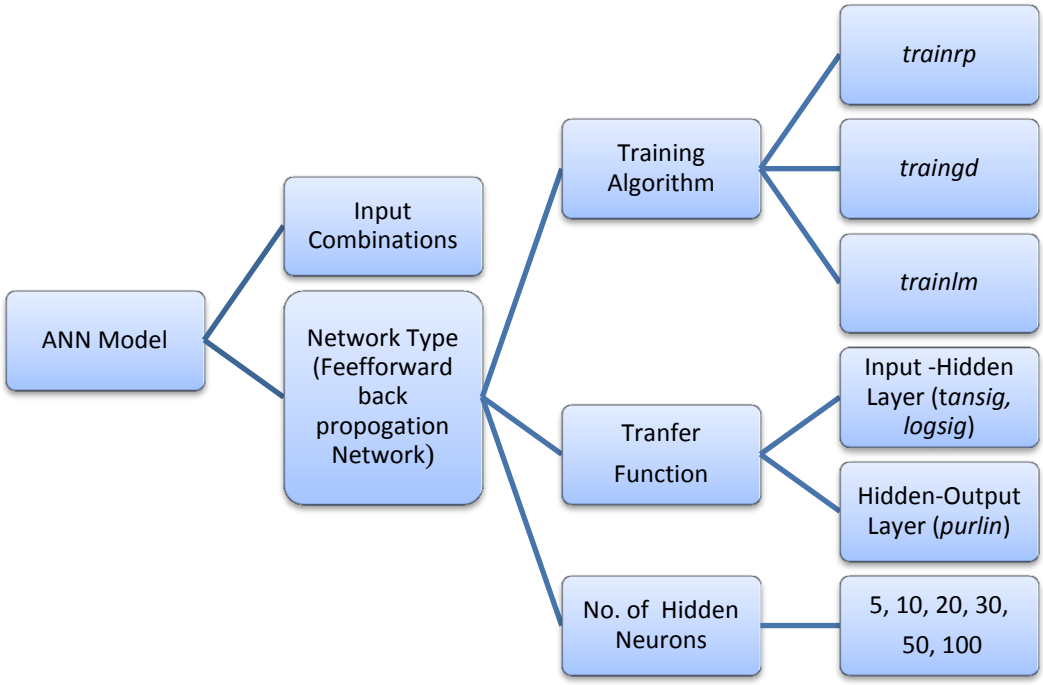


Figure 5-2 Building of ANN Model

Once the models have been generated, the most important process of training the network is followed. For the supervised training of ANNs, the division of data is typically done so that there are a greater number of training sets than simulation sets. Random Data Division (*dividerand*) was used for ANN training in order to allow the maximum use of the data for training. This is a typical data split for ANN model development. The data for each beach are randomly divided into training, validation and testing period in a ratio of 60:20:20 (Beale et al., 2013).

For this research work, the Levenberg-Marquardt algorithm (*trainlm*) was selected as training function with feedforward network's default performance function mean square error (MSE) for all beaches. Random initial weights were assigned to network interconnections. The network was trained through iterations via the gradient-descent method with momentum correction; the model learning rate was 0.1 and the momentum term was 0.01 as default values of feedforward network. To prevent data over fitting, the model learning is stopped if the error

between the prediction and the validation data increases for 1000 iterations to avoid data overfitting using trial and error method.

Using all these parameters as one of the variables total of 32 ANN models for Sunnyside beach, 36 models for Rouge beach and 33 models for Marie Curtis Park East beach were generated. All these models have been listed in Appendix A in Table A-1, Table A-2 and Table A-3 for Sunnyside beach, Rouge beach and Marie Curtis Park East beach respectively. It should be noted that as the results obtained during testing the models using log transformed input variables except previous day E. coli were poor and thus discarded and not listed in the above mentioned tables.

Once the models have been built and trained, they are further simulated by introducing new sets of inputs and then outputs are compared with targets using different assessment methods as discussed in the next section.

5.3 Simulation of the Network

Simulation of the artificial neural network is very important step in order to make sure the trained network can generalize well and produce desired outputs when new data is presented to it. For this research, E. coli data and other required explanatory variables for particular beach of August 2012 were used for simulating the ANN models. The performance of ANNs should be assessed both quantitatively and qualitatively, as discussed in the following sub-sections.

5.3.1 Quantitative Assessment

Several quantitative measures of model performance should be used because of the incapability of a single criterion to briefly evaluate model performance (Harmel and Smith, 2007, Krause et al., 2005, Legates et al., 1999). The model performance is assessed

quantitatively by calculating the (1) Correlation coefficient (CORR) Eq. (4-1) and (2) Root mean square error (RMSE) between the observed and predicted lnEC. RMSE is a measure of the deviation of the predicted values to the observations, calculated as:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(y_k - Y_k)^2}{N}} \qquad (5\text{-}1)$$

RMSE uses the units of the variable of interest and describes the average error of the model. Lower values of RMSE indicate better agreement between simulated results and observed data (Mass and Ahlfeld, 2007, Legates et al., 1999).

5.3.2 Qualitative Assessment

Visual inspection and interpretation of scatter plots of observed data and model output provide qualitative information. For example, questions such as does the model always over or under estimates values, or does it perform well within one range of values and poorly outside that range can be answered. The goal of this research is to assess model performance relative to the water quality standards or guidelines for E. coli that is followed in Toronto, the location of the study area. Therefore, the actual value of the modeled concentration is less important than whether or not it falls in a certain concentration range. As a result, the numbers of correct categorizations are defined as the number of occurrences for which the observed and predicted indicator organism concentrations are lying in the same group (i.e. below threshold or above threshold). The number of false positives (i.e., when the predicted value is above the threshold concentration, but the observed value is below) and false negatives (i.e., when the predicted value is below the critical concentration, but the observed value is above) are also useful in evaluating model performance when the intended application of the model does not

necessitate the precise estimate of an indicator organism concentration but does requires the ability to predict relative to a critical concentration, such as a water quality standard. (Thoe et al., 2012, Mass and Ahlfeld, 2007) also incorporated this type of model performance criteria in their work.

5.4 Summary

As discussed in this chapter, pre-processing is a crucial step to follow before ANN model generation. Efficiency of ANN models depends upon the number of different parameters such as input parameters combinations, training algorithms, training functions and the number of neurons. Lastly, the model performance can be assessed with the help of qualitative and quantitative methods. As the research work focuses on beach water quality based on threshold, 100 E. coli count / 100 ml of beach water, the actual value of the modeled concentration is less important than whether or not it falls in a certain concentration range.

6 DISCUSSION OF RESULTS

6.1 Introduction

This chapter presents the results generated by the ANN nowcasting models for E. coli concentrations in water at the three Toronto beaches investigated in this study. The results of this research contain graphical representations and statistical analysis of the data. Results generated from ANN models are compared with the actual historic data of E. coli concentrations provided by the Toronto Public Health department. Training results and simulation results of best ANN models for each study beach are presented in this chapter.

Different types of input group, transfer function, hidden neuron and training algorithm were examined during model development. Those models are contained in Table A-1 for Sunnyside beach, Table A-2 for Rouge beach and Table A-3 for Marie Curtis Park East beach in Appendix A. The best ANN models were screened out for nowcasting of E. coli based on performance criteria discussed in Section 5.2 were used to check the prediction capability of each developed model. As it might be recalled, while running the ANN models, the input data were divided into three sets as training, validation and testing, although model screening was performed taking the RMSE and CORR value of the training data (training + validation + test) set and the performance for the simulation data (the last 31 days E. coli data) since this set is never used for training of the ANN. As a results for this particular research work, training data set and simulation data set were used for screening and prediction performance, the threshold for screening out best models among all developed models was selected based on RMSE and visual assessments.

6.2 General Results for all Three Beaches

The performance statistics presented in Table 6-1, Table 6-2 and Table 6-3 provide an overall assessment of the ANN models investigated for the effect of input normalization, different training algorithms, training functions and the number of hidden neurons. This investigation was common for all three beaches regardless of the combination of input parameters used. All tables show the RMSE and CORR of each model's performance. As we know, CORR value reveals underlying relationships between the output data and target data. Correlation analysis becomes more realistic when a large amount of data is available for comparison. Due to particular reason, simulation data CORR being smaller sample size (31 samples) was not considered as a decision making criteria.

Table 6-1 Comparison of Sunnyside Beach ANN models

No	Model	Transfer function	Training algorithm	No. of neurons	Input Normalization	Training data		Simulation data	
						RMSE	CORR	RMSE	CORR
1	M1	tansig purelin	trainlm	30	NO	1.420	0.320	1.276	0.262
2	M5	tansig purelin	trainlm	20	YES	0.830	0.700	0.856	0.356
3	M5F	tansig purelin	trainlm	30	YES	1.012	0.598	0.943	0.196
4	M5A	tansig purelin	trainlm	50	YES	0.819	0.786	0.893	0.307
5	M5B	tansig purelin	trainlm	100	YES	0.949	0.698	0.905	0.266
6	M5C	logsig purelin	trainlm	30	YES	0.999	0.657	0.899	0.297
7	M5D	tansig purelin	trainrp	30	YES	0.945	0.701	0.887	0.329
8	M5E	tansig purelin	traingd	30	YES	1.048	0.612	0.917	0.218
9	MLR	tansig purelin	trainlm	30	YES	0.613	0.629	1.091	0.157

Table 6-2 Comparison of Rouge Beach ANN models

No	Model	Transfer function	Training algorithm	No. of neurons	Input Normalization	Training data		Simulation data	
						RMSE	CORR	RMSE	CORR
1	R1	tansig purelin	Tarinlm	20	NO	41.47	0.0624	55.122	0.080
2	R5	tansig purelin	Tarinlm	20	YES	0.861	0.6635	1.02	0.529
3	R5A	logsig purelin	Tarinlm	20	YES	0.75	0.583	0.87	0.574
4	R5B	tansig purelin	Tarinlm	50	YES	0.977	0.640	1.33	0.678
5	R5C	logsig purelin	Tarinlm	50	YES	0.86	0.62	0.9	0.6702
6	R5C	tansig purelin	Tarinrp	50	YES	0.95	0.558	1.11	0.4
7	R5D	logsig purelin	Tarinrp	50	YES	0.9	0.604	1.2	0.591
8	MLR	tansig purelin	Tarinlm	50	YES	0.747	0.571	2.53	0.424

Table 6-3 Comparison of Marie Curtis Park East Beach ANN models

No	Model	Transfer function	Training algorithm	No. of neurons	Input Normalization	Training data		Simulation data	
						RMSE	CORR	RMSE	CORR
1	M1	tansig purelin	Trainlm	20	YES	0.732	0.578	2.420	0.205
2	M1A	tansig purelin	Trainlm	30	YES	0.740	0.450	1.670	0.370
3	M1B	tansig purelin	Trainlm	50	YES	0.782	0.501	1.246	0.480
4	M1C	tansig purelin	Trainlm	100	YES	0.750	0.430	2.130	0.346
5	M2	logsig purelin	Trainrp	20	YES	0.700	0.320	1.560	0.180
2	M2A	logsig purelin	Trainrp	30	YES	1.243	0.3160	1.859	0.257
7	M2B	logsig purelin	Trainrp	50	YES	1.022	0.239	2.130	0.298
8	M2C	logsig purelin	Trainrp	100	YES	0.893	0.320	1.954	0.112
9	M12	tansig purelin	Trainlm	20	YES	0.794	0.646	1.570	0.150

10	M12A	tansig purelin	Trainlm	50	YES	0.812	0.666	1.440	0.385
11	M12B	logsig purelin	Trainlm	50	YES	0.809	0.616	1.860	0.057
12	M12C	tansig purelin	Trainlm	100	YES	1.230	0.617	3.057	0.016
13	MLR	tansig purelin	Trainlm	50	YES	0.678	0.533	2.850	0.265

Input Normalization: In order to test the usefulness of normalization of input variables, model M1 for Sunnyside beach was developed without applying normalization *(mapminmax)* function as shown in Table 6-1 and the same approach was applied to other two beaches as well. It was observed that model M1 had high RMSE and low CORR value compared to other ANN models which used normalized input variables. Based on these results it was decided that normalization of input variables would be the first step during ANN model development as all inputs have different units. As discussed in Section 5.1 this approach was accepted by (Heydari et al., 2013, Thoe et al., 2012, Shamisi et al., 2011) during their studies for beach water quality.

Training Algorithm: As mentioned before in Chapter-3, *tarinlm* is most used training algorithm for all three beaches but in order to verify and to understand the difference in terms of performance, some other training algorithm were tried keeping other criteria's same. From the results presented in Table 6-1, Table 6-2 and Table 6-3 it is evident that *tarinlm* has shown better results than *tarinrp* and *traingd* for all three beaches in terms of RMSE and CORR. This training algorithm is generally fastest training function among others and is the default training function for feedforward networks (Beale et al., 2013), for that reason *trainlm* was used for the research work as it performs better on function fitting (nonlinear regression) problems too.

Transfer Function: The default transfer function of Neutral Network Toolbox for Levenberg-Marquardt algorithm (*trainlm)* from input to hidden layers is Hyperbolic Tangent (*tansig)* and from hidden to output layer is Linear (*purelin),* even though *tansig* is the fastest training function compared to the other two (Beale et al., 2013). To understand the influence of each training functions, different training functions were employed during model development form input to hidden layer. The performance of those models (Table 6-1, Table 6-2 and Table 6-3) indicated that *tansig* function produced better results for E. coli prediction. Therefore, *tansig* was applied from input to hidden layer and *purelin* was applied from hidden to output layers for all models. Thoe et al. (2012) also applied them same approach during their work and found the same results in terms of training function selection.

Numbers of Neurons: For each models, performance in terms of the RMSE and CORR of the training and simulation data sets was calculated to determine the appropriate number of hidden neurons to provide adequate generalization while avoiding overfitting. ANN models using 5, 10, 20, 30, 50 and 100 hidden neurons were assessed using trial and error method during model development. Out of which models with 5 and 10 hidden neurons revealed poor results and hence were not investigated further. Table 6-1, Table 6-2 and Table 6-3 demonstrate that increasing the number of hidden neurons more than 20, decreased model error during the training as measured by RMSE. However, considerably higher values of RMSE were observed when ANN models with more than 50 hidden neurons were applied to the simulation data, indicating overfitting. It was concluded that the best performance was demonstrated in models using 30 hidden neurons for Sunnyside beach, 50 hidden neurons for Rouge beach and 50 hidden neurons for Marie Curtis Park East beach.

