2014-08-26

Environmental Risk Mapping for Contamination of Drinking Water Wells Post Flood in Southern Alberta

Eccles, Kristin

UNIVERSITY OF CALGARY

Environmental Risk Mapping for Contamination of Drinking Water Wells Post Flood in Southern Alberta

by

Kristin Eccles

A THESIS PROPOSAL

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN GEOEGRAPHY

CALGARY, ALBERTA

AUGUST, 2014

Abstract

The objective of this research was to determine if there were more cases of contamination in 2013 than in previous years. To determine private groundwater wells in the Calgary Health Zone were impacted by the flood in June 2013, and finally determine what environmental variables influence contamination during a flooding event. The analysis utilizes, test results of total coliform and *E.coli* of private water wells were obtained though Alberta Health Services' Provincial Laboratory (ProvLab) for the period of June 19th to September 30th, 2013. The analysis was completed using ArcGIS 10.2 and R 3.0.2. The results of the regression indicate that total coliform contamination was not impacted by the flood, however, *E.coli* contamination was impacted by floodways, flood fringe, farms, and intermittent water (sloughs).

Keywords: Flood, Water Quality, Spatial Analysis, Geographically Weighted Regression

Acknowledgement

I am grateful for the support from many individuals and groups, in that the completion of this thesis would not be possible without them.

First and foremost, I would like to sincerely thank my supervisor, Dr. Stefania Bertazzon for her never ending support and guidance throughout my whole Master's educations.

Secondly, I would also like to thank my co-supervisor Dr. Sylvia Checkley for the opportunity and support of this interdisciplinary research between the department of Geography and Veterinary Medicine.

I would also like to thanks the Department of Geography for the financial support and various individuals from the department including the SESAH research group for feedback on my research. I would also like to thank my committee members, Dr. Herman Barkema, and Dr. Darren Sjogren for your valuable suggestions and guidance in the various stages of this thesis.

I would also like to thank Alberta Health Services and the Provincial Laboratory for allowing me to be a collaborator.

Lastly, I would like to thank my family and friends for supporting me in all of my endeavours.

Thank you!

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapters One: Introduction

Problem Identification

Canadians are fortunate to have plentiful sources of clean and safe drinking water. Most Canadians have access to this water through publically implemented and monitored municipal water treatment systems. However, as Canada is the second largest country in the world by landmass (9,984,670 km^2), with varying distributions of settlements throughout the country, in some regions citizens do not have access to these municipal water resources. This is more common in rural areas. It is estimated that between three to four million Canadian residents rely on private sources for drinking water, which are usually sourced from groundwater systems (Charrois, 2010). Of those three to four million, about six hundred thousand of them reside in Alberta (Alberta Health and Wellness, 2004).

As most well water systems are not municipally regulated, it is more difficult to ensure the quality of water coming from these private groundwater sources. It is the responsibility of the well owner to ensure quality through testing, and if required, treatment of the water as the integrity of the well water can become compromised through pathways of both microbial and chemical contamination (Charrois, 2010). Alberta Health Services (AHS) recommends that water wells deeper than 50 feet (15.24 meters) should have a bacteriological test twice a year and water wells shallower than 50 feet should have the bacteriological test four times a year. The guidelines also recommended that a chemical analysis be performed once every two to five years (AHS, 2009). However, as testing and treatment is left up to individual due diligence, the testing rates may be much lower than what is recommended by AHS. Non-compliance increases the risk that a greater number of individuals will be exposed to gastroenteric viruses, which can be present in groundwater (Borchardt et al., 2004).

Surface water and groundwater are highly interconnected. Surface water infiltrates into the ground, recharging the groundwater system, and then can be extracted from the ground for the purpose of drinking water (though the borehole/ well). Due to this

interconnectedness, it is possible for groundwater sources to become contaminated via surface water sources. This occurs when water recharging the groundwater source is contaminated. The main mechanisms through which this can occur are if the soil that the recharge water is filtering though is contaminated or if the water is contaminated prior to entering the ground. It is also possible for groundwater to become contaminated when a hazardous substance, such as a chemical leachate, infiltrates though the soil and into the groundwater (EPA, 2011). The amount of time it takes for the recharge water to reach the groundwater systems is dependent on many factors including soil type and surficial geology (USGS, 2014). While there are many different mechanisms of groundwater contamination, the means of contamination that will be focused on in this study will be contamination as the result of overland flooding.

Globally, flooding is the most commonly occurring natural disaster, accounting for 40% of all natural disasters annually (Alderman et al., 2012). There can be notable issues when overland floods occur, such as the increased velocity and volume of water changing the topography of rivers and streams. Large magnitude flooding can change a riverbed, decreasing the ability for the riverbed to filter out contaminated water as it infiltrates into the ground. Additionally, as the magnitude of flooding increases, a greater volume of the clogging layer along the riverbed is removed. This layer normally functions to slow down the rate of water infiltration into the groundwater system. Resultantly, this allows an above normal volume of water to infiltrate into the ground. The large water volumes during floods also increase the risk of carrying point source, and non-point sources of contamination into water wells (Hiscock and Grischeck, 2002).

In an urban setting, heavy rainfall that results in overland flooding has the potential to over burden the sewer systems. When this occurs, a backflow of contaminated water can spill into the city street. Adding this contaminated sewer water to the already existent overland flooding increases the risk for potential for contamination (Hiscock and Grischeck, 2002). This can cause negative repercussions in rural settings as the water can move from the urban environment to a more rural setting where overland flooding is more likely to increase the amount of runoff contaminated with animal faecal

matter. Additionally, the higher than normal river volume creates more turbid water conditions in the rivers. This effect increases with the magnitude of the flood. The increase in turbidity increases the risk that any microorganisms that have previously settled out of water system for example in the rock bed of a river will become resuspended to then be transported further downstream (Hofstra, 2011).

In June of 2013, Southern Alberta was the site of devastating flood that claimed the title of Canada's most costly natural disaster. It is expected that the cost to recover from this flood will be roughly six billion dollars. The flood affected over 100,000 Albertans in 30 different communities of which 40,000 people were evacuated from their homes. Infrastructure including over 1000km of road, bridges and houses were destroyed by the flood (Alberta Government, 2013).

Although the month of June is typically the wettest month in Southern Alberta, this flood was a result of many factors. Contributing factors include a higher than normal snowfalls in the mountains from October 2012 to March 2013, excess amount of precipitation during the early spring where the Bow and Elbow River watershed received more than 300mm of precipitation, and an already wet spring leaving soils saturated having no room to absorb any additional precipitation (City of Calgary, 2013; Environment Canada 2013a). Additionally, leading up to the flood, the mountains received an extended period of rain caused by a slow moving low front. This accelerated the snow melt at a rate 25% faster than normal. As much of the ground in the mountains was still frozen, water was forced to run off into the rivers instead of infiltrate into the ground (Environment Canada, 2013a). At its peak discharge rate, it is estimated that the Bow River was flowing around 1700m3/sec, 7.5 times greater than the mean discharge rate since 1911 (City of Calgary, 2013).

Research Purpose

The objective of this research is to determine which environmental factors are associated with the contamination of private drinking water wells and drinking water quality and safety in the Calgary Health Zone in the time immediately following the June 2013 flood. It is important to determine which wells are at a higher risk of becoming

contaminated with bacteria and therefore which wells are more likely to pose a public health risk to individuals living in rural areas. This information will be used to produce a risk assessment map. This map will demonstrate areas where there is higher risk of drinking water well contamination during a flood from total coliforms or *E. coli* bacteria in the Calgary Health Zone. Ultimately this analysis will be used to facilitate the transfer of knowledge so that homeowners who live in areas that are more prone to well water contamination during a flood can be better prepared. This study can also be used as a resource for public health investigators at AHS to plan for, and identify private drinking water wells in high-risk areas. This knowledge will help to identify individuals who may require more assistance with remediation of a contaminated drinking water well.

Research Objectives

The overarching objective of this research is to model factors that influence private drinking water well contamination during a flood in the Calgary Health Zone. This will be achieved through both descriptive and quantitative methods.

1. Historical Analysis
    a. Historical analysis of test results from June 19th, 2013-September 30th, 2014 compared to the same time period dating back to 2005.
2. Descriptive Analysis
    a. Maps of the spatial distribution of positive and negative well test results and well attributes.
3. Quantitative Analysis
    b. Statistical model to determine what environmental attributes influence water well contamination during a flood.
    c. Based on the regression model produce, a risk map/ risk surface will be created for the Calgary Health Zone, showing areas that are at low, moderate, and high risk of drinking water well contamination.

Chapter Two: Background

Groundwater and Vulnerability in Alberta

Traditionally, water sources are categorized into surface and groundwater. However, this descretization of a continuous system, the hydrologic cycle, detracts from the notion that this cycle is continuous with water constantly moving. Rather, the hydrologic cycle should be thought of as dynamic rather than static (Bear, 2012). Figure 2 demonstrates the interconnectedness between surface and ground water, as well as various types of groundwater sources that are used as drinking water sources.

Generally, the term groundwater refers to any type of water in the ground. Surface water from sources of precipitation, or surface water bodies such as lakes and stream will infiltrate into the ground. Surface water serves as the recharge for ground water. This water is then retained within the pore spaces of ground sediment or rock that allows for water in varying quantities to be stored (EPA, 2012).

Groundwater can be structured very differently depending on the geology of the region. Specifically in Alberta, the first layer of sediment subsurface is typically unconsolidated sediment, which is comprised of unlithified till containing clay, silt, sand and gravel. Under the surficial sediments, bedrock, which was deposited about 75 million years ago, can be found in Alberta. The rock types typically found in this layer are sedimentary rock, which includes shale, siltstone, sandstone, mudstone, claystone and in some areas even coal (AARD, 2010). In some regions of the province, this sedimentary rock can be exposed, whereas in others this rock is buried more than 100 meters below the surface. In some areas, sand and gravel deposits are present between the surficial layer of till and the bedrock. Bedrock can range from being highly fractured (e.g. sandstone), which allows groundwater to flow more freely than unfractured bedrock (e.g. shale). Highly fractured rock, where groundwater is able to flow more freely, can be defined as an aquifer. Unfractured rock that is able to hold water in the porous spaces, but cannot transmit the water at a fast enough rate to sustain a water supply, is considered an aquiclude or aquitard. Aquifers have high hydraulic conductivity while,

aquiclude or aquitard are said to have low hydraulic conductivity (AARD, 2010; Bear, 2012).

The water table is found within the upper surface, closer to ground level. In Alberta, the water table is typically within 5 meters of the surface (AARD, 2010). However, as the topography of the land changes so can the depth of the water table. The properties of the sediment under the water table and above the aquifer dictate how quickly water is able to move through the system. Regions that have sediment with large connected pore space will allow the water to flow more quickly between various regions within the ground water system. Typically in Alberta, the sediment properties are in between that of an aquifer and an aquiclude (AARD, 2010).

The susceptibility of groundwater to contamination from surface sources can vary greatly from place to place and as mentioned above, all groundwater sources are not the same. There are two main types of aquifers; these are confined and unconfined. Confine aquifers, are aquifers that occur under geologic formations that have a very low hydraulic conductivity, making it much more difficult for water, and contaminants to pass through. Water in this type of aquifer has more protection from contaminants than others. Conversely, unconfined aquifers are aquifers that are not overlaid with rock having low permeability. Typically, these wells are shallower and occur closer to the surface. Not having a protective layer with low permeability between the aquifer and the surface makes these types of aquifers more susceptible to contamination from surface sources (AARD, 2010).