6.3 Quantitative Assessment

Out of all models generated, comparatively well performing models were selected for Sunnyside beach, Rouge beach and Marie Curtis Park East Beach as shown in Table 6-4, Table 6-5 and Table 6-6 respectively. The same tables shows the parameters i.e. transfer functions, training function and number of hidden neurons used for generating those models. For the quantitative assessment purpose, all models RMSE and CORR for training and simulation data have been shown in the same tables. For the generation and assessment of confusion matrix, used to visualize the performance of individual models in their ability to correctly predict conditions requiring beach posting, 100 E. coli count / 100 ml of beach water are used as the threshold concentrations.

Sunnyside Beach

After running all 32 models shown in Table A-1, Table 6-4 displays the best performing models based on quantitative assessment for the Sunnyside beach. Based on the performance matrix values shown in Table 6-4 it is clear that models M7B, M9 and M15 outperform others due to their most favourable input combinations. The input parameters for models M9 and M15 were common except that M15 had an additional input data of lake level. Performance parameter values for M15 for the training set were such that CORR value was 0.838 (the highest value) and RMSE was 0.536 which is significantly lower compared to other input combinations. When simulation data set was considered, CORR was 0.276, with RMSE value relatively higher than for other models. As mentioned earlier this lower values of CORR could have been happened because of smaller sample size available for simulation data.

Table 6-4 Best Performing ANN models for Sunnyside beach

			Training data					Simulation data				
No	**Model**	**Input combination**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**
1	M4	st.fl(3),w.ht,w.spd,slr, past 48hrs rain(3)	78%	17%	5%	0.982	0.672	71%	16%	13%	0.920	0.204
2	M5F	Humber st.fl, w.spd, slr,Hy041 past 48hrs rain, pr. lnEC	74%	15%	11%	1.012	0.598	70%	14%	16%	0.943	0.196
3	M7B	st.fl(3),w.spd,slr,past 48hrs rain (HY041,T.W.2), Pr. lnEC, l.l.	79%	15%	6%	1.165	0.477	74%	10%	16%	0.108	0.151
4	M9	st.fl(3),w.spd,slr, HY041 past 48hrs rain,pr. lnEC	81%	13%	6%	0.677	0.777	71%	26%	3%	0.939	0.442
5	M15	st.fl(3),w.spd,slr, HY041 past 48hrs rain, pr. lnEC, l.l.	82.5%	13.5%	4%	0.536	0.838	65%	22%	13%	1.374	0.276

For all models- Transfer function (tansig- purelin), Training algorithm (trainlm), No of hidden neurons- 30

*past 48hrs rain(3)**-past 48hrs rain of station HY041,T.W.2 and T.W.11, st.fl(3)*-streamflow of Humber river,Mimico creek and black creek, pr. lnEC -previous day ln E. coli count, w.spd-wind speed, w.ht-wave height , l.l.-lake level, slr-solar radiation*

Lower values of RMSE and values of CORR closer to 1 indicate better agreement between simulated results and observed data. The other models M4 and M5F have comparatively low CORR and high RMSE value. This may be due to lack of essential input parameters that are necessary to capture the underlying pattern between water quality parameters and E. coli concentrations or the existence of different driving forces and relationships between variables. The results of model M5F signifies that considering Humber River streamflow only out of all streamflow is not enough and it is essential to consider all three river streamflows to obtain optimized results.

Rouge Beach

After running all 36 models shown in Table A-2, Table 6-5 shows the best performing models based on quantitative assessment for Rouge beach. The same enlists the results for the

model performance criteria described above, as well as the percent of correct classifications and the number of false positives and false negatives for the E. coli concentrations prediction.

Table 6-5 Best Performing ANN models for Rouge beach

			Training data					Simulation data				
No	**Model**	**Input combination**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**
1	R4C	st.fl,w.dir,w.ht	75%	21%	4%	0.867	0.576	71%	29%	0%	1.43	0.151
2	R6	st.fl,w.dir,w.ht,w.spd,slr, past 48hrs rain(HY044,HY070), pr. lnEC	79%	15%	6%	0.854	0.683	65%	29%	6%	1.89	0.447
3	R7A	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044,HY070), pr. lnEC	79%	15%	6%	0.922	0.661	65%	29%	6%	1.26	0.279
4	R9	st.fl,slr,past 48hrs rain(HY044,HY070), pr. lnEC	76%	17%	7%	0.873	0.570	68%	19%	13%	1.95	0.6
5	R11	st.fl,past 48hrs rain(HY044,HY070), pr. lnEC	75%	20%	5%	0.89	0.547	71%	16%	13%	1.98	0.654

For all models- Transfer function (tansig- purelin), Training algorithm (trainlm), No of hidden neurons- 50

st.fl-streamflow of Rouge River, pr.lnEC-previous day ln E. coli count, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, l.l.-lake level, slr-solar radiation

From the Table 6-5 it was found that input combinations of models R6, R7A and R9 performed the best for the training and simulation data sets. For models R6 and R7A the input parameters were common, except that the R6 model has an additional input data of wind speed. Performance parameter values for R6 for the training set was such that CORR was 0.683 (the highest value) and RMSE was 0.854, which was lowest compared to other models with other input combinations. When simulation data set was considered, CORR was 0.447, with relatively high RMSE value of 1.89 for ANN model R6, whereas the model R9 exhibits higher CORR value of 0.600. The other models R4C and R11 have comparatively low CORR and high RMSE

value. It could be concluded from the results comparison of model R4C and R6 that solar radiation, past 48hrs rain and previous day E. coli concentrations have more effect on the E. coli concentrations in the water at the Rouge beach.

Marie Curtis Park East Beach

After running all 33 models shown in Table A-3, below are the best performing models for Marie Curtis Park East beach. Table 6-6 show the results for the model performance criteria described above, as well as the percentage of correct classifications and the number of false positives and false negatives for the E. coli prediction.

Table 6-6 Best Performing ANN models for Marie Curtis Park East Beach

			Training data					**Simulation data**				
No	**Model**	**Input combination**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**	**Correct classification**	**False -**	**False +**	**RMSE**	**CORR**
1	M7	st.fl,w.ht,past 48hrs rain (HY025,HY033), pr. lnEC	72%	19%	9%	0.780	0.581	68%	10%	23%	3.060	0.532
2	M12A	st.fl,w.ht,w.dir,slr, past 48hrs rain(HY025,HY033), pr. lnEC,l.l.	74%	18%	8%	0.812	0.666	71%	16%	13%	1.440	0.385
3	M13A	st.fl,slr,past 48hrs rain (HY025,HY033), pr. lnEC, l.l.	75%	11%	13%	0.860	0.648	68%	19%	13%	1.507	0.431

For all models- Transfer function (tansig- purelin), Training algorithm (trainlm), No of hidden neurons- 50

st.fl-streamflow of Etobicoke creek, pr. lnEC-previous day ln E. coli count, w.dir-wind direction, w.ht-wave height, l.l.-lake level, slr-solar radiation

Examination of Table 6-6 shows that input combinations of models M12A and M13A were better among other input combinations based on the performance matrix values for the training and simulation data sets. The performance parameters of all three models are good in terms of RMSE and CORR but M13A shows better overall results when comparing training and

simulation sets together. It is apparent that solar radiation, lake level, wave height and wind direction are the good explanatory variable for ANN modelling for the E. coli concentrations in water at the Marie Curtis Park East beach.

6.4 Qualitative Assessment

In parallel with quantitative assessment, the actual performance of the model to predict the exceedance of water quality threshold (e.g. 100 E. coli count / 100 ml of beach water) was also assessed with the help of qualitative assessment. As discussed previously, using performance matrix all the results were divided into percentage of correct classification as well as the number of false negatives and false positives. Detailed discussion of qualitative assessment for all three beaches has described as under.

Sunnyside Beach

Based on the details presented in Table 6-4, compared to other models both model M7B, M9 and M15 yield much better ability to correctly predict E. coli concentrations. This means that the models are useful in predicting the water quality threshold exceedance and issue correct beach advisory notes to public under poor water quality. At the same time for models M9 and M15, false negative values are slightly higher than false positive values, indicating the problem of 'false alarm' (issuing incorrect beach advisory notes when beach is actually clean) can also be alleviated. This implies that the models have sufficient precision to be used for operational nowcasting of E. coli concentrations at Sunnyside beach. Models M7B, M9 and M15 are able to correctly predict 79%, 81% and 82.5% instances of concentrations exceeding the beach water quality standard, respectively. These numbers are promising.

Rouge Beach

Based on the information presented in Table 6-5, R6 and R7A models are able to correctly predict at least 78% instances of concentration exceeding beach water quality standards i.e. 100 E. coli count / 100 ml of beach water, whereas model R9 achieves 68% correct classification and nearly equal percentage of false negatives and false positives. As noticed in the case of Sunnyside beach, models for this beach also exhibited slightly higher value of false negatives. Out of all models, R6, R7A and R9 models have sufficient precision to be used for operational nowcasting of E. coli concentrations at Rouge beach. Although the difference of performance being minor the higher correct classification for models R6, R7A and R9 show that the ANN models are better performing when solar radiation is considered as one of the input variables.

Marie Curtis Park East Beach

Table 6-6 shows the qualitative assessment of best performing models out of 33 models developed for Marie Curtis Park East beach. Based on the information presented, the ANN models correctly predict 72% to 75% of the occurrences of concentrations greater than 100 E. coli count / 100 ml of beach water, depending on the input parameters used. For all models the number of false positives and false negatives are nearly equal and less than 20% of the number of observed values. Overall improvement was observed when solar radiation and lake level characteristics are added to the input parameters. Out of all models, M12A and M13A are better performing models, able to correctly classify at least 74% occurrences in the training phase and minimum 68% in the simulation phase.

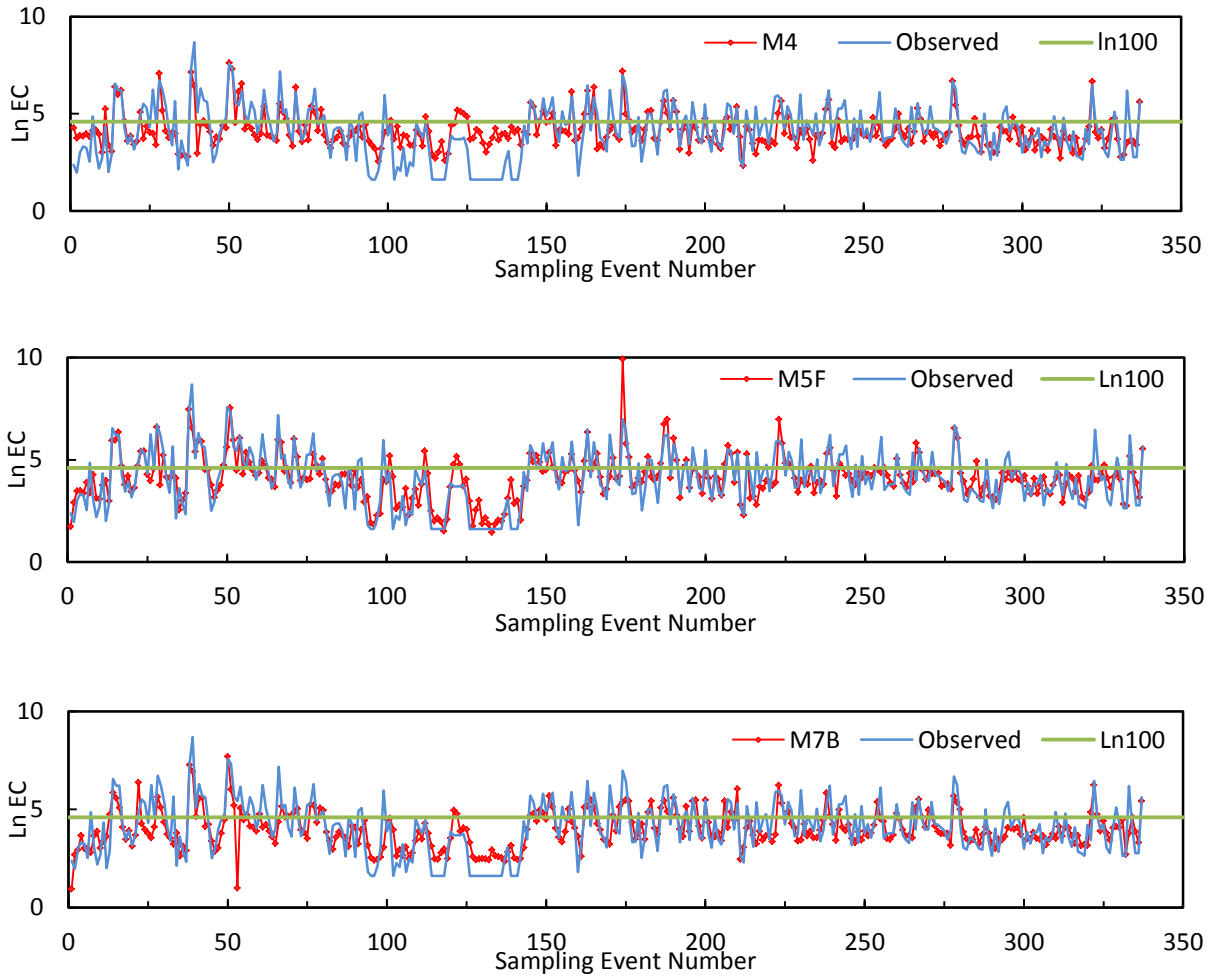
6.5 Visual Assessment

Examination of the performance statistics for different ANN models for each beach would suggest similar model prediction capability. However, examination of graphical representation illustrate why graphical evaluation of model results is important to understand model behavior and assessing suitability of the model for a particular application along with qualitative and quantitative assessment. These figures provide time series plot and a point-by-point comparison of observed (x-axis) and simulated (y-axis) E. coli concentrations for selected ANN model scenarios. In addition to the calculation of statistics, plotting the time series of observed and predicted values can be advantageous and help to diagnose strengths or weaknesses of a model.