Surficially, many factors can play a role in the vulnerability of groundwater. The vulnerability of a groundwater sources are said to be independent of the pollutant and instead relies upon environmental factors such as land use and soil variables (AARD, 2010). Surficial qualities in combination with the aquifer qualities are associated with the risk of contamination of groundwater. In particular, unconfined shallow aquifers comprised of sand/ gravel are at a higher risk of contamination groundwater contamination due to the proximity to pollutants, lack of protection from an impermeable rock layers, and quick recharge from the surface. Areas are at an even higher risk of

contamination if there are sources of pollution in the area, particularly agriculture. Due to the high volume of farming in Alberta, this is a problem particularly in southern Alberta (AARD, 2010). Groundwater analyses in Alberta for nitrate, *E. coli*, and total coliforms reveal that contaminates can leach into the water table, which could affect the water quality of wells for drinking water that are close to farming operations (McCallum et al., 2008; Olson et al., 2008). Conversely, the safest aquifers are those that are confined with a very thick (30m or greater) layer of unfractured clay or shale, due to the extremely low permeability of these materials. In such groundwater systems, it can take 1000 years for water to permeate the rock and travel through to the aquifer (AARD, 2010).

A common tool used to determine groundwater vulnerability is the DRASTIC method which is an application designed by the National Water Well Association (NWWA) and the United States Environmental Protection Agency (USEPA). This index serves as a relative evaluation tool. The assumptions of this model are that the contaminant is introduced to the groundwater from the surface. The mobility is due to precipitation. The pollutant is mobile in water and the area of the model is greater than 100 acres (0.41 km2). The acronym stands for: Depth to water, Net Recharge, Aquifer Media, Soils, Topography, Impact of vadose zone, and Hydraulic Conductivity. This index is used by introducing predetermined rankings (r) and weightings (w) to each of the variables. This equation can be seen in Equation 1.

$$\textit{Drastic Index} = D_r D_w + R_r R_w + A_r A_w + S_r S_w + T_r T_w + I_r I_w + C_r C_w \qquad \textbf{Equation 1}$$

The greater the depth (D) to the available groundwater, the longer it takes contamination to reach the aquifer. This value is typically based on water well log data. The net recharge (R) value is the amount of water that enters the aquifer. While an increased amount of water recharge can dilute the pollutant, at the same time, recharge is a major pathway for contamination transportation. The net recharge can be estimated based on a mass water balance using precipitation, evaporation, and runoff values as inputs and outputs. The aquifer media is rated based on the composition of each layer in the media. The higher the permeability of all the media, the higher the risk of contamination. The type of soil can affect the types of pollutants that can pass through.

The soil type can also affect the microbial, and breakdown processes of both chemical and microbial contaminants. Table 1 is an example of the proposed ratings for different soil types.

Table 1. Soil Type Rating System Proposed by Aller et al. (1987)

Range	Rating
Thin or absent	10
Gravel	10
Sand	9
Peat	8
Shrinking and/or Aggregated Clay	7
Sandy Loam	4
Loam	5
Silty Loam	4
Clay Loam	3
Muck	2
Non-shrinking and non-aggregated clay	1

Topography (T) can also have an impact on groundwater quality. Particularly the slope of the land can be associated with runoff, which can determine whether the contaminant is more likely to be part of the runoff or more likely to infiltrate into the ground. Low degrees of slope will have less runoff and therefore any contaminants would be more likely to infiltrate into the ground. The vadose zone (V), also referred to as the unsaturated zone, is associated with how water is able to more within the groundwater system. Zones with a higher permeability are at a higher risk of contamination as pollutants are able to move more freely between zones. Finally, hydraulic conductivity (C) refers to the fractures of the aquifer, which is associated with the movement of water within the ground. The higher the hydraulic conductivity, the higher the risk for contamination is. Overall, the higher the final value of the DRASTIC index, the more susceptible the aquifer is to becoming contaminated.

Although this method was developed over 35 years ago, this method is still widely used today (Babiker et al., 2005; Panagopoulos et al., 2006; Rahman, 2008; Neshat et al., 2013). It is particularly favoured in applications of geographic information systems (GIS) as it allows for efficient analysis and capacity to handle large data sets that cover a wide

geographic area (Babiker et al., 2005). Additionally, GIS allows for visualization of the final classification. Advancements to this model include using this method in conjunction with statistics and geostatistical techniques to enhance the ratings and weightings in the model (Panagopoulos, Antonakos, and Lambrakis, 2006). A study by Neshat et al., (2013) demonstrated the DRASTIC model was an accurate method for determining the risk of non-point source pollution especially in areas where agriculture was a major industry.

Alberta Agriculture and Rural Development have completed aquifer vulnerability mapping for the province. Although the DRASTIC method was not used, a similar method was utilized. The modified version of the Aquifer Vulnerability Index (AVI) developed by Van Stempvoort et al. (1992 and 1993). This method uses inputs similar to those used in the DRASTIC method, such as the permeability of geologic features, depth to aquifer, and surrounding geologic material. After combining the various inputs, the vulnerability of the aquifer in each region was ranked as high risk, medium risk and low risk. Here, high risk means that it takes a short amount of time for surface water to percolate through the ground material and into the aquifer. This type of contamination can take a few years to occur, while low risk of contamination areas would take thousands of years for contaminated water to percolate through the ground material to reach the aquifer (ARSD, 2005).

While it is recognized that the intent of this research is not to produce an aquifer vulnerability assessment or specifically use the DRASTIC method. As mentioned above, surface and groundwater do not function independent of each other and due to this it is important to take into consideration all factors that may play a role in modeling water well contamination. Therefore, factors that make an aquifer more vulnerable to contamination such as variables used in the DRASTIC method will be taken into consideration in this study.

Total Coliform

'Total coliforms' is the general classification for rod-shaped (bacilli) gram-negative non-spore forming bacteria. Typically, these bacteria are aerobic but can also be facultative anaerobic, having the capability to survive in both aerobic and anaerobic

environments (World Health Organization, 2004). A gram stain can be used to distinguish gram-negative from gram-positive bacteria. Gram-negative bacteria show up as a red colour as the crystal violet stain used in a gram stain is not able to fully penetrate the thick multilayer the cell well (Maloy et al., 1994). It is important to distinguish the difference between gram- negative and gram-positive bacterium as they have different structures and internal processes, and therefore need different approaches for treatment.

Coliform bacteria are wide spread in the environment and include but are not limited to the genera *Escherichia, Citrobacter, Klebsiella, Enterobacter, Serritia*, and *Hafnia*. These genera can be found in faecal environments such as the intestines of warm-blooded animals as well as the natural environment including in the structure of plants and soil. While the presence of total coliforms does not necessarily indicate that a harmful bacterium is present, it is a good indication of the cleanliness of an environment (World Health Organization, 2004). Due to this, a total coliforms count is a widely used indicator of potable water quality in North America. The total allowable limit for total coliforms in drinking water is zero total coliforms per 100 millilitres (Weiner, 2012).

Faecal Coliforms

A subgroup of total coliforms, used to indicate faecal contamination, faecal coliforms encompass bacteria that are of both faecal and non-faecal genera. For example, *Klebsiella* coliforms are classified under the faecal sub-group, however the major source for this genus is typically textile factories including pulp and paper mills. However, the source of most faecal coliforms contamination comes from farms in the form of animal waste or from septic beds in the form of human sewage. Faecal coliforms are another indicator used to measure water quality (Wiener, 2012). In this thesis, the specific faecal coliforms species that will be focused on is *Escherichia coli (E. coli*).

Escherichia coli

This bacterium is a subgroup of the faecal coliforms group, and is commonly found in both human and animal faecal matter. It is very rare to find the presence of *E. coli* without also finding faecal pollution. Presence or absence of *E. coli* is considered the most reliable indicator of faecal contamination. As a result, testing for this bacterium is

regarded as the optimal choice for drinking water surveillance (World Health Organization, 2004). Certain strains of *E. coli*, for example *E. coli* 0157:H7, can cause gastroenteritis related outbreaks and even deaths as occurred in Walkerton, Ontario in 2000 (Hrudey et al., 2003). The maximum allowable limit of *E. coli* in drinking water is also zero colonies per 100 millilitres as well (World Health Organization, 2004).

Environmental Variables

Many studies that have examined the contamination of surface waters with Total Coliforms (TC), Escherichia coli (EC), and other microorganisms, as a result of environmental conditions such as run-off, increased precipitation, and effects of seasonality (Dorners et al., 2007; Hofstra, 2011). One of Canada's recent incidents of well water contamination that has happened in the recent past of Canada occurred in Walkerton, Ontario. This incident can provide valuable information on how the environment can impact the contamination of water wells that are sources of drinking water. First, Walkerton is situated on a surface where the geologic conditions are considered very poor. This means there is a large amount of highly fractured rock surrounding the water wells allowing recharge between surface and ground water to occur very quickly. In addition to the poor geologic conditions, one of the wells was very shallow. Another factor in this occurrence of contamination was the distance between water wells and proximity to farming operation. Also, these wells were located within a close proximity to two different farms that served as a potential for point source faecal contamination. In particular it was previously known that certain wells were susceptible to surface water influence. These environmental factors, in addition to 70 mm of rain that fell within a few days were enough to carry the *E. coli* into the drinking water well supplying the town's drinking water causing contamination. In this case, even though groundwater was delivered to residents of the town through municipal facilities, the chlorine treatment of the water was improperly managed leading the treatment method ineffective (Hrudey et al., 2003).

A study conducted by Wallender et al., (2013) looked at factors that contribute to contamination of untreated groundwater system and gastroenteric disease outbreaks.

This study was completed by assessing disease outbreaks that were associated with the untreated groundwater used for drinking and then examining the factors that contributed to these outbreaks. Commonly, these wells were contaminated with either human waste (from septic systems) or animal waste. This study found that most often the improper well design, unsatisfactory upkeep of the well, and the improper location of the well (inadequate setback from septic tanks, and waste site) were contributing factors to contamination. A geologic factor found to be significant was lithology that allowed for the rapid transport of water from the surface into the groundwater system. This was measured using a groundwater vulnerability maps. The environmental factors found to be significant were heavy periods of rainfall, which lead to flooding events. Additionally, wells located in depressions or downhill were more likely to become contaminated.

Richardson et al., (2009) examined the relationship between climatic factors, environmental factors, and the contamination of *E. coli* and TC in drinking water wells in England. First, a univariate analysis was undertaken using a chi-squared test to determine significance of each variable. Then only the variables that were significant at the 98% confidence interval (CI) were included in the multivariable logistic regression model. Significant variables included region or sample, year of sample, month of sample, supply type (domestic or commercial), treatment (treated or untreated), the reason for sampling, water source, amount of rainfall on previous day, rurality (population density), density of cattle and density of sheep per km2. Presence absence data was used as the dependent variable in a logistic regression and the environmental variables listed above were used as independent variables. The variables in the multivariate analysis were chosen based on a backwards selection method, eliminating the insignificant ($p>0.1$) variables one at a time. The multivariate model determined the year and month of the sample to be significant. To account for the uneven number of samples taken at each individual water well, the number of samples taken from the site inversely weighted the sampling locations. The variables that proved to be significant in the regression model includes year and month, classification, source, treatment, amount of rainfall on previous day, and sheep density (km2). These results demonstrated seasonality to groundwater contamination where

levels of contamination are highest in spring. Another interesting correlation is that sheep were considered significant but cattle were not. The authors deduced this to the relationship to an *E. coli* to manure ratio. While cattle do produce more manure, sheep actually have two times the concentration of *E. coli* in their faeces than cattle. Additionally, the amount of rainfall one day before the sample was the climatic factor with the most significance.

Climate Change

The Intergovernmental Panel on Climate Change (IPCC) is an intergovernmental body that reviews and synthesizes peer reviewed and non- peer review research pertaining to climate change and climate monitoring. This organization produces synthesized reports pertaining to the changing climate (IPCC, 2010). These reports state that extreme weather events such as heavy rainfall events, and tropical cyclones are increasing specifically in the North Atlantic Ocean (McMichael et al., 2006; Bouwer, 2011). The scientific community attributes this increase in natural disasters to anthropogenic drivers of climate change. Based on climate modeling, forecasters predict that by 2100, global, temperature will rise between 1.4°C and 5.8°C. Higher latitudes will experience a greater degree of climate change than central latitudes. This is an issue as temperatures increase climate becomes more variable (McMichael et al., 2006).

Heavy rainfall events and resultant floods are increasing in frequency and magnitude as the climate increases (McMichael et al., 2006). Flooding is the most frequently occurring natural disaster, which at the local scale is affected by the amount of rainfall received, topography and resulting run-off, water evaporation, and in coastal regions, sea level. These local scale impacts can be mitigated through land use practices including urbanization, and forestry methods, which will be elaborated on in the further in this section. At a global scale, the driver of these types of events is predominantly the El Niño Southern Oscillation (ENSO). Driven by ocean temperatures, this cycle affects, atmospheric pressure, trade wind patterns, and resultantly weather, particularly rainfall. As the temperature rises, these ENSO episodes are lasting longer and are producing more variable climatic conditions. Over last 30 years, the number of individuals who have been

affected by this weather events related to the ENSO has greatly increased (McMichael et al., 2006).

In some areas, local populations are able to adapt to the changing frequency and magnitude of naturals disasters occurring through various adaption methods (McMichael et al., 2006). However, in some regions the rate at which these natural disasters are occurring is more frequent than risk reduction measures can be achieved. This can prove to be disastrous when taking into account the economic burden that is incurred from increased natural disasters (Bouwer, 2011). This can also lead to health implications when a population is stressed beyond their ability to adapt to the changing environment (McMichael et al., 2006).

Flooding and Urbanization

While this thesis focuses on flooding in rural areas and related water well contamination, rural flooding cannot be considered independent of urban influences. Urbanization is a growing trend globally with the population of urban residents increasing by about 60 million people a year. It is estimated that six out of every 10 people will live in an urban city by the year 2030, and seven of 10 in the year 2050 (WHO, 2014). With a growing urban population, urban centers are growing due to urban sprawl to accommodate the influx of people. The method by which urbanization and urban sprawl occur is through the development of natural land into features of a built environment containing many impervious surfaces (Lee et al., 2006; Sheng and Wilson, 2009). There is evidence that increased urbanization alters the natural landscape of watersheds, modifying the risk of flooding in certain areas. Studies show that urbanization on a watershed can change the geomorphology of river channels, vegetation that serves riparian functions along riverbanks, and stream flow characteristics. As well, increased area of impervious surface increases the frequency and magnitude of flooding. By changing the natural features, the relationship between the hydrologic cycle and the earth is altered. Traditionally, precipitation will be absorbed into the surface of the earth, however, if an impervious surface is constructed, water can no longer absorb into the ground, and instead is forced to run off. This can cause increased pooling of water in low

laying areas as well as increase the amount of pollution that is carried downstream (White and Greer, 2006; Sheng and Wilson, 2009). With evidence demonstrating that climate change is increasing the frequency and magnitude of rainfall events associated with flooding, and that urban sprawl is changing the dynamics of how the earth is able to respond to increased rainfall, the impacts that these two influences have will only continue to increase over time.

Health Effects of Flooding

There are both long-term and short-term health effects that result from flooding, although the long-term effects are less well understood (Alderman et al., 2012). A study conducted by Pitt (2008) post flood in the counties of Hull, Worcestershire, and Gloucestershire, United Kingdom, reported that 64% of individuals affected by the flood reported adverse health effects. The adverse mental health effects included anxiety, heightened stress, and depression and adverse physical affects which included dermatitis, chest infections, and heightened episodes of asthma and arthritis. In this study, 70% of individuals who had to evacuate their residencies because of the flood reported adverse effects on their mental and physical health. Long-term health issues resulting from the flood were stress and anxiety related debt issues, anxiety due to the loss of security and home (Gray, 2008).

Water associated with overland flooding serves as a vector to carry microbe-causing illnesses from one location to another. There is also strong evidence to suggest a link between heavy rainfall events and waterborne outbreaks of disease (Hrudy et al., 2003; Hofstra, 2011). Examples of the link between heavy rainfall causing flooding and contamination of drinking water sourced from a water well that have results in outbreaks of disease and death can be seen below. One of the largest outbreaks in the United States was in Milwaukee, Wisconsin in 1993, which resulted in 403,000 cases of gastroenteritis and 54 deaths caused by the contamination of ineffectively treated municipal water after a heavy rainfall (Auld et al., 2004). In Walkerton, Ontario in 2000 there was an outbreak of over 2,300 cases of gastroenteritis of which 65 cases needed hospitalization, and seven deaths occurred following a heavy rain event (Hrudey et al., 2003). Not only does

contamination of drinking water well pose a risk of harm to individuals, but it also puts a strain on the health care system with increased number individuals requiring hospitalizations. As the frequency and magnitude of natural disasters such as flooding, heavy rainfalls, and hurricanes increase due to global warming (Few, 2004; Hofstra, 2011), it is important to examine water contamination as a results of these events and determine risk factors that may contribute to the contamination of private drinking water wells.

Regression Modeling

Aspatial Regression Models

Regression modeling is a commonly used tool across many disciplines. The main goal of such analysis is to model the relationship between a dependent variable and one or more independent (also called explanatory) variables. Although there are many different types of regression, the simplest multivariable linear regression can be seen below in Equation 2.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon \qquad \textbf{Equation 2}$$

In the above equation, Y (the dependent variable) is modeled based on the equation for the line, wherein the beta values (β) are the estimated parameters, x is the value for each independent variable, and ε is an error term. The β parameters are estimated using the ordinary least squares (OLS) method. This method optimizes the equation of the line to minimize the sum of vertical squares between the known observation and the line created. However, this type of regression can only be used if all assumptions of the OLS method are met. The assumptions for this regression are as follows. The relationship must be linear. The square matrix of the independent variables must be invertible; therefore, there must be no multicolinearity between the independent variables (high correlation is generally defined as correlation greater than 0.7). The error term must be homoscedastic, meaning that the error terms must have a constant variance. The error must also be independently distributed, meaning that there is no (spatial) autocorrelation. If all these assumptions are met then the estimates are is considered Best Linear Unbiased Estimators (BLUE) they have low variance, making the

estimates more reliable. Here, unbiased means that the mean of the estimator equals the true parameter that is, there is no systematic error, best means with minimal variance (which implies spatial stationarity) and no spatial autocorrelation. The assumptions on the spatial properties of the error will be addressed further in this section (Chatterjee and Hadi, 2006).

When data is not normally distributed and does not meet the aspatial assumptions of an OLS linear regression, an alternative method that is commonly used is a generalized linear model (GLM). A non-normal distribution of data is seen in binary data and can be seen in count data. A GLM method can be used with both continuous and discrete data. This model uses less stringent regression assumptions so different data types can be utilized. There are three parts to a GLM. The first is a random component, which is a function of the response variable and probability distribution. The second is a systematic component, which is the relationship between the explanatory variables. The third is the link function, which is the relationship between the random and the systematic components. It is the link function that allows the assumptions of a traditional regression to be less stringent and the ability to model data and relationships that do not meet OLS assumptions (Breslow, 1996). The general equation for a GLM can be seen below in Equation 3.

$$YM = BX + E \qquad \textbf{Equation 3}$$

In this equation, Y is the value of the dependent variable, which is a function of the beta value (B) multiplied by the value of the independent variable (X) as well as the error term (E). The M term is unique to GLMs, and is the matrix of the coefficients that defines the transformation of the dependent variable. This term is specific to the type of GLM chosen, corresponding to the link function.

Although GLMs are a less stringent version of the traditional linear regression model, there are still assumptions that must be met in order to produce the most accurate and most interpretable model. The first, assumption, much like the linear regression, is that the residuals must be homoscedastic, normally distributed, and independent of each other. This is due to the method used to calculate the inferential statistics in the GLM

models, the chi- squared test. This test has assumptions of its own. The chi-squared test requires the distribution of the error to be independently, identically, and normally distributed. If the error is distributed in a heteroscedastic fashion, rather than homoscedastic, this produces a higher degree of variability in the error, which produces a less reliable model. This also applied to the independence of the residuals (Dobson, 2002). Mitigating the issues associated with data that exhibit properties of spatial dependence and heteroscedasticity will be explored further on in this section.

There are many different types of GLMs available for use. One assumption of the GLM pertains to choosing the correct identity of the GLM for the data that is being used. This assumption states that the identity must be specified properly using the correct link function. The type of data being used determines the link that should be used in the GLM. For example if the data is count data and has a Poisson distribution, a log function should be used but if the data is binary, then the logit link function should be used (Breslow, 1996).

Another assumption of the GLM is that the variance function, which relates the mean to the variances, must be correctly specified. Examples of variance functions are log, logit, probit. As well more specifically for a Poisson GLM, the data must be equidispersed, meaning that the mean must equal the variance. If the data does not met the assumption of equidispersion, the data could either be overdispersed where the variance is greater than the mean, or underdispersed where the mean is greater than the variance (McCullagh and Nelder, 1989). If the data is overdispersed a quasipoisson model, or a negative binomial model should be used (Ver Hoef and Boveng, 2007), whereas if the data is underdispersed a Conway Maxwell Poisson distribution can be used (Shmueli et al., 2005).The equation used for a Poisson regression specified with a log link can be seen below in Equation 4.

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \textbf{Equation 4}$$

Very similar to a simple multi-linear regression, the dependent and independent variables are related linearly. However due to the non-normal distribution of the data, the

dependent and independent variables are related through the log function (ln). In this equation it is the log of the dependent variables, and is the calculated as the exponent of the natural logarithm (Euler–Mascheroni constant = 2.71828) for the independent variables. Specific to Poisson models, an offset can also be added to the regression, which enables the model to account for an uneven sample/count distribution that can be present in the dependent variable. Poisson regression models are also referred to as log-linear models as the offset must be logged when fitting the regression equation to constrain the offset to one (Dobson, 2002).

Regression Modeling with Spatial Data

As mentioned above, there can be many problems associated with the statistical reliance when using spatial data (Brunsdon et al., 1998; Brunsdon et al., 1999; Getis and Aldstadt, 2010). Spatial data is known to commonly violate the previously mentioned assumptions of a traditional regression model. In an aspatial OLS of GLM regression, the assumption that the error be independently an identically distributed is commonly violated by spatial data.