Sunnyside Beach

Figure 6-1 represent the time-series plots of observed and predicted E. coli concentrations for M4, M5F, M7B, M9 and M15 models during training periods. Figure 6-2 and Figure 6-3 shows the time series plot and scatter plot of all five ANN models for the simulation period. All time series plot of Figure 6-1 and Figure 6-2 show sampling event number on X-axis and corresponding lnEC value on Y-axis. These event numbers corresponds to consecutive days of year 2008 to 2012 from June-August which is consistent with all three beaches. In Figure 6-3 the scatter plots are divided into quadrants differentiating true negatives, false positives, true positives and false negatives. The visual inspection of Figure 6-1, Figure 6-2 and Figure 6-3 reveals how ANN models were successful in predicting E. coli concentrations in the period representing the time frame for the training and simulation data. Figure 6-1 shows that the developed ANN models predicts well for the training dataset with model M7B, M9 and M15

being slightly higher. Numbers of outliers are less in the case of Models M9 and M15. Figure 6-2 and Figure 6-3 reveals that during simulation models M4 and M5F fail to have the same numbers of correctly classified E. coli counts as models M7B, M9 and M15. Apart from that, generally Model M4 and M5F over predicted E. coli concentrations for values less than 100 E. coli count / 100 ml of beach water. In contrast, Models M7B, M9 and M15 are very well able to predict safe state as well as exceedance of water quality standards. These results are consistent with their lower value of RMSE and higher CORR values (Table 6-4). Sometimes, depending on model application, a tendency to over or under predict concentrations in a particular range may be allowable or even preferable.



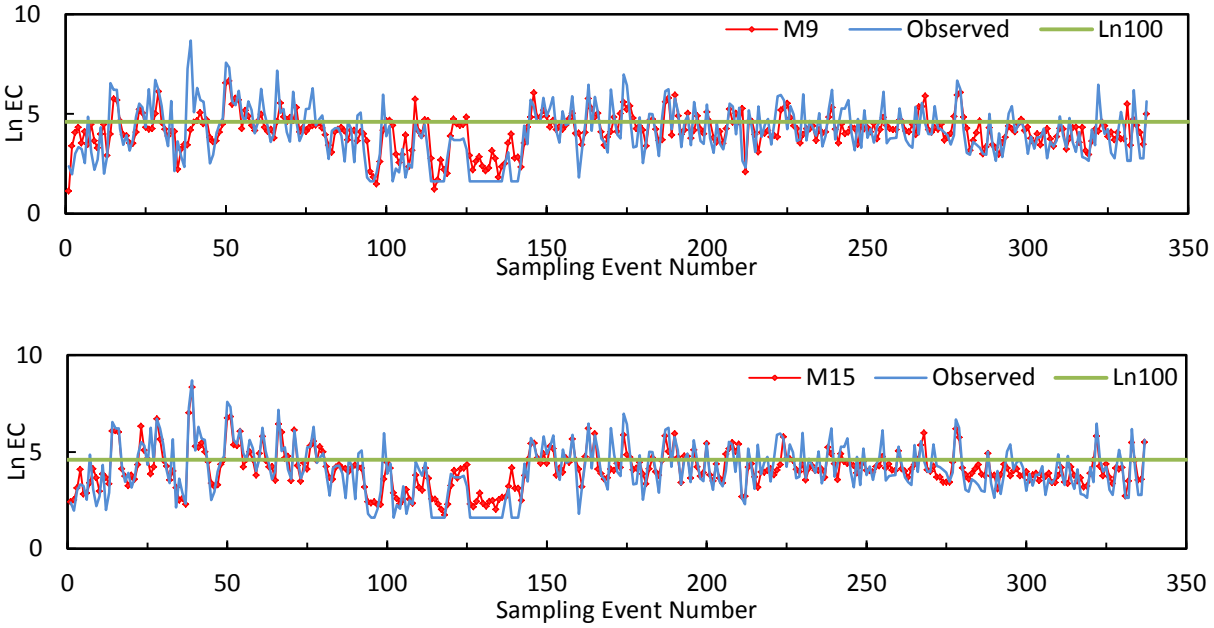
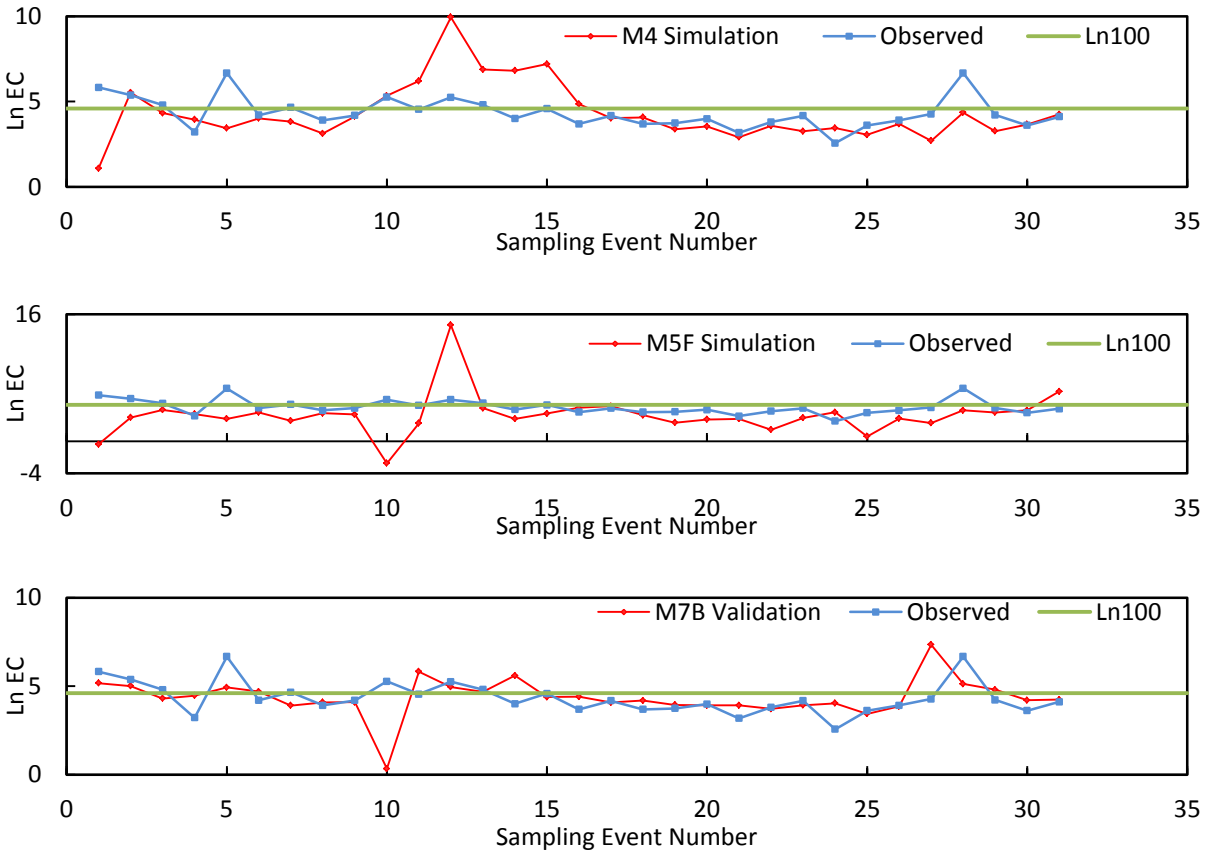


Figure 6-1 Comparison of predicted and observed lnEC concentration at Sunnyside beach for the testing period



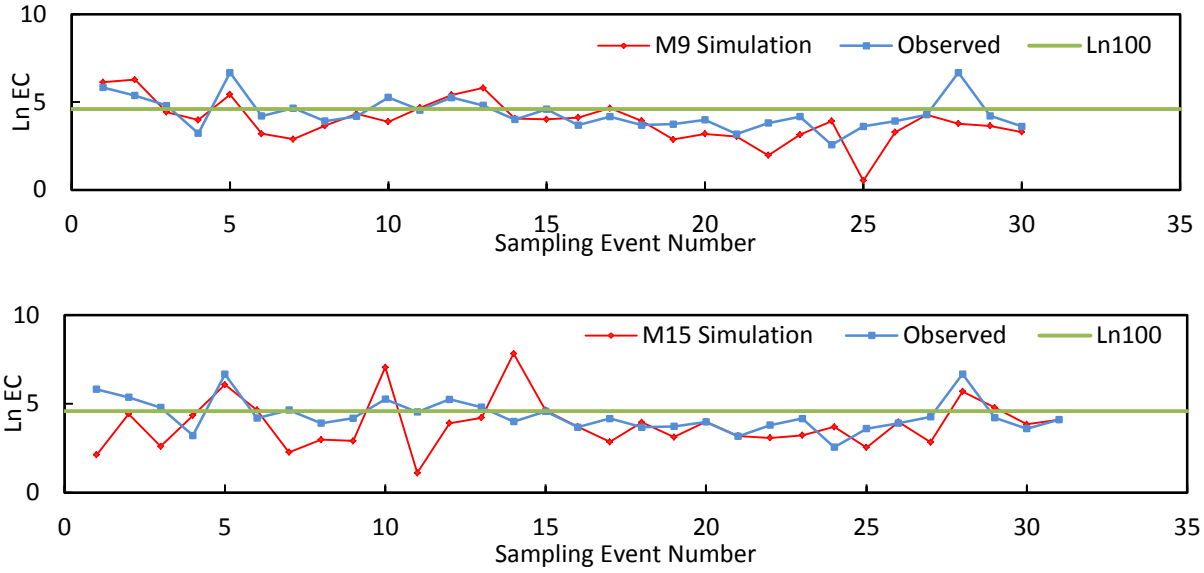
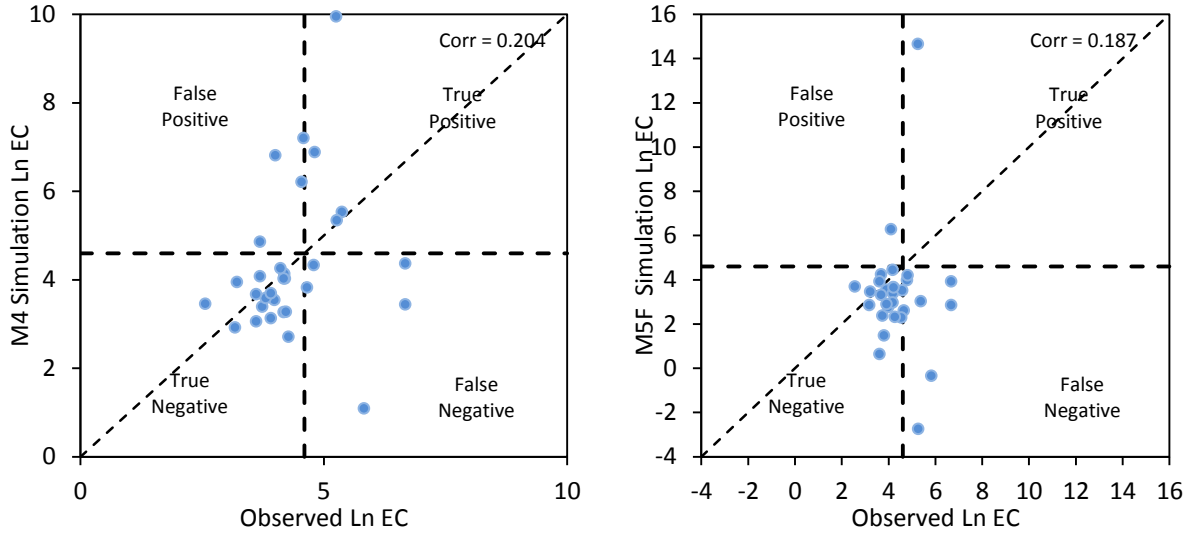


Figure 6-2 Time series plot of predicted and observed lnEC concentration for the simulation period, Aug-2012