This can be informally described by the so- called Tobler's first law of geography that states, "[e]verything is related to everything else, but near things are more related than distant things" (Tobler, 1970: 236). This spatial phenomenon often causes the error in a regression utilizing spatial data to not have an independent distribution. As a result, the model no longer has the smallest amount of variance. Therefore, the beta value is less reliable as a result of the high variance. Methods of spatial autoregression (SAR) are commonly used to properly address the issues associated with spatial autocorrelation. These methods take into account the effect that neighbouring observations have on a given observation. By taking into account the effect that neighbours have on any observation, the model is able to better capture the underlying spatial processes that are likely associated with the dependent variable, hence, with the observed spatial autocorrelation (He et al., 2003). An autoregressive model can be described in Equation 5.

$$Y = \rho WY + \varepsilon \qquad \textbf{Equation 5}$$

In this equation, ρ (the Greek letter rho), denotes the autoregressive parameter of the equation, and will range between negative one and positive one. W is the spatial weights matrix, which can be defined based on proximity of neighbouring points, shared borders, nearest neighbour, or contiguity matrices. Y is the dependent variable and ε is the error (Plant, 2012). The basis of an autoregressive function is that the spatial dependence is removed from the regression by multiplying the independent variables by the inverse of the model's spatial dependence. In doing so, the spatial dependence is removed from the regression model, allowing for a reduction in the variance (Fotheringham and Rogerson, 2008).

It is important to note that SAR does not use OLS to produce beta values, but instead uses a maximum likelihood estimator (MLE). Due to autoregressive nature of the model, where the regression is performed on itself (spatially lagged), and this model must be estimated with a MLE or generalized least square. With these methods some of the assumptions are released and the regression can be estimated, yielding estimates with optimal properties. More commonly, MLE is used. This method predicts the beta values based on the probability density function of the error to maximize the likelihood of the function. However, MLE is only suitable for larger datasets, as the estimator can be biased for small samples. As this method does not yield typical R^2, the goodness of fit of a model is frequently assessed using information criteria such as the corrected Akaike information criterion (AICc).

Another assumption that is commonly violated by spatial data is the error identical distribution meaning that the event must be stationary over space. When a global regression is applied to a study area, it assumes that the relationship between dependent and independent variables is constant over the whole region. However, when the relationship is non-stationary the regression will over predict the ß coefficients in some areas, and underestimate them in others (Brunsdon, 1998). Non-stationarity is strictly a function of spatial data (Brunsdon et al., 1999). The requirements for stationarity are that the data has a constant mean, constant variance, constant covariance, and no directionality. These conditions ensure that the data has equal intensity and does not vary

over space. When these conditions are not met, the non-stationarity creates a heteroscedastic error (ε), which will increase the variance of the error and weaken the statistical inferential power of the model (Fotheringham et al., 2003).

Geographically weighted regression (GWR) is commonly used in order to deal with this issue of non-stationarity in regression modeling and reduce the variance in the model. A GWR is performed at a local level, and is able to address the issue of non-stationarity by taking into consideration the changing relationships of the variable over space. In contrast, a typical regression is performed at a global level. A general GWR equation can be seen below in Equation 6.

$$Y_i = \beta_{0i}(u) + \beta_{1i}(u)x_{1i} + \beta_{2i}(u)x_{2i} + \ldots + \beta_{mi}(u)x_{mi} \qquad \textbf{Equation 6}$$

In a GWR the dependent variable (yi) is locally explained by ßi, which represents the local coefficients of the independent variables at a given location (u). The GWR yields many local regressions, typically the same number of regressions, as there are spatial sample points. By decreasing the size from a global regression (simple linear regression) to a local geographically weighted regression, this causes each regression that is performed to be stationary, allowing the error to become identically distributed through the model. This decreases the variance of the model, which increases the reliability of the beta estimates. Model selection or geographically weighted regressions can include simple linear regressions, Poisson regression models for assessing the geographic distribution of rare event over a geographic region, and logistical regressions, which uses non-parametric nominal data (Charlton et al., 2009). Bandwidths and kernels are used to determine the range of the estimation for each local regression (Fotheringham et al., 2003).

The bandwidth of the kernel can either be fixed, where it is based on a certain distance, or it can be adaptive, where it is based on a predetermined number of units around each observation point. The latter is also referred to as nearest neighbour distance (Getis and Aldstadt, 2010). This bandwidth can have a large impact on the matrix and the resultant weighting function. If the kernel is too large than it will encompass most of the study area, making it more comparable to a global estimate. This would defeat the purpose of a local GWR. Conversely, if the bandwidth is too small, then this can result in

having very few data points within each region. This would cause the degrees of freedom to be small, creating a large standard error, which would render the model less statistically significant. It can be difficult to choose the appropriate bandwidth. Often the selection of a bandwidth is data led. As with all spatial weighted matrices, these created bandwidths are arbitrary and are not present in nature. By creating these arbitrary breaks, the relationship seen between dependent and independent variables, are significant globally, but may not be significant locally. In part due to this issue, a GWR is better suited to large dataset (Brunsdon et al., 1998).

A popular model selection for a GWR is the use of a Gaussian kernel with a distance decay function. Here the further away a point is from the central event in the kernel, the less influence it has on the model (Brunsdon et al., 1998). This kernel is used to determine the weighting function produced through the matrix. Higher weight is placed on events that have a greater influence. The weighting function always had a sum of one. Any point outside the bandwidth would be weighted as zero. The weighing function for the coefficients (ß) can be seen below in Equation 7.

$$\hat{\beta} = (X^T W_i X)^{-1} X^T W_i y \qquad \textbf{Equation 7}$$

In this equation, a matrix weighting function determines the estimated ß coefficient. In this equation Wi refers to a series of weighting functions determined by the chosen kernel, which is used to derive the diagonal of a created square matrix. The whole term $(X^TW_iX)^{-1}$ is the inverse of the variance-covariance matrix. The X^TW_iy term is the weighted variance covariance matrix of the dependent variable (Charlton et al., 2009). GWR has applications in different types of regression modeling and can also be applied to GLM models where a geographically weighted generalized linear model (GWGLM) is created. This model is an integration of the GLM and GWR equations above.

While a GWR can adequately address the issues of non-stationarity, which can weaken the statistical power, there are some issues to be cognizant of when using this model. This method does not address the other regression assumptions frequently violated by spatial data; that the error be independently distributed. As previously discussed, non-independent error distribution is generally associated with spatial

autocorrelation. While not all spatial data is spatially autocorrelated, most of this time this is the case (Brunsdon et al., 1999). The GWR model fails to take the potential for spatial autocorrelation into consideration, and can sometimes even aggravate it. This is due to the small regions created by the GWR. The events within that each local area may have increased similarity compared to the global area, therefore inducing spatial autocorrelation. Therefore, the model produced by in a GWR may not be completely BLUE and it may be possible that the statistics are still weakened by spatial autocorrelation.

Furthermore, when there is spatial autocorrelation of the model's error terms, this can cause problems of undersmoothing which can affect the weighting function of the kernel, ultimately altering the number of events that are taken into consideration for each local regression. This can cause inaccurate estimations of the ß coefficients, which creates a higher variance of the error terms in the model (Brunsdon et al., 1999). Due to these potential issues, it is important to also assess the spatial autocorrelation that is produced in the error terms before and after modeling. Prior to modeling a test for global spatial autocorrelation can be completed. The global spatial autocorrelation used the Moran's I (Spatial Analyst Toolbox) analysis. The formula for this analysis can be seen below in Equation 8.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \qquad \textbf{Equation 8}$$

The Moran's I index is calculated based on a number of spatial units in the index (N) determined by the number of nearest neighbours. The spatial units are determined based on the sum of a spatial weights matrix (w_{ij}), the variable of interest at each individual point (X_i), and the mean of all X variables (Xbar). The conceptualization of spatial relationships was set as fixed distance band and the distance method was set as Euclidian distance. The output for this analysis is a report, as the analysis is a global analysis, it produces one value for the entire area.

Local method of spatial autocorrelation detection such as Getis and Ord's G statistic (Ord and Getis, 1995) can be completed after the GWR is completed to assess the spatial autocorrelation of the residuals. Local indicators of spatial association (LISA) were

used to determine the local spatial autocorrelation; a cluster analysis utilizing Getis and Ord's G*. The formula for this analysis can be seen below in Equation 9.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \qquad \textbf{Equation 9}$$

In Equation 9, the G statistic is calculated based on the sum of the weighted distances between attributes at two different points (x_i and x_j). This is completed for all spatial variables (n), added together, and then divided by summed values without the weighting to produce the G statistic. These tests can give some indication as to whether there are any spatially autocorrelated variables and where the spatial autocorrelation occurs. Resultantly, identified areas of spatial autocorrelation can also indicate where the statistics of the model may be more deficient due to the violation of regression assumptions (Fotheringham et al., 2002).

Another area where error could occur in a GWR (Charlton et al., 2009) is that the parameters that are run in the local regressions are decided at the global level instead of at the local level. Due to this, it is possible that some of the variables decided at the global level hold no statistical significance at the local level or variables excluded at the global level are actually significant at the local level.

Applications of Spatial Regressions in Environmental Modeling

GWR is a common method being used in environmental modeling due to the non-stationary processes of environmental. Additionally a GWR can be useful for policy as it allows interventions to be tailored to a focused area where it is most needed or will be most beneficial (Auchincloss et al., 2012). Below examples within the literature will be highlighted.

Due to variability in pollution distribution, a GWR is well suited to such studies. Mennis and Jordan (2005) used a multi variable GWR to assess environmental equality based on air pollutants and various socioeconomic factors. For the GWR model, a Gaussian model was selected. Although the authors experimented with different bandwidths, selection was ultimately data led, as this method produced the lowest cross validation score. They noted that smaller bandwidths were usually indicative of higher

spatial variability that occurred in the independent variables. Additionally, they pointed out that it is important to ensure that bandwidth is not smaller than the size of the local geographic regions being studied. These authors highlight the usefulness of combining the technique of GWR with chloropleth mapping to reveal the area of non-stationarity of the model. However, the authors refer to the GWR technique as an exploratory tool used to show relations of non-stationarity. Using these methods, they were able to demonstrate areas where the air pollution had a greater influence on poverty, for example. The local R^2 values ranged from 0.05 to 0.97. While this article made a clear distinction between spatial autocorrelation and non-stationarity, spatial autocorrelation in the GWR model was not discussed.

Hu et al., (2012) used a geographically weighted regression to estimate ground level particulate matter smaller than 2.5 micrometers (PM 2.5) concentrations as PM 2.5 likely emitted from anthropogenic sources has a demonstrated correlation with heart related illnesses. First, a Pearson's correlation analysis was performed to ensure there was no multicolinearity among the independent variables. Variables with high multicolinearity (>0.7) were removed from the model. This study used an adaptive bandwidth to accommodate for the uneven distribution among the data points in study area. To choose the best model of the GWR, the Akaike Information Criterion (AIC) was used. The model with the lowest AIC value was accepted as the best model. The GWR model was compared to an OLS regression and it was demonstrated that GWR produced a better r-squared value and had the largest relative accuracy after an accuracy assessment of the predicted values was completed. By modeling PM 2.5 using a GWR, this study demonstrated that meteorological variables that influence the distribution of PM 2.5 vary over space. This article addressed issues of spatial autocorrelation produced by the GWR model by completing a test for spatial autocorrelation among the residuals. This test demonstrated that were was no significant autocorrelation among the residuals. This confirmed that the GWR was able to incorporate any of the error potentially associated with spatial autocorrelation into the parameters of the GWR.