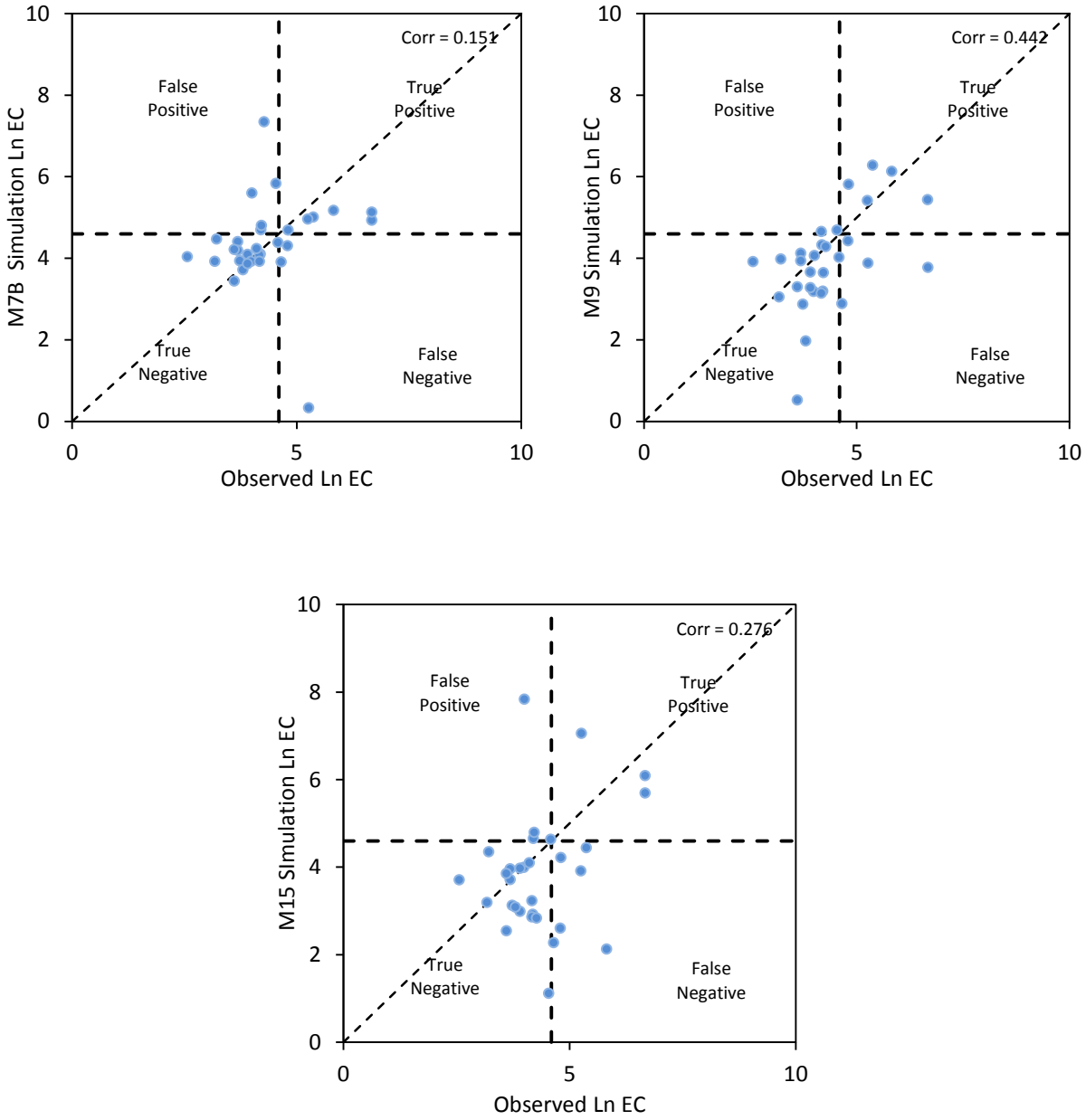
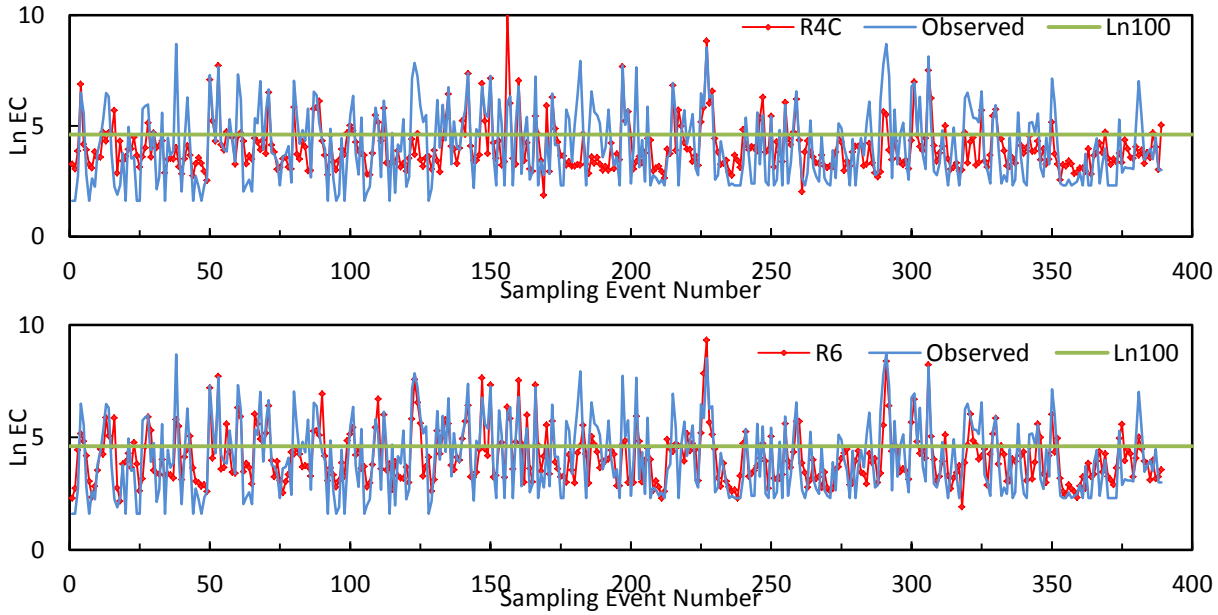


Figure 6-3 Scatter plot of predicted and observed lnEC concentration for the simulation period, Aug-2012

Rouge Beach

Figure 6-4 represent the time-series plots of observed and predicted E. coli concentrations for R4C, R6, R7A, R9 and R11 models during training periods. Figure 6-5 and Figure 6-6 shows the time series plot and linear regression plot of these four ANN models for simulation period. Figure 6-4 shows that the models R4C and R11 have a number of occurrences where the predicted values are extremely high or low compared to observed values, whereas models R6, R7A and R9 are able to closely predict the observed values. As seen in Figure 6-5 and Figure 6-6. Model R6, R7A and R9 are very well able to predict safe state as well as exceedance of water quality standards i.e. 100 E. coli count / 100 ml of beach water. As it can be seen from the same Figures, Models R6, R7A and R9 perform well during simulation with model R9 having a slightly higher value of correct classification.



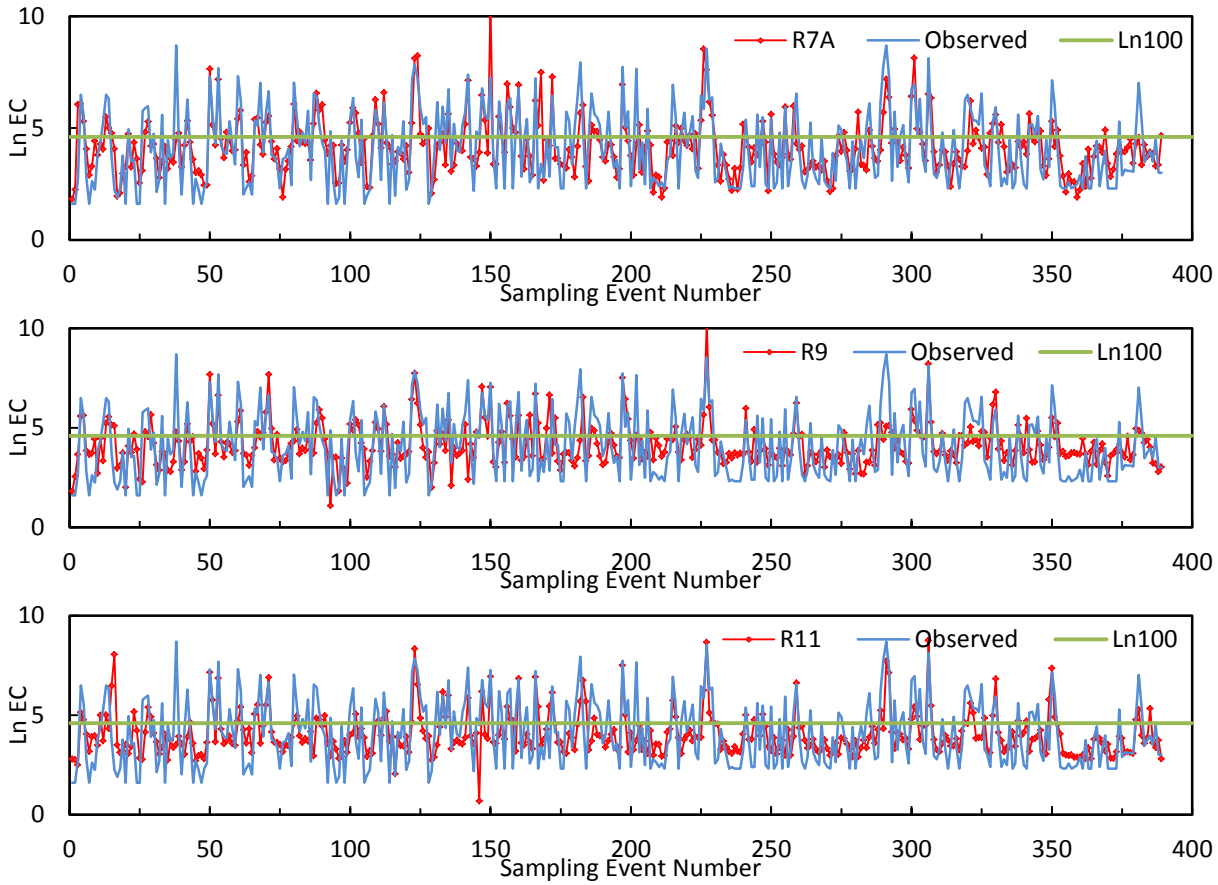
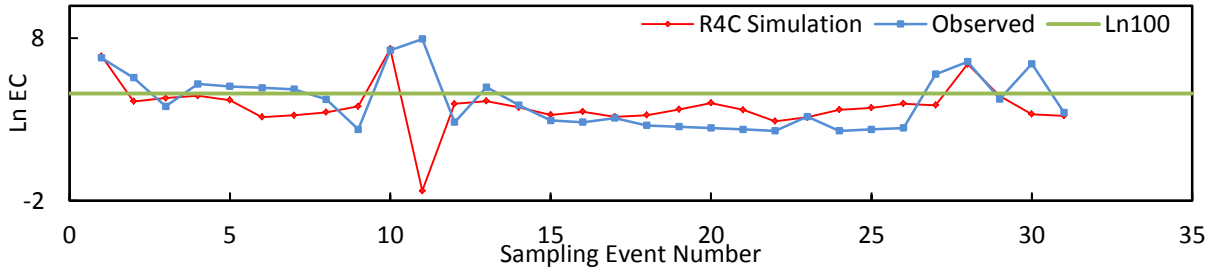


Figure 6-4 Comparison of predicted and observed lnEC concentration for Rouge beach's ANN models for the testing period, 2008-2012(jun-july-aug)



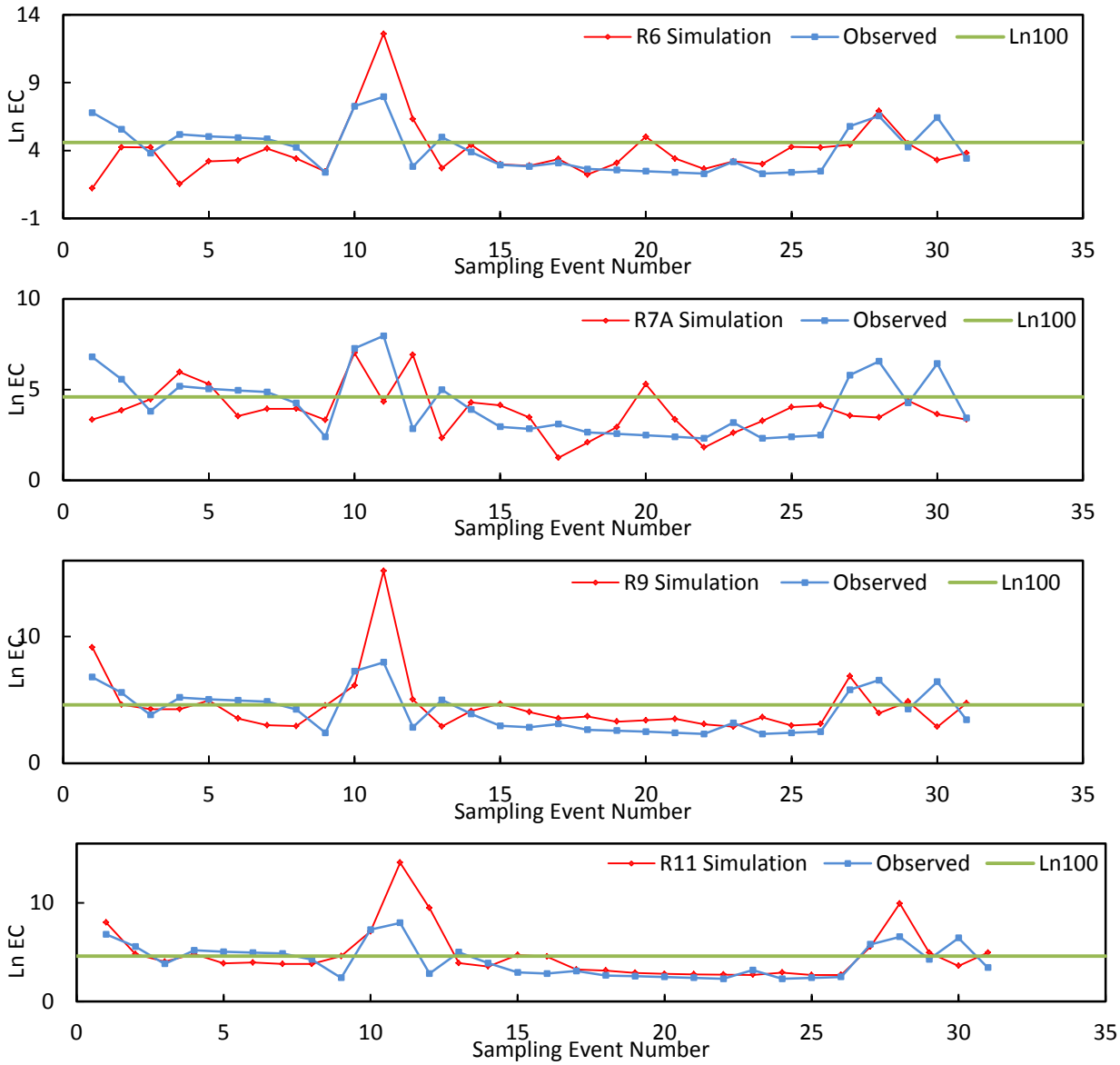
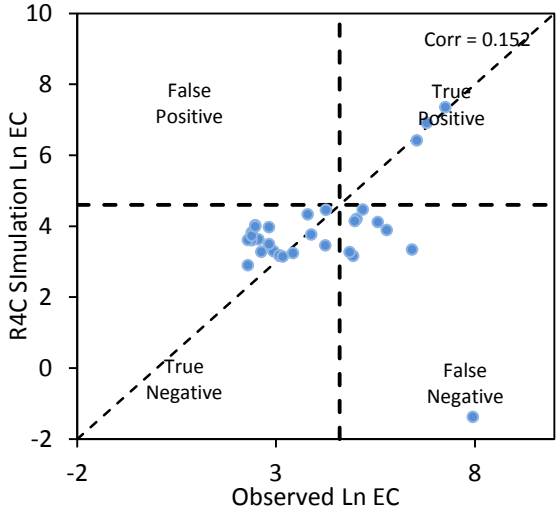
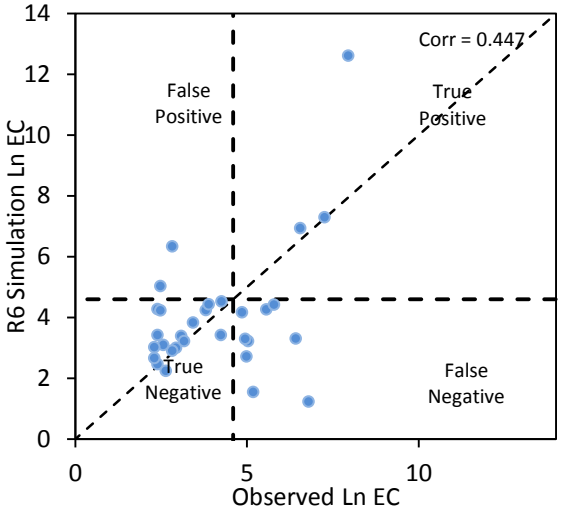
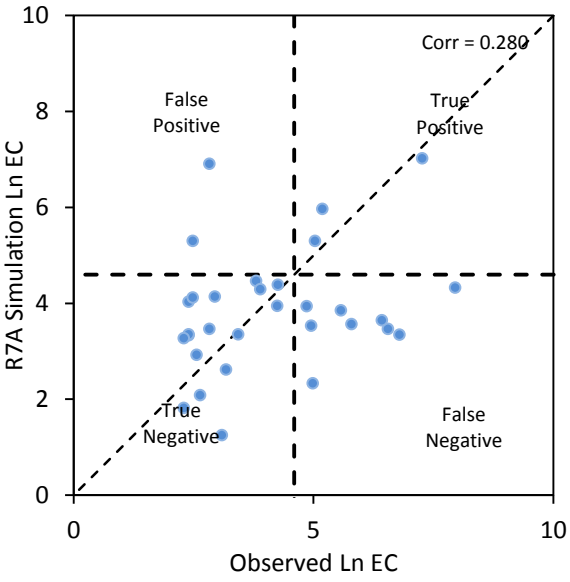
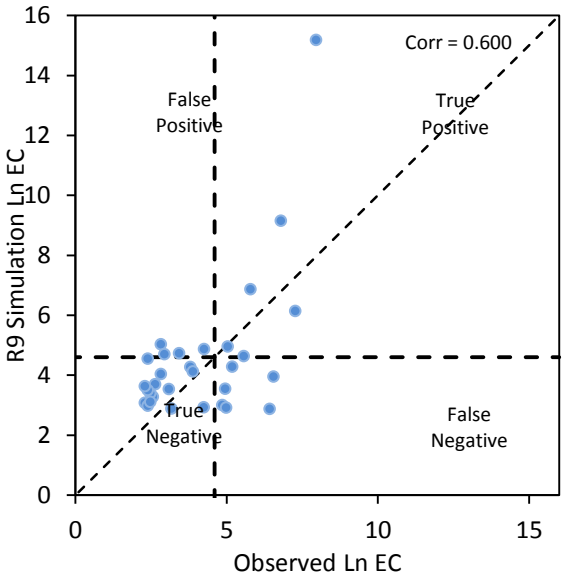