Another area where spatial regressions are commonly used is in ecological modeling due to the spatial autocorrelation and spatial non-stationarity within the environmental data. Kupfer and Farris (2007) evaluated the use of an OLS global regression model and the use of a GWR to predict patterns of montane ponderosa basal areas in Sguaro National Park, Arizona. In the model, environmental variables were included such as elements of topography (slope, elevation, steepness, and aspect), as well as frequency of wildfire in each area and vegetation. The authors noticed that between the aspatial OLS regression and GWR, there were changes in the both the significance of variables and the direction of the relationship. This gives evidence that there are local controls, which act upon patterns of montane ponderosa basal areas that are different from what is seen in the global model. They conclude by saying the GWR model is a superior method, as it is able to provide information on the fine scale relationships that the global model used over large areas missed. The authors found that overall the GWR model was able to reduce the residuals, reduce the spatial autocorrelation of the residuals, as well as create a model that was able to show the variation is relationships at a local level.

It is also possible to integrate non-parametric data sources with a GWR utilizing a geographically weighted generalized linear model (GWGLM). Erener and Düzgün (2010) compare both aspatial and spatial methods of regression modeling to determine what method is best when assessing the relationship between landslide occurrence (binary variable) and various environment factors such as topography, geological parameters, land cover, and triggering factors such as rainfall. The methods compared were a logistic regression, a spatial regression, and a geographically weighted regression. The data was prepared from various sources and data types; remotely sensed imagery (Landsat TM) was used for the topological variables, methods of interpolation were utilized to achieve a continuous surface of rainfall values and GIS layers. After the environmental variables were prepared, they were then converted from vector data to a grid that had 30x30m cells for the analysis. Once the regressions were completed, the probability was reclassified into low, medium, and high categories using natural breaks. The authors

conclude that the spatial autoregressive model has better predictive power than the non-spatial logistic model. The GWR was used on both the spatial and non-spatial regression to assess how the variable coefficients differ at the local level. Overall, the authors found that the combined spatial methods enhanced the overall predictability of the environmental landslide susceptibility modeling.

Although the method of a geographically weighted regression does present with some problems, the above studies demonstrates the ability of this method to mitigate spatial non-stationarity. This method is most appropriate for datasets that have little spatial autocorrelation but exhibit non-stationary processes. In order to create models with minimal variance, it is essential that the processes that occur in spatial data are corrected to decrease the variance in model. Many studies, which utilized environmental data, favour the use of a GWR due to its ability to handle non-stationary data. The studies above discussed the different methods that were used for model selection, and kernel bandwidth selection. As well, these studies draw attention to two other attractive GWR applications for visualization using geographic information systems (GIS) and for policy strategy.

Chapter Three: Methodology

Research Design

Within geography, there tends to be a divide between physical geography and human geography and, although contained within the same discipline, very different methods are used in these two sub disciplines (Pitman, 2005). Physical geography, like many other hard sciences, typically uses quantitative methodology. Physical geography also views the environment as a large system comprised of complex relationships (Egner and Elverfeldt, 2008). This approach to physical geography is viewed as deductive and confirmatory. The well-accepted research design of science uses methods of classical statistics to validate models. This takes a traditional deductive hypothetical reasoning approach to research (Haines-Young and Petch, 1986). Conversely, human geography typically uses qualitative methodology. This approach is more inductive and exploratory in nature (Johnson and Christensen, 2008; Pitman, 2005).

The field of geographic information systems (GIS) has grown and is a method that many geographers are starting to utilize in research. Traditionally GIS comes from geographic information science. Coming from a deductive epistemology the methodology in GIS is traditionally also rooted in positivism, as there is heavy reliance on calculations, mathematical formulas, and statistical models. Traditionally, this falls into a reductionist framework, which is the philosophical position that believes a complex system to be nothing but the sum of its parts. However, one of the issues with this philosophical position is that model systems in nature can be very complex, for example, well water contamination. Modeling in a reductionist manor, often is not able to capture the complex relationships that are seen within system (Wilson and Poore, 2009).

Ludwig von Bertalanffy first introduced systems theory in the 1940's. Contrary to traditional reductionism in systems theory, a system can be defined as "a set of interconnected parts which function as a complex whole" (Philips, 1992:195). Typically, environmental systems are open as energy is transient, and outputs of one system can feed into the input of another system (Egner and Elverfeldt, 2008). However, one of the issues when using systems theory especially in a GIS, is that decisions must be made on

which variables to include in the system that one is trying to represent. Geographers use general systems theory in order to make representations of the complex world within a GIS. Specifically, systems theory is an interdisciplinary method that uses mathematical models to organize, and describe complex environments. Due to the abilities of this methodology and the ability to use this method within a GIS, the design of this research will be based in systems theory. Particularly what makes this method well suited for a GIS is that every part of the system is seen as an information system and GIS allows for the easy integration of all these information systems (inputs) into the larger system (Goodchild, 2004). Within systems theory, the system is always greater than the sum of its parts. Due to the complex interrelationships of the system, this method attempts to look at the system as a whole, rather than just one aspect of the system. However, with saying that, systems can be modeled as part of subsystems, where each part of the system can be placed within other systems. Here the output from one system will be the input to another system (Strahler, 1980).

Area of Study

The study area is located in Southern Alberta. Specifically, the area of interest is the Calgary Health Zone, which is depicted below in Figure 1 This zone contains one of Alberta's two metropolitan centers, the city of Calgary, as well as smaller municipalities such as Canmore, Banff, and High River. As of 2011, this health zone had a population just over 1.4 million people, which is the most populous health zone in Alberta. The Spray River, Elbow River, and Bow River are the major rivers in the Calgary Health Zone. The study area is approximately 205 km in length and approximately 206 km in width. The study area is just under 40 000 km^2.

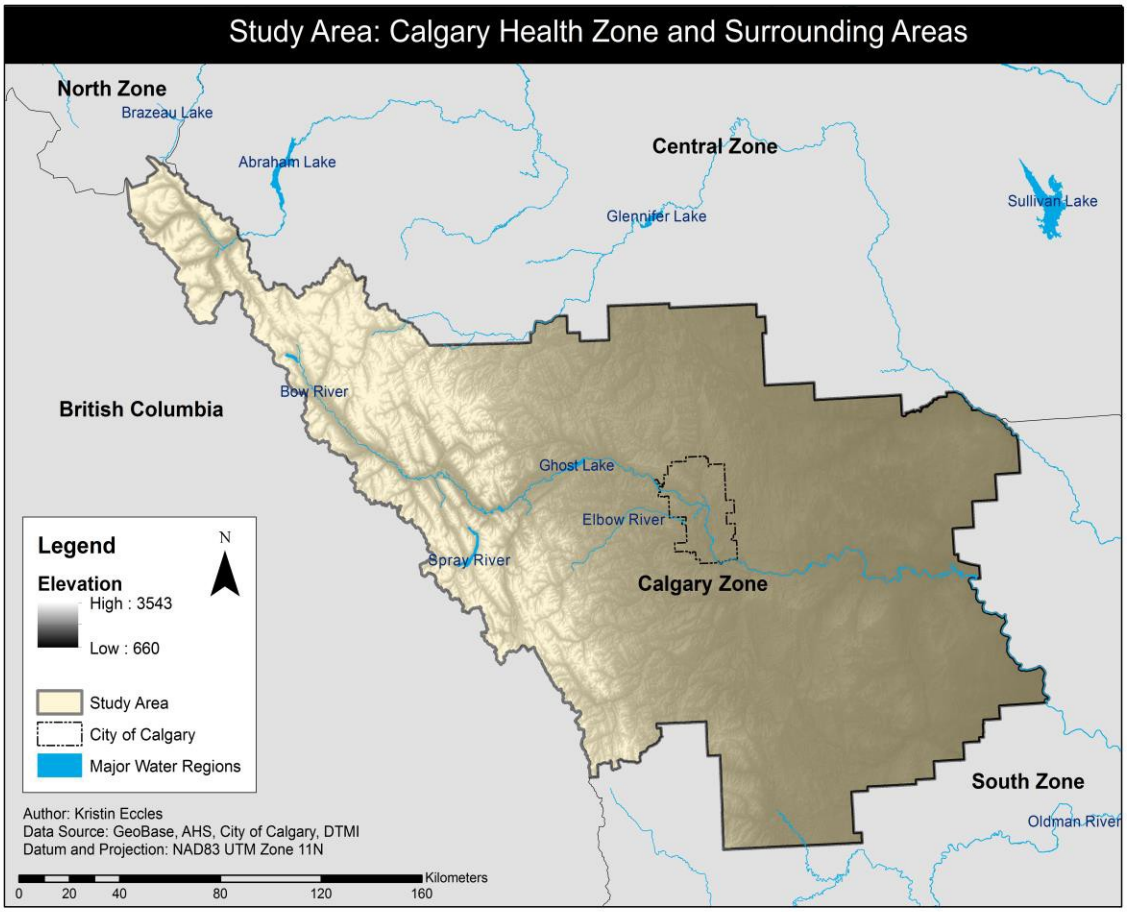

Figure 1. Map of study area, the Calgary Health Zone

Data

Historical Data

Results of private drinking water well sample for total coliforms and Escherichia coli (*E. coli*) were obtained dating back to June 19th, 2005. This data was obtained from Alberta's Provincial Laboratory for Public Health (ProvLab). The total coliforms and *E. coli* results for untreated drinking water wells from private rural residences are based on voluntary samples submitted to ProvLab. The method of analysis the laboratory uses is a Colilert Enzyme Substrate. This produces a binary positive or negative result for both total coliforms and *E. coli*.

Dependent Variables: Coliform Data

For the regression modeling the start and end date of samples that was used are June 19th and September 30th, 2013. The resolution of this data is reported at the quarter

section (1600m2). A summary of the ProvLab sample results can be seen below in Table 2. Although there are 1266 sample results, only 839 (66.3%) of the tested samples are geolocated and are usable in a geographic information system (GIS).

Table 2. Summary of Public Untreated Drinking Water Well Results

Result	Total Samples	Percentage	Geocoded Samples	Percentage
$TC^{-}EC^{-}$	856	67.5%	540	64.4%
$TC^{+}EC^{+}$	92	7.3%	65	7.7%
$TC^{+-}EC^{-}$	320	25.2%	234	27.9%
Total	1268	100%	850	100%

When preparing data set for analysis, first the meridian, range, township, section, and quarter section values were truncated into a unique quarter section identification number (PID) that would be used to join the quarter section shapefile to the ProvLab data. However, there were 79 sets of duplicate records where two or more samples had been reported in the same quarter section. This occurs since many quarter sections have more than one water well located within it. In order to create a join between the quarter section shapefile and the data, the data had to be reclassified so the attribute ID was unique. This involved summing the number of positive sample results occurring at all wells and the total number of samples taken at all wells in each quarter section. This data was all recorded to the unique PID of each quarter section. This method aggregated the individual results of each well up to a quarter section.

After the data was reclassified, there were 530 quarter sections with data available to be used in the analysis. However, only 470 samples could be joined to the unique PID quarter section file in ArcGIS 10.2 (ArcGIS © ver 10.2, Environmental Systems Research Institute, Redlands, CA, USA). This is likely due to reporting errors in the land location of the water wells when the homeowners were submitting samples.

Independent Variables: Environmental Data

The environmental variables that were used in this analysis were collected from various sources. Table 3 provides a summary of the data files, the source, and the datum and projection of all data files that were obtained for use. The majority of the files were obtained in the format of a shapefile (.sph). Shapefiles are the file type used in ESRI's

ArcGIS. The datum of the shapefile refers to which reference method is used to model the shape of the earth, also known as the reference ellipsoid. The projection of the shapefile refers to the coordinate system used to represent the three dimensional earth on a two dimensional surface. It can be seen below that the datum and projection for most shapefiles are Geographic Coordinate System (GCS) 1980 North American 1983 using a North American Datum (NAD) 1983.