Figure 6-5 Time series plot of predicted and observed lnEC concentration for Rouge beach for the simulation period, Aug-2012


Corr = 0.152
False Positive
True Positive
True Negative
False Negative
R4C SImulation Ln EC
Observed Ln EC
10
8
6
4
2
0
-2
-2
3
8


Corr = 0.447
False Positive
True Positive
True Negative
False Negative
R6 Simulation Ln EC
Observed Ln EC
14
12
10
8
6
4
2
0
0
5
10


Corr = 0.280
False Positive
True Positive
True Negative
False Negative
R7A Simulation Ln EC
Observed Ln EC
10
8
6
4
2
0
0
5
10


Corr = 0.600
False Positive
True Positive
True Negative
False Negative
R9 Simulation Ln EC
Observed Ln EC
16
14
12
10
8
6
4
2
0
0
5
10
15

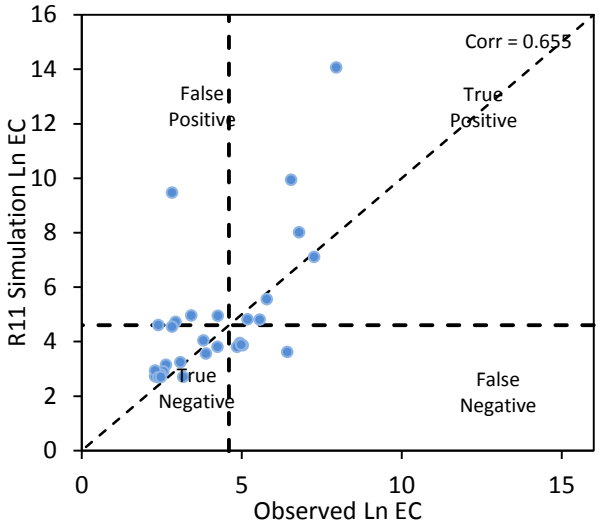


Figure 6-6 Scatter plot of predicted and observed lnEC concentration for Rouge beach for the simulation period, Aug-2012

Marie Curtis Park East Beach

Figure 6-7 represent the time-series plots of observed and predicted E. coli concentrations for M7, M12A and M13A models during training periods. Figure 6-8 and Figure 6-9 shows the time series plot and linear regression plot of these ANN models for simulation period. Figure 6-7 shows that the ANN developed for all models predicts well for the training dataset with models M12A and M13A being slightly higher. Numbers of outliers are lower in the case of Models M12A and M13A. Figure 6-8 and Figure 6-9 reveals that model M7 fails to have similar numbers of correctly classified counts of E. coli as models M12A and M13A for given water quality standards i.e. 100 E. coli count / 100 ml of beach water. Even though M12A and M13A being superior out of all models they are apparently not the most optimized model for the particular beach. This is may be due to a variety of reasons including lack of essential input parameters that are necessary to capture the underlying pattern between water quality and

indicator organism concentrations. As the goal of this research is to develop a model with the data available in real-time not all possible explanatory variable could have been incorporated.

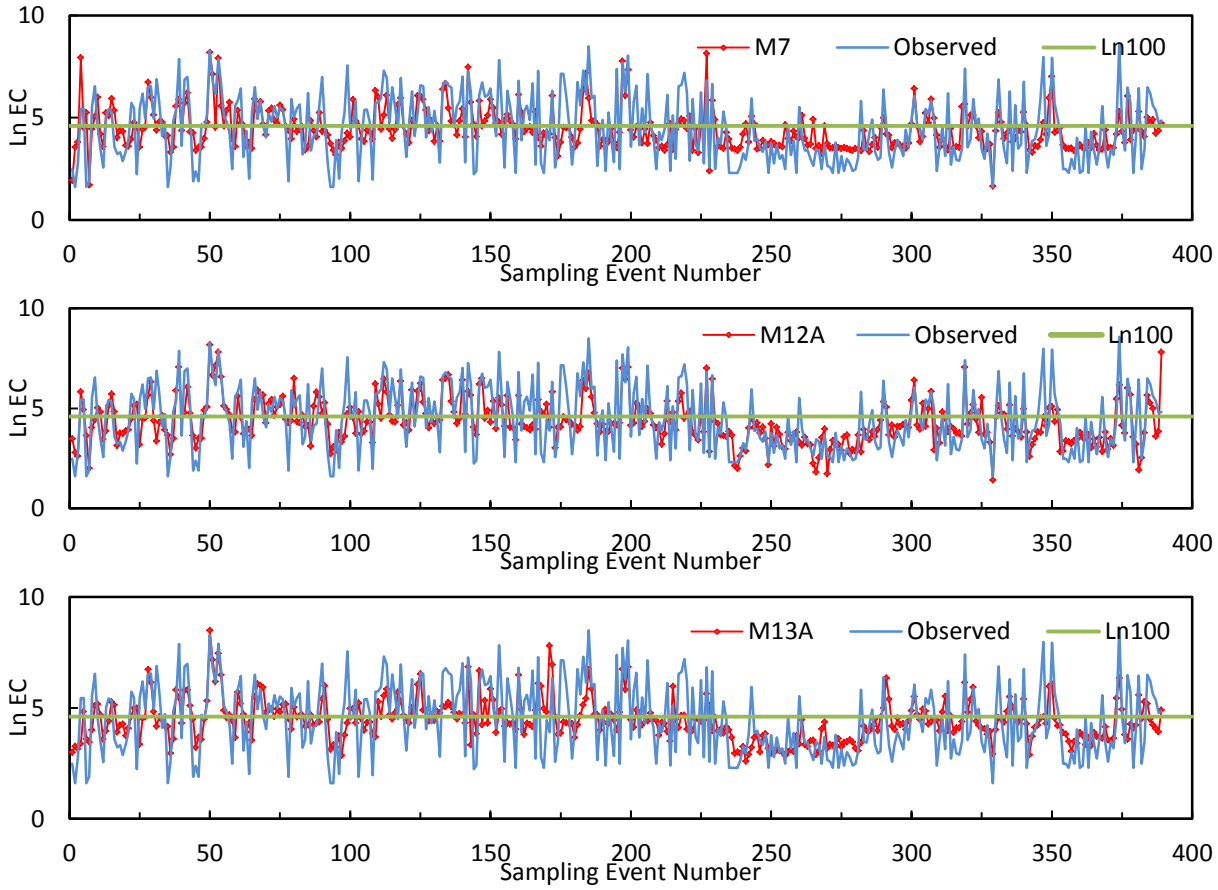
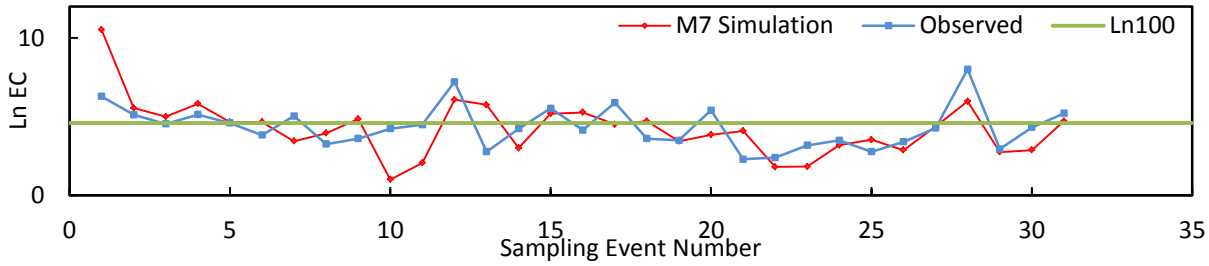


Figure 6-7 Comparison of predicted and observed lnEC concentration at Marie Curtis Park East beach's ANN models for the testing period, 2008-2012(jun-july-aug)



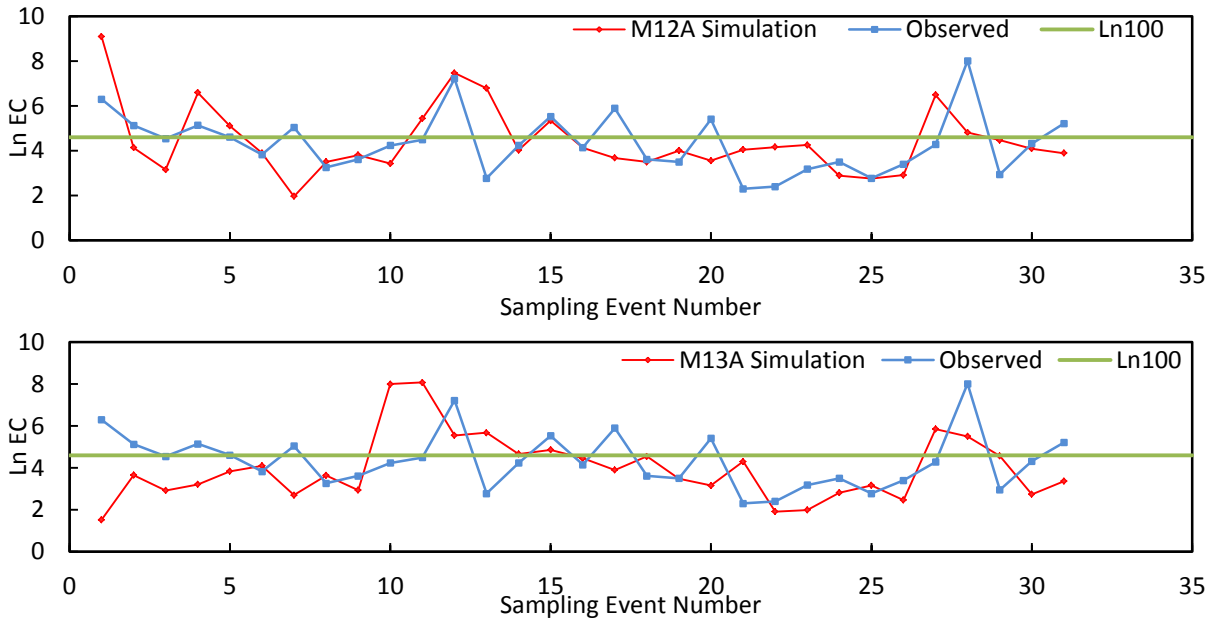
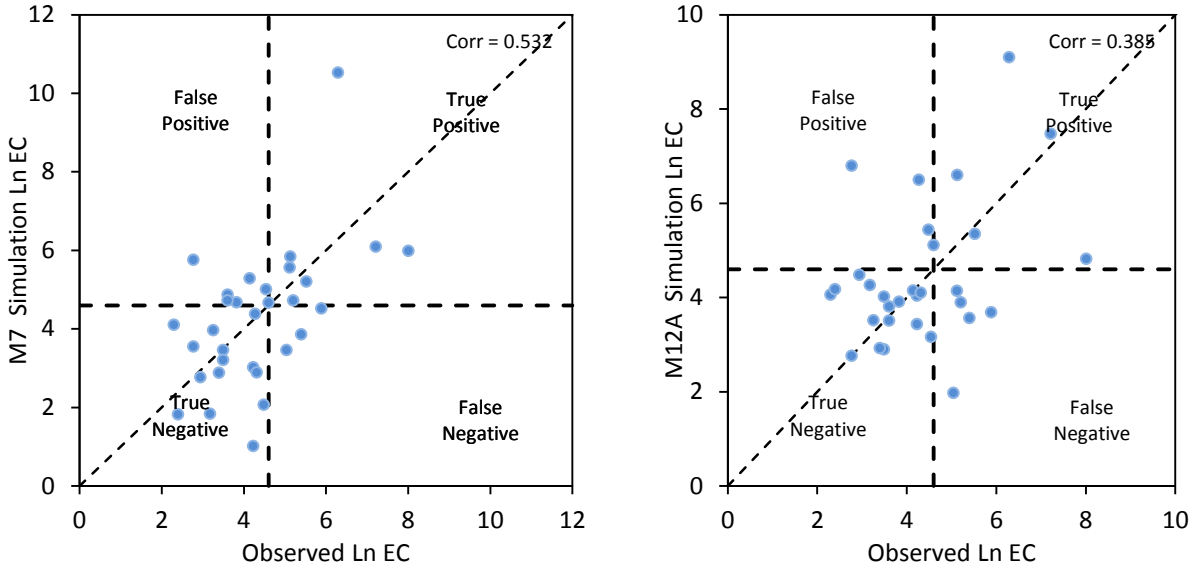


Figure 6-8 Time series plot of predicted and observed lnEC concentration for Marie Curtis Park East Beach for the training period, Aug-2012



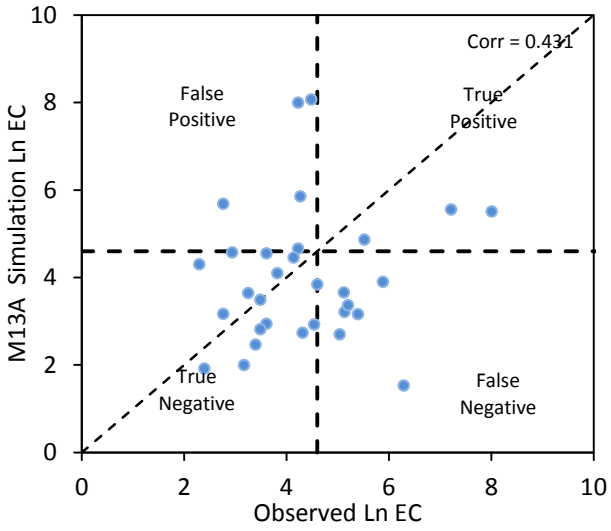


Figure 6-9 Scatter plot of predicted and observed lnEC concentration for Marie Curtis Park East Beach for the training period, Aug-2012

From the above discussion, where a comparison of models with similar performance metrics resulted in different performance in terms of under or over- prediction relative to a water quality standard, illustrates the need for a combination of relative and absolute statistics as well as graphical assessment of model output to evaluate their performance. Had CORR and RMSE been the only statistics used to evaluate the models, the differences in performance between ANN models would be essentially undetectable without graphical examination of the results. The same has helped understand model behavior for the given threshold value of beach water quality standard and assessing suitability of the model for a particular application along with qualitative and quantitative assessment.

6.6 Model Performance Evaluation

ANN model evaluation was performed using a simulation data set by computing the number of correct classification, false negatives and false positives and compared with the persistence models.

Sunnyside Beach

Table 6-7 and Figure 6-10 shows the comparison between best performing ANN models and currently used persistence model i.e. prior day E. coli, during simulation period (August 2012) for Sunnyside Beach. The scatter plot shows that the best performing ANN models predict decreased false positive values and increased the correct classification nearly by 15%, making it more protective of public health than the use of prior day E. coli i.e. the persistence model. However, in certain instances the ANN models leads to increase in false negatives.

Table 6-7 ANN Model performance statistics for Sunnyside beach with the persistence model during Simulation Period (August 2012)

Model	Correct Classification	False -	False +
M4	71%	16%	13%
M5F	70%	14%	16%
M7B	74%	10%	16%
M9	71%	26%	3%
M15	65%	22%	13%
Persistence	65%	19%	16%

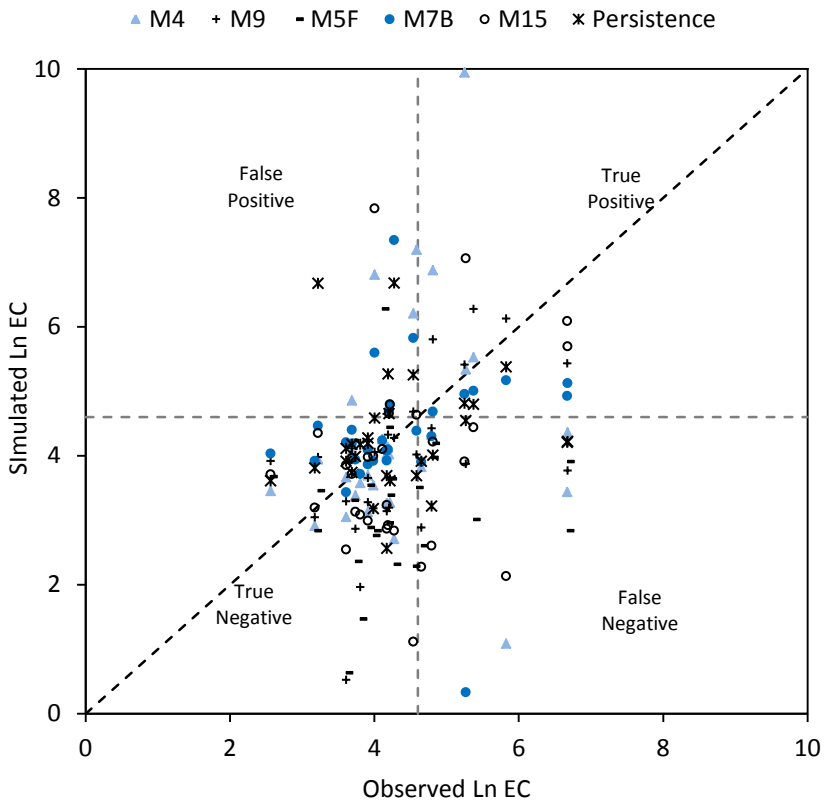


Figure 6-10 Performance evaluation for best performing ANN models and the persistence model for Sunnyside Beach during Simulation Period (August 2012)

As seen in Table 6-7 overall correct classification percentage is slightly higher for the ANN models and they balance the rates of true positives and true negatives.

Rouge Beach

Table 6-8 and Figure 6-11 shows the comparison between best performing ANN models and currently used persistence model i.e. prior day E. coli, during simulation period (August 2012) for Rouge Beach. The scatter plot shows that the best performing ANN models predict decreased false positive values and increased the correct classification nearly by 10%, making it more protective of public health than the use of prior day E. coli i.e. the persistence model. However, in certain instances the ANN models leads to increase in false negatives.

Table 6-8 ANN Model performance statistics for Rouge beach with the persistence model during Simulation Period (August 2012)

Model	Correct Classification	False -	False +
R4C	71%	29%	0%
R6	65%	29%	6%
R7A	65%	29%	6%
R9	68%	19%	13%
R11	71%	16%	13%
Persistence	64%	19%	17%

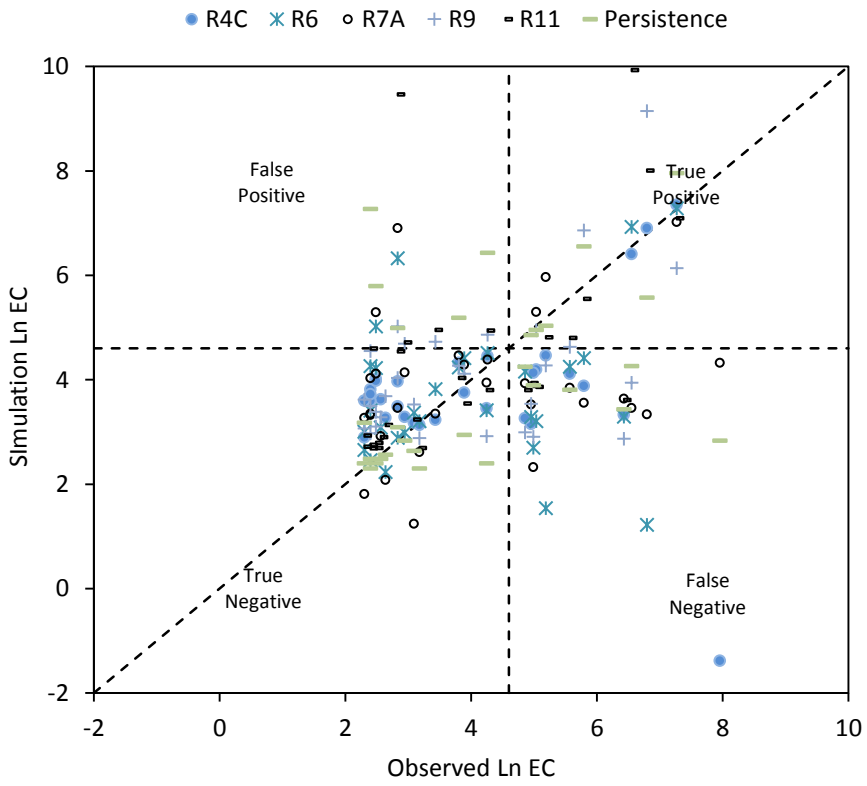


Figure 6-11 Performance evaluation for best performing ANN models and the persistence model for Rouge Beach during Simulation Period (August 2012)

As stated earlier model R9 was considered to be best performing model out of all ANN models for Rouge beach. As seen in Table 6-8, overall correct classification percentage is slightly higher

for R9 Model and it balances the rates of true positives and true negatives. Even though, R4C, R6 and R7A are having higher percentage of correct classification, significantly lower amount of false positives compare to false negatives higher the risks of public health.

Marie Curtis Park East Beach

Table 6-9 and Figure 6-12 shows the comparison between best performing ANN models and currently used persistence model during simulation period (August 2012) for Marie Curtis Park East Beach. The scatter plot shows that the best performing ANN models predict decreased false positive values and increased the correct classification nearly by 65%, making it significantly protective of public health than the use of prior day E. coli i.e. the persistence model. Unlike other two beaches, all the ANN models for Marie Curtis Park East Beaches exhibited lower number of false positives and false negatives.

Table 6-9 ANN Model performance statistics for Marie Curtis Park East beach with the persistence model during Simulation Period (August 2012)

Model	Correct Classification	False -	False +
M7	68%	10%	23%
M12A	71%	16%	13%
M13A	68%	19%	13%
Persistence	42%	29%	29%

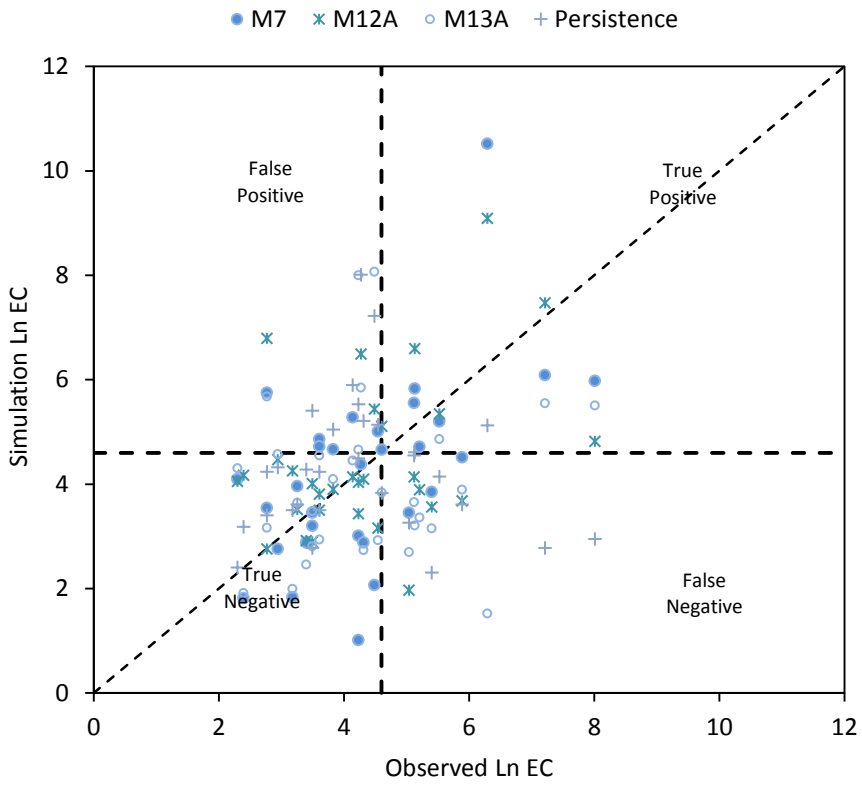


Figure 6-12 Performance evaluation for best performing ANN models and the persistence model for Marie Curtis Park East Beach during Simulation Period (August 2012)

It would be worth noting here that even though author thinks that there is a room for improvement in ANN models for particular beach by introducing some other explanatory variables, the models generated with current configuration significantly outperform than the persistence model.

6.7 Summary

Research conducted on the three out of eleven swimmable beaches of Toronto, because of the high concentrations of E. coli concentrations on these three beaches and as the City of Toronto requires an alternative nowcast model in the instances when sampling of water is not

possible due to extreme weather conditions. The use of multilayer feedforward ANNs to simulate concentrations of the E. coli was investigated. Correlation and graphical analysis was used to select explanatory input variables for all models. A total of 32 ANN models for Sunnyside beach, 36 models for Rouge beach and 33 models for Marie Curtis Park East beach were considered using readily available real time data for prediction of E. coli concentrations. Model goodness of fit was evaluated using quantitative and qualitative performance assessment and visual assessment of model results. In addition, model ability to correctly classify E. coli concentrations relative to Toronto based beach water quality standards was assessed.

Current use of persistent model for beach water quality monitoring showed poor performance and does not provide the precision necessary for the operational nowcasting of E. coli concentrations. While not being able to successfully predict precise concentrations, the ANN models demonstrated the ability to predict up to 75%-83% of the occurrences when E. coli concentrations would exceed beach water quality thresholds. This is a significant finding from a public safety standpoint, as it is often more critical to determine when E. coli concentrations exceed a standard than to know their exact E. coli concentration. Typically, the models performed best at correctly categorizing exceedance of the 100 E. coli count / 100 ml of beach water quality standards. Generally, higher numbers of false negatives compared to false positives for the E. coli models indicates that the problem of 'false alarm' can be alleviated but also indicates that the tendency of the models to underestimate values at high observed concentrations at times.