The water variables used were obtained through DMTI Spatial. Only polygons and polyline shapefiles were obtained for analysis. The obtained layers included major water regions, minor water regions, minor water lines, and minor intermittent/ slough regions. The metadata indicates the accuracy of these layers are National Topographic Data Base (NTDB) standard and are accurate down to sub-meter. These layers were last updated 2012.

The digital elevation model (DEM) was obtained from DMTI Spatial. The layer metadata indicates the DEM is based on the Canadian National Topographic System (NTS) and has a resolution of 30 meter pixels (1:50,000). The raster layer is current as of 2011.

Aquifer depth, obtained from IHS Energy, however no metadata is available for this layer, and therefore the resolution and age of the layer are unknown.

Hydraulic Connectivity was obtained from Alberta Agriculture and Rural Development (AGRASID) and as indicated by layer metadata, has a resolution of 1:100,000. Changes to the data were last made in 2001, but has been regularly maintained since then.

Through the Flood Hazard Identification Program, Alberta Environment and Sustainable Resource Development identifies areas of the province that are in flood hazard zones. The flood hazard mapping identifies areas that would experience flooding based on a 100 year flood. The delineation is broken down into 4 different classification. Floodway is considered part of the river channel. In this region, the water has the fastest flow of all zones, the water is the deepest, and resultantly, the most destructive. The flood fringe is an area that occurs outside of the floodway. During a flood this region will experience overland flood, however, the water moves at a slower pace, and usually less

than one meter deep. Within the dataset flood fringe and overland flooding are classified separately, as overland flooding is considered to be a special case of flood fringe by the developers. Additionally, regions that are under review are available, but were not included in this research. These layers were last updated August 9th, 2013 (AESRD, 2013).

The farming variables obtained from the 2006 Agricultural Census and is reported by Soil Land Survey of Canada Polygons for both hectares of farm land, as well as number of farms within each polygon.

The land cover data was obtained from Alberta Biodiversity Monitoring Institute and is reported as square kilometers of each land category based on the classification of remotely sensed images. Included in this classification is land annually cultivated for crops, orchards, vineyards, bare agricultural soil, and grazing land for cattle. The accuracy of the land cover classification was 75% when there were 11 classifications and 88% when the classes were reduced into 5 general groups (ABMI, 2010).

More information on all layers used in the analysis can be found below in Table 3.

Table 3. Data Sources, Datum and Projection

Shapefile	Source	Datum	Projection
Well Test Results [Data File]	Alberta Health Services- ProvLab	N/A	N/A
Calgary Health Zone Boundary [shp]	Alberta Health Services	GCS North American 1983	UTM (10TM)
Township and Range Quarter Sections [shp]	AltaLIS Through SANDS*	GCS North American 1983	Unprojected
City of Calgary City Boundaries [shp]	City of Calgary Data Catalogue	WGS 1894	UTM (3TM)
Water (Including minor and major streams, rivers, and water bodies) [shp]	DMTI 2013 though SANDS	GCS WGS 1984	NAD 1983
DEM [raster]	GeoBase	GCS North	Unprojected
Aquifer Depth [shp]	IHS Through SANDS	GCS North	Unprojected
Hydraulic Conductivity [shp]	Alberta Agriculture and Rural Development (AGRASID)	GCS North American 1983	UTM (10TM)
Precipitation June 2013 [data file]	Environment Canada	Geographic Coordinate	Lat, Long

Flood Impacted Zone [shp]	Environment Canada	GCS North American 1983	Unprojected
Abandoned Wells [Data File]	Alberta Energy Regulators	N/A	N/A
Farms [shp]	2006 Agricultural Census	GCS North American 1983	10TM AEP Forest
Land Cover Data [shp]	Alberta Biodiversity Monitoring Institute	GCS North American 1983	10TM AEP Forest
Population and Dwelling Density	2011 Census	GCS North American 1983	Unprojected
*SANDS= Spatial and Numeric Data Services			

Data Preprocessing

First, all files were converted to from the existing datum and projection to NAD 1983 CSRS UTM Zone 11N. This was completed using the Projections and Transformations tool (Data Management Toolbox). Vector files were converted using the Feature Project tool and raster layers were converted using the Raster Convert Coordinate Notation tool. To avoid reference errors, it is important that all used layers have the same datum and projection. The files were then added to a geodatabase using appropriate feature datasets and corresponding feature classes. The advantage to using a geodatabase is that it ensures that all layers have the same datum and projection. It also enforces data integrity of all layers in that no data is accidently deleted or altered, which would affect the integrity of the layer and resultantly the accuracy of the analysis (Longley et al., 2010).

Variable Creation

Interpolated Rainfall Surface

Data was obtained from Environment Canada in the form of a CSV file. This file was then imported in ArcGIS 10.2. To produce the most accurate interpolated surface, weather station values for the entire province were used; there are 242 stations recording total rainfall (millimeters) in June 2013. All weather stations that had a record of zero millimeters of rain were removed to ensure that only active stations were included in the interpolation; there were 229 remaining weather stations with 24 of the 229 weather stations residing within the Calgary Health Zone. The latitude and longitude coordinates of each weather station (geographic coordinate system WGS 1984) were converted to NAD

1983 CSRS UTM Zone 11N. Then, the interpolation method of universal Kriging (Geostatistical Wizard) was used to remove any surface trends present to allow for the most accurate predictions. This created a continuous surface of estimated rainfall values for the month of June between the measured locations.

Elevation Derived Variables

Digital Elevation Models (DEM), which are specialized raster tiles used for elevation, were downloaded from GeoBase. Tiles were extracted from regions 83A, 83B, 83C, 82N, 82O, 82P, 82I, and 82J. The tiles where then converted from individual tiles into one continuous, mosaic image using the Mosaic tool (Data Management Toolbox). The pixels located between different images were averaged to create a seamless mosaic image. After, the mosaic image was clipped to match the size of the Calgary Health Zone by using the Extract by Mask Tool (Spatial Analyst), the raster was then converted to NAD 1983 CSRS UTM Zone 11N. Next, the slope of the Calgary Health Zone was derived using the DEM as the input in the Slope Calculator (3D Analyst toolbox). Lastly, using the DEM of the Calgary Health Zone, the aspect was created. The aspect, also known as the directionality of the surface, has pixel values of negative one to indicate a flat surface and values between zero and 360 degrees to indicate the direction of the slope. This variable was calculated using the Aspect tool (Spatial Analyst toolbox).

Data Files

All of the remaining data files had the legal land description (LLD), based on the Alberta Township Survey, included. The text data files were imported into Excel 2013 (Microsoft Excel© ver 2013, Microsoft Corporation, Redmond, WA, USA). In excel the LLD were rearranged to match the PID (meridian, township, range, section, and quarter). The quarter section value was then convert to the numerical equivalents; NE=80, NW=70, SE= 50, SW=60. These values were then concatenated to form the unique PID. This allowed the information to be joined to the quarter section shapefile. The abandoned wells data from the Alberta Energy Regulators also included information pertaining to lot, block, and plan numbers. This is finer resolution than the quarter section. To have the same resolution of all other data, the LLD was aggregated up to the quarter section. The

aggregation process can be seen below in Figure 2. After the unique PID was created, the files could then be added into ArcGIS and joined based on the PID. The shapefile was then exported to a new shapefile for use in the analysis.

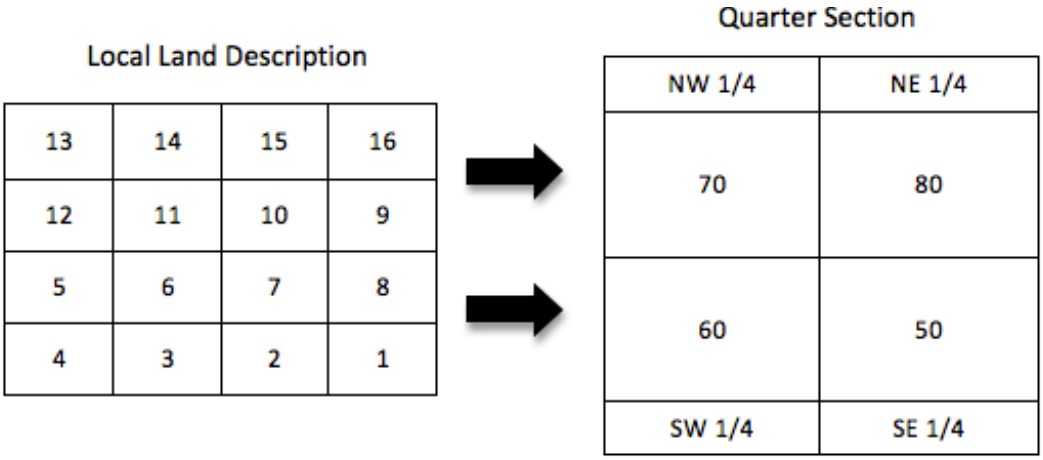

Figure 2. Aggregation of Dominion Land Survey Land Locations.

Data Extraction

First, the quarter sections with data were converted from polygon features to point features using the Feature to Point (Data Management Toolbox) which produced a center point (centroid) in each quarter section. Then, buffers (Analysis Toolbox) were created around these points. The first buffer had a radius of 400m. This first buffer has the same diameter of a quarter section. Other buffers used have radii of 800 meters, 1600 meters, 3200 meters, and 6400 meters.

The geoprocessing intersect tool was used to find all variable elements within each quarter section. For example, the quarter section layer was intersected with minor water lines (a polyline feature). As some quarter sections included different segments of minor water lines, the sum of the length of these segments was calculated using the summarize function. This tool added the total length of the minor water lines in each quarter section, which was organized by PID of each quarter section (e.g. 5070281980). If the layer that was being summarized was a polygon, instead of having the total length calculated, the total area was calculated (e.g. Major Water Regions). The summarized data was saved as a text file.

To find the distance from the quarter section to the nearest feature, for example, distance to the nearest major waterway, the Near tool (Analysis Toolbox) was used to calculate this distance. The calculated distances were saved as a text file. As the intersect tool is not compatible with raster data sets, data from the DEM, slope, aspect, and rainfall raster data sets were extracted using the Feature to Point (Data Management Toolbox) tool. This tool placed a point in the center of each quarter section so raster data could be extracted to this point. There were a total of 60, 811 quarter sections in the Calgary Health Zone. Then, the Extract Values to Points (Spatial Analyst Toolbox) tool where the underlying feature values were extracted to the point file. This was then saved as a text file.

After data extraction, all text files were added to ArcGIS 10.2 and using the centroid shapefile that contained the PID for all quarter sections, all of the text files were joined using the joining feature of the unique PID. All records were kept (opposed to only keeping matching records). This enabled null values, which would later be converted to zero values, to be placed in the column where there was no data. No data would occur if the feature in question, for example major water regions, were not present within the buffered area (e.g. 400 meter buffer). The newly created table containing all created and calculated variables was then exported to a text file, which was then imported into Excel 2013 for further analysis. This file serves as the master sheet to be used in the regression modeling.

Historical Analysis Methodology

First, a visual representation of the percentage of positive test results for total coliforms and *E. coli* that occurred during the study period, June 19th to September 30th for the previous eight years (2005-2013) were graphed in Excel 2013. To determine if the percentage of positive samples that occurred during the study period, was statistically different from the same time period in previous years, a Wilcoxon Rank Sum Test (z-test) was used in S+ 8.2 (S+® 8.2, ver 8.2, TIBCO Spotfire, Palo Alto, CA, USA).