In addition, the chapter yielded the following findings:

- It is evident that past 48hrs rain of station HY041 and T.W.2, streamflow of Humber River, Mimico Creek and Black Creek, previous day E. coli concentrations, wind speed, solar radiation and lake level are the governing key factors for E. coli concentrations at the Sunnyside beach.
- For the Rouge Beach, wind direction, wave height, wind speed, solar radiation, streamflow of Rouge River, past 48hrs rain of station HY044 and HY070 and previous day E. coli counts are the variables that affected the E. coli concentrations the most.
- The streamflow of Etobicoke Creek, previous day E. coli counts, wind direction, wave height, lake level, past 48hrs rain of station HY025 and HY033 and solar radiation are the good explanatory variable for nowcasting of E. coli concentrations at the Marie Curtis Park East beach.
- For model performance assessment at least three methods should be used to find out the best models among the all developed models. In this research RMSE and CORR were used for the quantitative assessment, confusion matrix consisting of the number of correct classification, false negative and false positive was used for the qualitative assessment and the time-series and regression plots were used as visual assessment to find the best ANN models.
- In every instance ANN models are better to predict the changes in E. coli concentrations when compared with the persistence model.
- There could be further improvement possible if some other explanatory variables are added. Due to unavailability of the data (e.g. Turbidity) in real-time some parameters

were not considered as an input even though they appear to be highly correlated with E. coli concentrations.

7 CONCLUSIONS AND RECOMENDATIONS

7.1 Summary and Conclusions

This research work has studied the usage of artificial neural networks as an alternative method for E. coli concentrations prediction. The existing persistence model for the beach posting based on previous day's E. coli concentrations and MLR model based on E. coli concentrations with related parameters were studied with their advantages and disadvantages.

All ANN models were developed keeping in mind that all the input data should be available in real-time, as the intended application of the methodology is to generate a nowcast of E. coli concentration prediction in the beach water using real-time data provided by different government authorities. E. coli results were compared to the actual historic data provided by the Toronto Public Health. The benefit of using real-time data is that the operator of this model does not have to wait until some test results are available, which could be an input parameter for the model for the following day prediction.

Models investigated in this thesis belong to the feedforward back-propagation neural network architecture. A representative ANN-based nowcast model for E. coli concentration prediction for recreational beaches was developed. The simulation results obtained prove that satisfactory performance has been achieved for all three beaches by the best performing ANN models for the respective beaches. As further illustrated, depending on the application of the neural network and the size of the training data set, size of the ANN (the number of hidden layers and number of neurons per hidden layer) keeps varying. The importance of choosing the most appropriate ANN configuration, in order to get the best performance from the network, has

been stressed upon the threshold value (i.e. 100 E. coli count / 100 ml of beach water) in this work.

Some important conclusions that can be drawn from this thesis are:

- The persistence model used for E. coli concentrations predictions may not handle extreme weather conditions as sampling could not be possible. Whereas the results of this study demonstrate that it is possible to predict the E. coli concentrations at Toronto beaches using ANN models regardless of weather conditions. Probably, this is one of the biggest advantages of ANNs over traditional modelling techniques.
- The performance comparison for models using logarithmically transformed explanatory data except previous day E. coli did not yield good results. This finding adds to the research work of (Bowden et al., 2003, Thoe et al., 2012) and suggests that logarithmic transformation of explanatory variables may be of some limited use in terms of multilayer feedforward ANN model performance. Nevertheless natural log transformed of previous day E. coli proved to be an important explanatory variable.
- Considering rain gauge stations of the related watershed area for particular beaches rather considering only one rain gauge station of Toronto Pearson International Airport (YYZ) is better accounting for the distribution of rainfall during wet weather. As apparent from the results of ANN models considering this criteria has certainly improved the nowcasting ability of the same.
- R values obtained from ANN models in this research work are comparable with the values stated in review of literature performed to predict different parameters in Beaches. Total correct prediction on average 65-83% for training and 65-75% during simulation was obtained, compared to 60-75%.

- The best performing ANN models were evaluated for a simulation data set by computing the number of correct classification, false negatives and false positives and compared with currently used persistence model for all three study beaches and found that ANN models are performing better at least by 10-65% than the persistence model.
- ANN is found to be a viable, easy and economical alternative for E. coli concentrations predication in real time. It is very essential to investigate and analyze the advantages of a particular neural network structure and learning algorithm before choosing it for an application because there should be a trade-off between the training characteristics and the performance factors of any artificial neural network.
- The training algorithms, input normalization, transfer function and the number of hidden neurons are the main criteria to be decided prior to work with any type of ANNs modelling.
- The finally selected ANN model, saved as a MATLAB project, can be used as a predictive tool for nowcasting/forecasting E. coli concentrations in beach waters using the *sim* function in MATLAB on daily base.

Overall, the ANN model results for E. coli concentration predictions in beach waters showed higher percentage of correct classification with nearly equal amount of false negatives and false positives which are typically reflective for all model results. The success of ANNs in other applications and results presented here provides an example of ANNs capability as a tool for predicting concentrations of E. coli. As all best performing models are predicting E. coli concentration differently in the same instance, further selection out of best performing models might be required. This selection can be performed by employing these models along with

persistence model for prolong period of time. And based on the accuracy to better predict E. coli concentration for the particular beach, optimized ANN model can be chosen.

The results presented in this research work add to the limited research on the use of ANNs for predicting E. coli concentrations in beach water and show that ANNs were able to successfully identify when E. coli concentrations exceeded the beach water quality standard that relate to public health standards. In particular, the results of this work show promise for ANNs to be used to correctly predict when beach water quality standards for E. coli are exceeded, which is a model application with relevance to researchers and practitioners involved in watershed management and water supply and recreational water protection.

7.2 Recommendations

This section provides the important recommendations that would be useful in future for beach water predictive models in Toronto.

- The methodology should be applied on more recent data sets (2011 – 2013) to refine the models and their performance.
- Concentrations of E. coli show rapid temporal variation, so data such as rainfall amount, flow of river, lake level and turbidity collected with shorter lag time will generally yield better results.
- ANN model generation guideline provided in this study can be used to develop an artificial measurement instrument. For example, the ANN model can be integrated to the code of an in-situ measurement device to predict Turbidity based on other parameters.

- There are parameters that have shown correlation to the E. coli concentrations i.e. turbidity. Therefore, in future studies, inclusion of more input parameters may result in better representation of the system on beaches.
- Models generated for Marie Curtis Park East beach to predict E. coli concentration are in a good range but still there is a room for improvement. As the main aim was to use real-time input parameters for the models, some input parameters which were affecting the beach water quality were omitted.
- As a possible extension to this work, it would be quite useful to analyze all the possible neural network architectures and to provide a comparative analysis on each of the architectures and their performance characteristics. The possible neural network architectures that can be analyzed apart from back propagation neural networks are radial basis neural network (RBF).

In addition, to the recommendations listed above, it is critical to remember that use of data driven models, like ANN, require periodic reassessment of the water quality conditions. As over the period of time both the physiographic of the lake and management practices of beach water quality may change, re-evaluation of existing models and/or inclusion of readily available new input parameters may be necessary to maintain or improve predictive performance.

8 APPENDIX

Table A- 1 Summary of ANN models developed for Sunnyside beach using different input parameters combination, training algorithm, transfer function and hidden neurons

No	Model	Input combination	Target	Transfer function	Training algorithm	No. of neurons
1	M1	ln st.fl(3),ln w.ht,ln w.spd,slr,past 48hrs rain(3)**	lnEC	tansig purelin	trainlm	30
2	M2	ln st.fl(3),ln w.ht,ln w.spd,slr,past 48hrs rain(3)**	lnEC	tansig purelin	trainrp	30
3	M3	ln st.fl(3),ln w.ht,ln w.spd,slr,past 48hrs rain(3)**	lnEC	tansig purelin	trainlm	30
4	M1A	ln st.fl(3),ln w.ht,ln w.spd,slr,past 48hrs rain(3)**	lnEC	tansig purelin	trainlm	30
5	M4	st.fl(3),w.ht,w.spd,slr,past 48hrs rain(3)	lnEC	tansig purelin	trainlm	30
6	M4A	st.fl(3),w.ht,w.spd,slr,past 48hrs rain(3)	lnEC	logsig purelin	trainlm	30
7	M5	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	20
8	M5A	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	50
9	M5B	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	100
10	M5C	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	logsig purelin	trainlm	30
11	M5D	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainrp	30
12	M5E	Humber st.fl,w.spd.,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	traingd	30
13	M5F	Humber st.fl,w.spd,slr,HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	30
14	M6	st.fl(3),w.spd,,slr,past 48hrs rain(3), pr. lnEC	lnEC	tansig purelin	trainlm	30
15	M6A	st.fl(3),w.spd,,slr, past 48hrs rain(3), pr. lnEC	lnEC	tansig purelin	trainlm	50
16	M6B	st.fl(3),w.spd,,slr,past 48hrs rain(3), pr. lnEC	lnEC	tansig purelin	trainrp	30
17	M6C	st.fl(3),w.spd,,slr,past 48hrs rain(3), pr. lnEC	lnEC	tansig purelin	trainrp	50
18	M7	st.fl(3),w.spd,slr,past 48hrs rain (HY041,T.W.2), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	5
19	M7A	st.fl(3),w.spd,slr,past 48hrs rain (HY041,T.W.2), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
20	M7B	st.fl(3),w.spd,slr,past 48hrs rain (HY041,T.W.2), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	30
21	M7C	st.fl(3),w.spd,slr,past 48hrs rain (HY041,T.W.2),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50

22	M8	st.fl(3),w.spd,slr,past 48hrs rain (3), pr. lnEC, l.l.	lnEC	tansig purelin	trainlm	30
23	M9	st.fl(3),w.spd,slr,HY041 past 48hrs rain, pr. lnEC	lnEC	tansig purelin	trainlm	30
24	M10	st.fl(Humber,Mimico),w.spd,slr,HY041 past 48hrs rain, pr. lnEC	lnEC	tansig purelin	trainlm	30
25	M11	st.fl(Humber,Mimico),w.spd,slr,HY041 past 48hrs rain	lnEC	tansig purelin	trainlm	30
26	M13	st.fl(3),slr,past 48hrs rain(HY041,T.W.2), pr. lnEC,l.l.,turbidity	lnEC	tansig purelin	trainlm	30
27	M14	st.fl(Humber,Mimico),slr,past 48hrs rain (HY041,T.W.2),pr. lnEC,turbidity	lnEC	tansig purelin	trainlm	30
28	M15	st.fl(3),w.spd,slr,HY041 past 48hrs rain, pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	30
29	M16	st.fl(3),w.spd,slr,past 48hrs rain(3),l.l..	lnEC	tansig purelin	trainlm	30
30	M17	st.fl(3),w.spd,past 48hrs rain (HY041,T.W.2),l.l..	lnEC	tansig purelin	trainlm	30
31	MLR	Humber st.fl, HY041 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	30
32	Mnorain	Humber st.fl,w.spd.,slr,pr. lnEC	lnEC	tansig purelin	trainlm	30

*past 48hrs rain(3)**-past 48hrs rain of station HY041,T.W.2 and T.W.11, st.fl(3)*-streamflow of Humber river,Mimico creek and black creek, pr.lnEC-previous day E. coli count, ln-natural logarithm, w.spd-wind speed,w.ht-wave height ,l.l.-lake level, slr-solar radiation*

Table A-2 Summary of ANN models developed for Rouge Beach using different input parameters combination, training algorithm, transfer function and hidden neurons

No	Model	Input combination	Target	Transfer function	Training algorithm	No. of neurons
1	R1	st.fl,w.dir,w.ht,past 48hrs rain(4)	E. coli	tansig purelin	trainlm	20
2	R1A	st.fl,w.dir,w.ht,past 48hrs rain(4)	E. coli	logsig purelin	trainlm	20
3	R2	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	tansig purelin	trainrp	20
4	R2A	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	logsig purelin	trainrp	20
5	R3	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	tansig purelin	trainlm	20
6	R3A	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	logsig purelin	trainlm	20
7	R3B	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	tansig purelin	trainlm	20
8	R3C	st.fl,w.dir,w.ht,past 48hrs rain(4)	lnEC	logsig purelin	trainlm	20
9	R4	st.fl,w.dir,w.ht	lnEC	tansig purelin	trainlm	10
10	R4A	st.fl,w.dir,w.ht	lnEC	tansig purelin	trainlm	20

11	R4B	st.fl,w.dir,w.ht	lnEC	tansig purelin	trainlm	30
12	R4C	st.fl,w.dir,w.ht	lnEC	tansig purelin	trainlm	50
13	R4D	st.fl,w.dir,w.ht	lnEC	tansig purelin	trainlm	100
14	R5	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	tansig purelin	trainlm	20
15	R5A	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	logsig purelin	trainlm	20
16	R5B	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	tansig purelin	trainlm	50
17	R5C	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	logsig purelin	trainlm	50
18	R5C	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	tansig purelin	trainrp	50
19	R5D	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY011,HY044),pr. lnEC	lnEC	logsig purelin	trainrp	50
20	R6	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	tansig purelin	trainlm	50
21	R6A	st.fl,w.dir,w.ht,w.spd,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	logsig purelin	trainlm	50
22	R7	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	tansig purelin	trainlm	50
23	R7A	st.fl,w.dir,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	logsig purelin	trainlm	50
24	R8	st.fl,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	tansig purelin	trainlm	50
25	R8A	st.fl,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	logsig purelin	trainlm	50
26	R9	st.fl,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	logsig purelin	trainlm	50
27	R10	st.fl, slr, past 48hrs rain(HY044,HY070)	lnEC	logsig purelin	trainlm	50
28	R11	st.fl,past 48hrs rain(HY044,HY070), pr. lnEC	lnEC	tansig purelin	trainlm	50
29	R12	w.dir,w.ht,slr,past 48hrs rain(HY044,HY070), pr. lnEC	lnEC	tansig purelin	trainlm	50
30	R12A	w.dir,w.ht,slr,past 48hrs rain(HY044,HY070),pr. lnEC	lnEC	logsig purelin	trainlm	50
31	R13	st.fl,w.dir,w.ht,slr,HY044 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	50
32	R14	st.fl,w.dir,w.ht,slr,HY070 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	50
33	MLR	st.fl,w.dir,pr. lnEC	lnEC	tansig purelin	trainlm	50