This test determines if the median percentage of positive test results that occurs in 2013 for total coliforms, and *E. coli* were greater than in previous years. As well this test

can indicate whether the study period (2013) has a greater number of positive test results than occurred in previous years. For this analysis the one-tailed version of the test was used and was tested at the 95% CI. This analysis was also completed for the number of samples submitted during the study period, compared to the previous eight years.

Descriptive Methodology

Spatial Distribution of Samples

To evaluate both the spatial distribution of the sample results for total coliforms and *E. coli* graphical and cartographic representations were created. Histograms were created in Excel 2013 to represent the number of sample processed each day between June 19th and September 30th. The first histogram comprises all samples. The next two histograms produced were total coliforms samples, and total *E. coli* samples. These histograms separated the positive and negative sample results. These three graphs demonstrated the temporal distribution of samples during the study period.

The pattern of all sample locations was also evaluated to assess the distribution of the locations. In ArcGIS 10.2 the nearest neighbour was calculated using Average Nearest Neighbour (Spatial Statistics Toolbox). As the input feature, the centroids of sample quarter section were used. For this analysis, the null hypothesis for the nearest neighbour analysis is that the sample locations are randomly distributed. This null hypothesis is either accepted or rejected based on the statistical significance of the z-score. If the z-score is not statistically significant, then the null was accepted, and could be concluded that samples are not clustered.

The spatial distributions of sample results were evaluated both globally and locally. In these analyses, total coliforms, and *E. coli* were evaluated separately. Spatial autocorrelation was calculated in ArcGIS 10.2 utilizing Moran's I, and Getis and Ord's G (Spatial Analyst Toolbox) was used to demonstrate hot spots (local clusters of positive test results), and cold spots (local clusters of negative test results).

For this analysis, the conceptualization of spatial relationships (weighting function) was set as fixed distance band and the distance method was set as Euclidian distance as

well. The output for this analysis is a cartographic representation depicting the clusters of positive (hot spot) and negative (cold spot) test results.

Analytical Methodology

Analytical Regression Modeling

The analytical regression modeling utilized the methods typically seen in Land Use Regression (LUR) modeling and other environmentally based regression models (Sliva and Dudley Williams, 2001; Wheeler et al., 2008), where buffers were created around the centroids of each quarter section that had sample results. The area or length of each feature contained within each buffer was calculated utilized in the methods outlined in the Data Extraction section.

To determine which buffered value would be included in the regression, a correlation analysis was performed between the number of positive test results and each of the independent variables, for the 400m, 800m, 1600, 3200m, 6400m buffers, as well as the distance to each feature calculations. For the correlation analysis, Spearman's correlation was used, as the data is non-parametric. For each variable, only the variable with the highest correlation coefficient was chosen to be included in the initial regression. There were 17 independent variables in this initial stage of the regression.

Total coliforms and *E. coli* were outcomes in two separate regression models. For both models, an offset of total number of samples submitted in each quarter section to account for the uneven sampling distribution. The offset was included as the log function of the number of samples to include the offset as an additive feature of the regression equation. Due to the assumed non-normal distribution, count nature of the data, and many zero values, a Poisson regression model was decided upon for both the total coliforms and the *E. coli* model. Additionally, due to the assumption of equidispersion among the dependent variables, a dispersion test was used (AER library) to determine if the Poisson regression was either overdispersed or underdispersed, and thus in violation of the assumption of equidispersion. If the data was underdispersed then a Conway–Maxwell Poisson distribution was used and if the data was over dispersed then a quasipoisson regression was used.

After the initial Poisson general linear regression model was run R 3.0.0 (R: A Language and Environment for Statistical Computing©, ver 3.0.0, R Core Team, Vienna, Austria). The correlation of the coefficients computed to ensure there was no multicolinearity between the independent variables. The correlation of the coefficients was used rather than a traditional correlation analysis due to the count nature of the data and the assumptions that surround the traditional correlation analysis methods, the assumption of normality. A conservative threshold was used and all variables that had a correlation of coefficient greater than 0.65 were removed. After highly cross-correlated variables were removed, a backwards stepwise method was use to select variables for the final model. This was conducted by removing the variables that were not significant at the 95% confidence interval, proceeding from least significant. After a variable was removed, the model was re- run to determine which variable would be removed next. This process was completed until all variables were statistically significant. Near the end of the variable selection methods, the threshold of a 95% CI was modified to be more flexible and included variables that have been deemed important in the literature review but did not meet the 95% CI criteria (Richardson et al., 2009).

To determine the quality of the model the residual deviance was used, as the quasi- poison regression model does not compute an AIC. As well, McFadden's pseudo R^2 was used to indicate the goodness-of-fit, which can be seen in Equation 10.

$$R^2_{McF} = 1 - \left(\frac{\ln(L_M)}{\ln(L_0)} \right) \qquad \textbf{Equation 10}$$

The heteroscedasticity of the residuals of the aspatial models were judged based on the Breusch-Pagan test (lmtest package). If there is heteroscedasticity of the residuals, this may be indicative of spatial clustering. However, this test statistic is based on a normal distribution, in which this data does not have, therefore this test is cautiously used. If the Breusch-Pagan test indicates heteroscedastic errors, then spatial methods are introduced. The method of a geographically weighted regression (GWR) was used if there was statistically significant spatial clustering.

Utilizing a Moran's I test, the spatial autocorrelation of the residuals at the global level was also assessed. This was to ensure that the assumptions of the model are not

being violated inducing added variance. If there was statistically significant spatial autocorrelation as detected by the Moran's I statistic then a spatial autoregression (SAR) was used to reduce the variance.

Risk Map Generation

To produce the risk map, the equation of the line that was created based on the results of the regression model, was used to predict the environmental risk of contamination in all quarter sections where results were not available. Using the significant variables from the regression models, first, the values for the variables extracted in ArcGIS 10.2. These values for all variables were then attached to each quarter section based on the PID. The values were extracted to an Excel spreadsheet. The extracted values and the beta values were entered into the equation seen below in Equation 11.

$$\widehat{Y} = exp^{(B_1X_1+B_2X_2\ldots\ldots B_iX_i)} \qquad \textbf{Equation 11}$$

In this equation, the predicted values in quarter section ($\hat{Y}$) are derived from multiplying the beta (β) values with the value for the variable (X). Then all the variables are added together and raised to the power of the natural logarithm (e=2.718). The predicted value for each quarter section was calculated in Excel 2013, and then is imported into ArcGIS 10.2. This unique PID was used to join the spreadsheet with the quarter section shapefile. To create the symbology of low medium and high risk, quartile breaks of the collected sample values are used; the first quartile is low environmental risk, the second and third quartiles are medium environmental risk, and the fourth quartile is high environmental risk.

Chapter Four: Results

Historical Comparison

Figure 3 demonstrates the number of samples submitted historically during the same study period of this research. This graph demonstrates that there were more samples submitted in 2013 than in the years 2006 through 2012. However, the number of samples seen in 2005 surpasses the number seen in 2013. Statistically, what is visually seen on the graph is confirmed using Wilcoxon Rank Sum test. The results from this test can be seen below in Table 4.

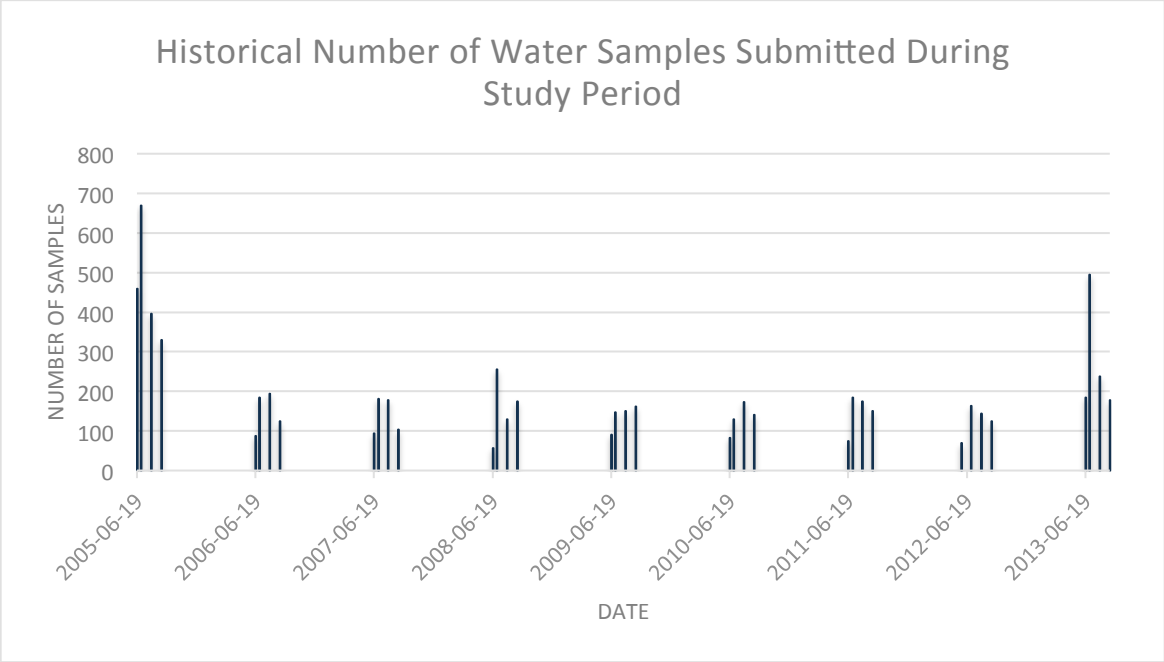

Figure 3. Number of water samples submitted for analysis during the study period (June 19th- September 30th) 2005-2013.

This analysis reveals that for the years 2006-2012, the median number of samples that were analyzed is statistically greater than 2013. However, in the year 2005, the number of samples that were analyzed is statistically greater than the number analyzed in 2013 ($z = 3.888$, p-value = 0.0001).