34	MLR-A	st.fl,w.dir,pr. lnEC	lnEC	logsig purelin	trainlm	50
35	MLR-B	st.fl,w.dir,pr. lnEC	lnEC	tansig purelin	trainrp	50
36	MLR-C	st.fl,w.dir,pr. lnEC	lnEC	logsig purelin	trainrp	50

st.fl-streamflow of Rouge River, pr.lnEC-previous day E. coli count, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, ln-natural logarithm, l.l.-lake level, slr-solar radiation

Table A-3 Summary of ANN models developed for Marie Curtis Park East Beach using different input parameters combination, training algorithm, transfer function and hidden neurons

No	Model	Input combination	Target	Transfer function	Training algorithm	No. of neurons
1	M1	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
2	M1A	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	30
3	M1B	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
4	M1C	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	100
5	M2	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	logsig purelin	trainrp	20
6	M2A	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	logsig purelin	trainrp	30
7	M2B	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	logsig purelin	trainrp	50
8	M2C	st.fl,w.ht,Rain 48hrs,Rain 48hrs,pr. lnEC,l.l.	lnEC	logsig purelin	trainrp	100
9	M7	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC	lnEC	tansig purelin	trainlm	20
10	M3	st.fl,past 48hrs rain (HY025,HY033),pr. lnEC	lnEC	tansig purelin	trainlm	20
11	M3A	st.fl,past 48hrs rain (HY025,HY033),pr. lnEC	lnEC	tansig purelin	trainlm	50
12	M4	w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
13	M5	st.fl,w.ht,w.spd,w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
14	M5A	st.fl,w.ht,w.spd,w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50

15	M6	st.fl,w.ht,w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
16	M6A	st.fl,w.ht,w.dir,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
17	M8	ln st.fl,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
18	M9	st.fl,w.ht,past 48hrs rain (HY025,HY033),pr. lnEC,,l.l.,w.temp	lnEC	tansig purelin	trainlm	20
19	M10	st.fl,w.ht,slr,pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
20	M17	st.fl,w.ht,slr,pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
21	M11	st.fl,w.ht,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
22	M11A	st.fl,w.ht,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
23	M12	st.fl,w.ht,w.dir,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
24	M12A	st.fl,w.ht,w.dir,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
25	M12B	st.fl,w.ht,w.dir,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	logsig purelin	trainlm	50
26	M12C	st.fl,w.ht,w.dir,slr,past 48hrs rain (HY025,HY033), pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	100
27	M13	st.fl,slr,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	20
28	M13A	st.fl,slr,past 48hrs rain (HY025,HY033),pr. lnEC,l.l.	lnEC	tansig purelin	trainlm	50
29	M14	st.fl,slr,past 48hrs rain (HY025,HY033),pr. lnEC	lnEC	tansig purelin	trainlm	50
30	M15	st.fl,slr,Hy025past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	50
31	M16	st.fl,slr,Hy033 past 48hrs rain,pr. lnEC	lnEC	tansig purelin	trainlm	50
32	M18	st.fl,slr,pr. lnEC	lnEC	tansig purelin	trainlm	50
33	MLR	st.fl,w.dir,past 48hrs rain (HY025,HY033)	lnEC	tansig purelin	trainlm	50

st.fl-streamflow of Etobicoke creek, pr.lnEC-previous day E. coli count, ln-natural logarithm, w.dir-wind direction, w.spd-wind speed, w.ht-wave height, l.l.-lake level, slr-solar radiation

9 REFERENCES

Blue-Flag-Canada [Online]. environmentaldefence.ca. Available: http://environmentaldefence.ca/issues/blue-flag-canada [Accessed February 17 2014].

Amaral, N. 2010. 2009 Surface water quality summary- regional watershed monitoring program.

Ashbolt, N. J., Schoen, M. E., Soller, J. A. & Roser, D. J. 2010. Predicting pathogen risks to aid beach management: the real value of quantitative microbial risk assessment (QMRA). *Water Research,* 44**,** 4692-703.

Basheer, I. A. & Hajmeer, M. 2000. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods,* 43**,** 3-31.

Beale, M. H., Hagan, M. T. & Demuth, H. B. 2013. Neural network toolbox-user's guide. Natick, MA, USA.: The MathWorks, Inc.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2003. Data tranformation for neural network models in water resources applications. *Journal of Hydroinformatics,* 4**,** 245-258.

Cabral, J. P. S. 2010. Water microbiology-bacterial pathogens and water. *Environmental Research and Public Health,* 7**,** 3657-3701.

Chan, S. N., Thoe, W. & Lee, J. H. 2013. Real-time forecasting of Hong Kong beach water quality by 3D deterministic model. *Water Res,* 47**,** 1631-47.

Chandramouli, V., Brion, G., Neelakantan, R. T. & LINGIREDDY, S. 2007. Backfilling missing microbial concentrations in a riverine database using artificial neural networks. *Water Research* 41**,** 217-227.

City of Toronto 2009. Great city, great beaches: Toronto beaches plan. Toronto: Toronto Water.

Costa, D. J. 2013. *Calculating geometric means* [Online]. Available: http://www.buzzardsbay.org/geomean.htm [Accessed December 2013].

Cybenkot, G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems,* 2**,** 303-314.

Eleria, A. & Vogel, R. M. 2005. Predicting fecal coliform bacteria levels in the Charles River, Massachusetts. *Journal of the American Water Resources Association,* 41**,** 1195-1209.

Enns, A. A., Vogel, L. J., Abdelzaher, A. M., Solo-Gabriele, H. M., Plano, L. R., Gidley, M. L., Phillips, M. C., Klaus, J. S., Piggot, A. M., Feng, Z., Reniers, A. J., Haus, B. K., Elmir, S. M., Zhang, Y., Jimenez, N. H., Abdel-Mottaleb, N., Schoor, M. E., Brown, A., Khan, S. Q., Dameron, A. S., Salazar, N. C. & Fleming, L. E. 2012. Spatial and temporal variation in indicator microbe sampling is influential in beach management decisions. *Water Res,* 46**,** 2237-46.

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

Francy, D. S., Gifford, A. M. & Darner, R. A. 2003. Escherichia coli at Ohio bathing beaches—Distribution, sources, wastewater indicators and predictive modeling. Reston, VA: Water-Resources Investigations Report 02-4285.

Francy, D. S., Stelzer, E. A., Duris, J. W., Brady, A. M., 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. *Appl Environ Microbiol,* 79**,** 1676-88.

Frick, W. E., Ge, Z. & Zepp, R. G. 2008. Nowcasting and forecasting concentrations of biological contaminants at beaches: A feasibility and case study. *Environmental Science & Technology,* 42**,** 4818–4824.

Gershenson, C. 2003. Artificial neural networks for beginners. [Accessed October, 2013].

Gironimo, L. D., Patterson, B. & Mckeown, D. D. 2009. Toronto Beaches Plan. Toronto,ON,Canada.

Hamelin, K., Bruant, G., El-Shaarawi, A., Hill, S., Edge, T. A., Bekal, S., Fairbrother, J. M., Harel, J., Maynard, C., Masson, L. & Brousseau, R. 2006. A virulence and antimicrobial resistance DNA microarray detects a high frequency of virulence genes in Escherichia coli isolates from Great Lakes recreational waters. *Appl Environ Microbiol,* 72**,** 4200-6.

Harmel, R. D. & Smith, L. P. 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *Journal of Hydrology,* 337**,** 326-336.

Haykin, S. 1994. *Neural Networks: A comprehensive foundation,* MacMillan College, New York, Prentice Hall PTR.

He, L. M. & He, Z. L. 2008. Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA. *Water Res,* 42**,** 2563-73.

Heberger, G. M., Durant, L. J., Oriel, A. K., Kirshen, H. P. & Minardi, L. 2008. Combining real-time bacteria models and uncertainty analysis for establishing health advisories for recreational waters. *Journal of Water Resources Planning and Management,* 134**,** 73-82.

Hellweger, L. F. 2007. Ensemble modeling of E. coli in the Charles River, Boston, Massachusetts, USA. *Water Science and Technology,* 56**,** 39-46.

Helsel, D. R. & Hirsch, R. M. 2002. Statistical methods in water resources. *In:* INTERIOR, U. S. D. O. T. (ed.) *Techniques of Water-Resources Investigations of the United States Geological Survey.* Hydrologic Analysis and Interpretation.

Heydari, M., Olyaie, E., Mohebzadeh, H. & Kisi, O. 2013. Development of a neural network technique for prediction of water quality parameters in the Delaware River, Pennsylvania. *Middle-East Journal of Scientific Research,* 13**,** 1367-1376.

Jain, A., Mao, J. & Mohiuddin, K. 1996. Artificial neural networks: A tutorial. *IEEE Computer,* 29**,** 31-44.

Khanna, T. 1996. Foundations of neural networks. *Addison-Wesley.*

Kinzelman, J. & Mcphail, C. 2012. Animal waste, water quality and human health. *In:* Dufour, A., Bartram, J., Bos, R. & Gannon, V. (eds.) *Exposure interventions.* IWA Publishing, London, UK.

Kisia, O. & Uncuoglub, E. 2005. Comparison of three back-propagation training algorithms for two case studies. *Indian Journal of Engineering & Materials Sciences,* 12**,** 434-442.

Krause, P., Boyle, D. P. & Base, F. 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences,* 5**,** 89-97.

Lavender, J. S. & Kinzelman, J. L. 2009. A cross comparison of QPCR to agar-based or defined substrate test methods for the determination of Escherichia coli and enterococci in municipal water quality monitoring programs. *Water Res,* 43**,** 4967-79.

Legates, D. R., Gregory, J. & Mccabe, J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research,* 35**,** 233-241.

Madigan, M. T., Martinko, J. M., Stahl, D. A. & Clark, D. P. 2012. *Brock biology of microorganisms*.

Mass, D. M. L. & Ahlfeld, D. 2007. The development and evaluation of Artificial Neural Networks for modeling indicator organism concentrations.

Mcculloch, W. S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics,* 5**,** 115-13.

MOE 1994. Water management policies & guidelines: Provincial water quality objectives Ontario, Canada: Ministry of Environment and Energy.

Motamarri, S. & Boccelli, D. L. 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. *Water Res,* 46**,** 4508-20.

Nevers, M. B. & Boehm., A. B. 2011. Modeling fate and transport of fecal bacteria in surface water. *The Fecal Bacteria***,** 165-188.

Nevers, M. B., Shively, D. A., Kleinheinz, G. T., Mcdermott, C. M., Schuster, W., Chomeau, V. & Whitman, R. L. 2009. Geographic relatedness and predictability of Escherichia coli along a peninsular beach complex of Lake Michigan. *J Environ Qual,* 38**,** 2357-64.

Nevers, M. B. & Whitman, R. L. 2005. Nowcast modeling of Escherichia coli concentrations at multiple urban beaches of southern Lake Michigan. *Water Res,* 39**,** 5250-60.

Olyphant, G. A. & Whitman, R. 2004. Elements of a predictive model for derermining beach clousers on a real time basis:The case of 63rd street beach chicago. *Environmental Monitoring and Assessment* 98**,** 175-190.

Reicherts, J. D. & Emerson, C. W. 2010. Monitoring bathing beach water quality using composite sampling. *Environ Monit Assess,* 168**,** 33-43.

Rosenblatt, F. 1962. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms,* Washington Spartan Books.

Sarle, W. S. 1994. Neural networks and statistical models. *Nineteenth Annual SAS Users Group International Conference.* SAS Institute Inc., NC, USA.

Shamisi, M. H. A., Assi, A. H. & Hejase, H. A. N. 2011. *Using MATLAB to develop artificial neural network models for predicting global solar radiation in Al Ain city – UAE,* UAE, InTech.

Shamshad, A., Parida, B. P., Khan, I. H., Isa, H. M. & Hussin, W. 2006. Modelling and simulation of monthly DO and BOD records of River Ganges: Box-Jenkins time series approach. *Spatial Hydrology*.

Smith, L. P., Carroll, C., Wilkins, B., Johnson, P., Gabhainn, S. N. & Smith, L. P. 1998. The effect of wind speed and direction on the distribution of sewage-associated bacteria. *The Society for Applied Microbiology,* 28**,** 184–188.

Thoe, W., Wong, S. H. C., Choi, K. W. & Lee, J. H. W. 2012. Daily prediction of marine beach water quality in Hong Kong. *Journal of Hydro-environment Research,* 6**,** 164-180.

TRCA 2009. Source Water Protection: Surface Water Quality Update. Toronto and Region Conservation Authority (TRCA).

Tufail, M., Ormsbee, L. & Teegavarapu, R. 2008. Artificial intelligence-based inductive models for prediction and classification of fecal coliform in surface waters. *Journal of Environmental Engineering,* 134**,** 789-799.

USEPA 2010. Predictive tools for beach notification : review and technical protocol. *In:* OFFICE OF WATER, O. O. S. A. T. (ed.). U.S. Environmental Protection Agency.

USEPA. 2012. *Water: monitoring & assessment.* [Online]. U.S. Environmental Protection Agency. Available: http://water.epa.gov/type/rsl/monitoring/vms511.cfm.

Varma, S. S. & Vijayan, N. 2009. Prediction of fecal coliform concentration in surface water using artificial neural networks. *10th National Conference on Technological Trends.* College of Engineering Trivandrum.

Wikipedia. 2013. *Artificial Neural Network* [Online]. Available: http://en.wikipedia.org/wiki/Artificial_neural_network [Accessed December 2013].

Zhang, G., Patuwo, B. E. & Hu, M. Y. 1998. Forecasting with artificial neural networks:The state of the art. *Elsevier Science,* 14**,** 35–62.

Zhang, W., Wang, J., Fan, J., Gao, D. & Ju, H. 2013. Effects of rainfall on microbial water quality on Qingdao No. 1 Bathing Beach, China. *Marine Pollution Bulletin,* 66**,** 185-90.

Zhang, Z., Deng, Z. & Rusch, K. A. 2012. Development of predictive models for determining enterococci levels at Gulf Coast beaches. *Water Res,* 46**,** 465-74.