Table 4. Number of Samples: One tailed (greater than) Wilcoxon Rank Sum Test (Z-test)			
Year	**Mean**	**T-value**	**P-value**
2005	19.82	-3.88	0.999
2006	5.664	3.522	0.0003
2007	5.346	3.803	0.0001
2008	5.923	3.298	0.0006
2009	5.087	3.940	0.0001
2010	5.041	3.942	0.0001
2011	5.606	3.726	0.0003
2012	4.74	4.322	0

Figure 4 compares the percentage of positive total coliforms results and the percentage of positive *E. coli* results for each day in the study period (June 19th to September 30th) inclusive of the years 2005 to 2013. This graph demonstrates that there is an increased number or drinking water wells that tested positive for total coliforms in 2013, seen in dark blue, compared to the seven previous years. Also notable is the high number of positive results in 2005. A similar pattern can be seen for the percentage of

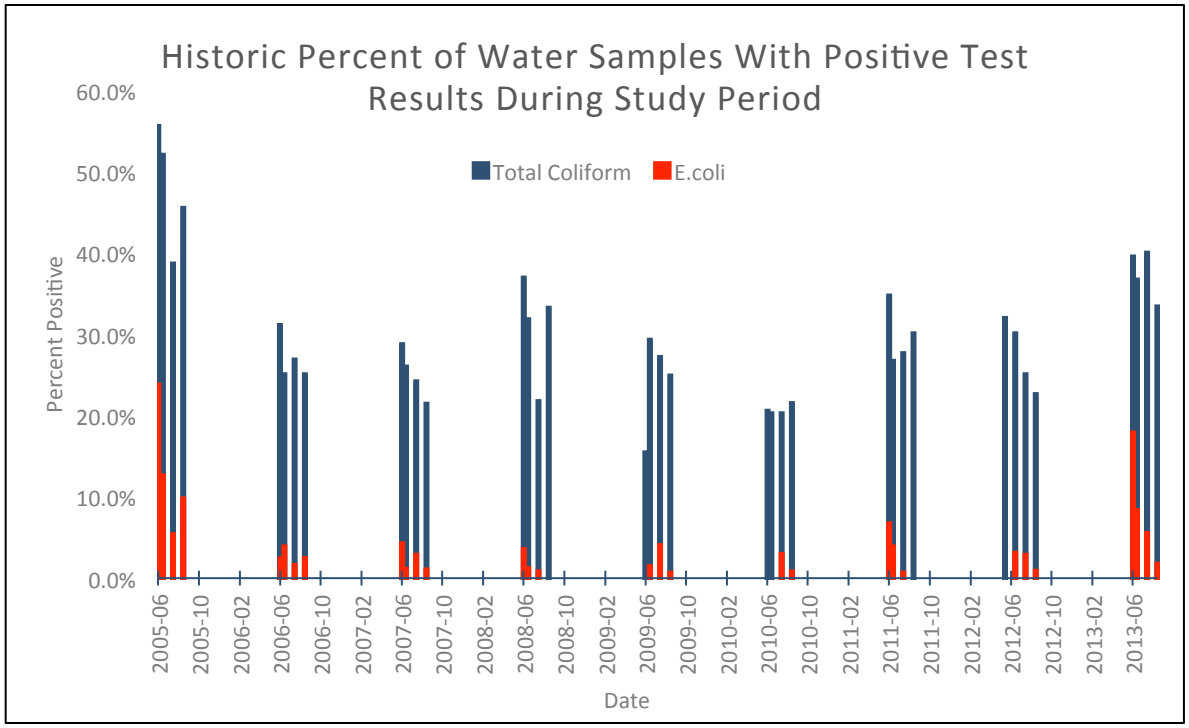

Figure 4. Percentage of Positive total coliform and E.coli results during the study period (June 19th- September 30th, 2005-2013

positive *E. coli* samples, seen in red, which is also larger in 2013 than it is in the previous seven years. Similar to total coliforms results, in 2005, the *E. coli* results also have a larger number of positive results compared to 2013, and the rest of the data set.

Statistically, the Wilcoxon Rank Sum Test (z-test) indicates that the percent of positive total coliforms results that occurs in 2013 is statistically different from years 2006-2012. Table 5 shows the results of the Wilcoxon Rank Sum Test (z-test) through a comparison of the median during the same time period (June 19th to September 30th) for the previous eight years. In this table, the p-values indicate that the median in 2013 is statistically larger than the median of all years, except 2005. In 2005, the p-value indicates that the percentage of positive total coliforms results seen in 2013 are not statistically larger than the results from 2005. However, the median of the total coliforms results in 2013 is not smaller than the percentage of total coliforms results seen in 2005 (z = 1.393, p-value = 0.08) at the 95% CI. Therefore it is concluded that the percentage of positive total coliforms results in 2005 is not statistically different from the results in 2013.

Table 5. Total Coliform: One tailed (greater than) Wilcoxon Rank Sum Test (z-test) results

Year	**Z-score**	**P-value**
2005	-1.395	0.919
2006	4.484	0
2007	3.812	0.0001
2008	3.391	0.0002
2009	4.652	0
2010	5.379	0
2011	3.760	0.0001
2012	3.3472	0.0004

Using a Wilcoxon Rank Sum Test (z-test), Table 6 demonstrates that the median percentage of positive *E. coli* test results that occurred during the study period in 2013 is statistically larger than the median percentage of positive *E. coli* results that occurred 2006-2012. However, this pattern did not hold true for the percentage of positive *E. coli* results that occurred in 2005. As indicated by the p-value, the percentage of positive *E. coli* results that occurred during the study period in 2013 was not statistically larger than the number that occurred in 2005. Rather, the median percentage that occurred in 2005 is

statistically larger (z = 3.434, p-value = 0.0003), than the median percentage of positive *E. coli* results that occurred in 2013.

Table 6. *E.coli*: One tailed (greater than) Wilcoxon Rank Sum Test (z-test) results		
Year	**Z-score**	**P-value**
2005	-3.46	0.997
2006	4.416	0
2007	4.525	0
2008	5.372	0
2009	5.004	0
2010	5.458	0
2011	4.591	0
2012	4.812	0

Descriptive Results

The purpose of the descriptive section is to obtain a better understanding of the total coliforms and *E. coli* test results for the study time period of this study, June 19th, 2013 to September 30th, 2013. Below in Figure 5 is a graph of samples by date. The highest volume of samples was seen in the first few weeks after the flood. This was then

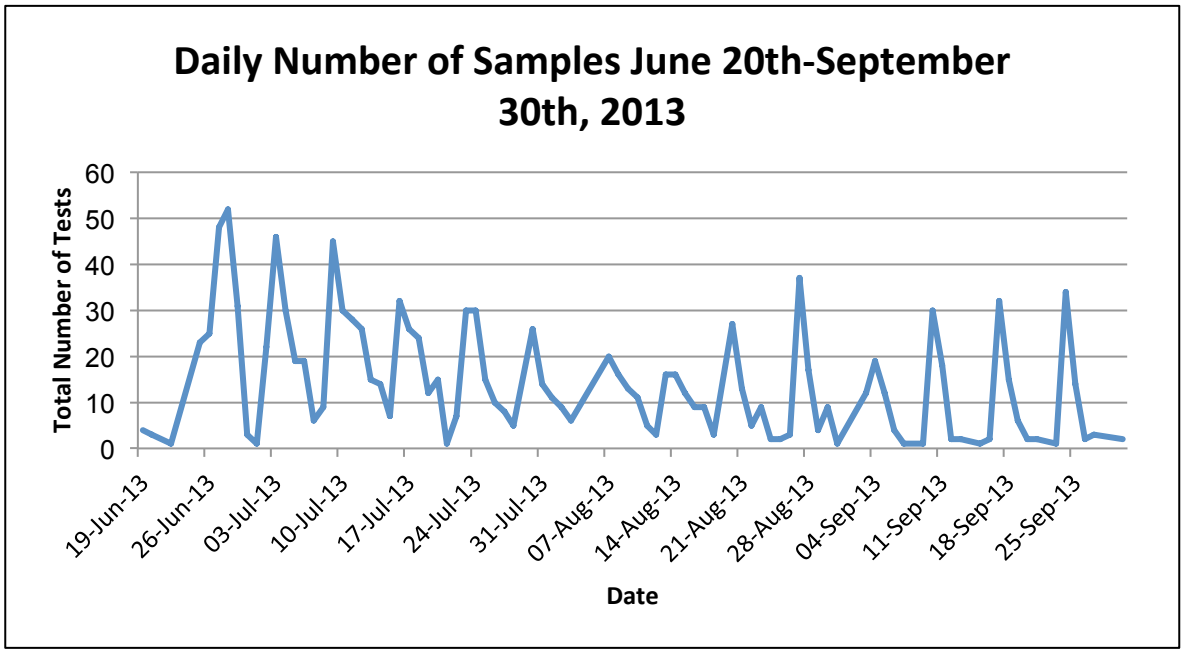

Figure 5. Total number of samples submitted each day for the study period.

followed by a decline of sample intake until August 29th where there was another small spike in sample intake. These volumes remained steady through the rest of the study period. The highest number of sample intake of 52 samples occurred June 28th, 2013.

After the samples are analyzed, the samples can then be separated by date as well as test result. Figure 6 and Figure 7 show the number samples that test positive and negative total coliforms and *E. coli* respectively. The graph in Figure 6, shows the number of positive test results for total coliforms seen in the weeks following the flood. After about two weeks the number of positive samples decreased from the high 20s to around 10 positive samples a day. The negative samples follows the same pattern that the positive samples take, for the most part having peaks and falls in the same places.

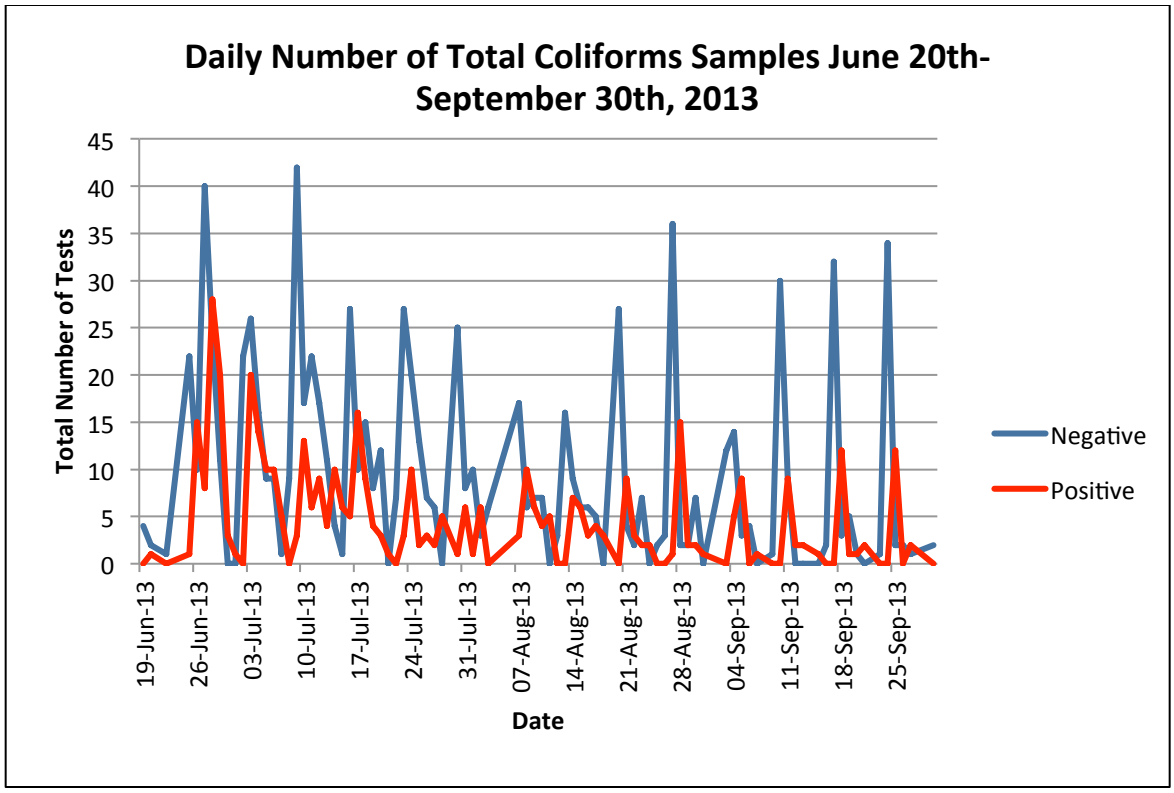

Figure 6. Total coliforms test results by date.

In Figure 7 the *E. coli* results during the study period. Generally, there is fewer than five samples per day that test positive through the study period; however, it is notable that in the two weeks following the flood, there were a higher number of samples testing positive. During this time, the highest number of positive samples was 12. The equal

distance between peaks suggests that there is a week effect with sample submission, where there are routinely more samples take and submitted at a certain time of the week.

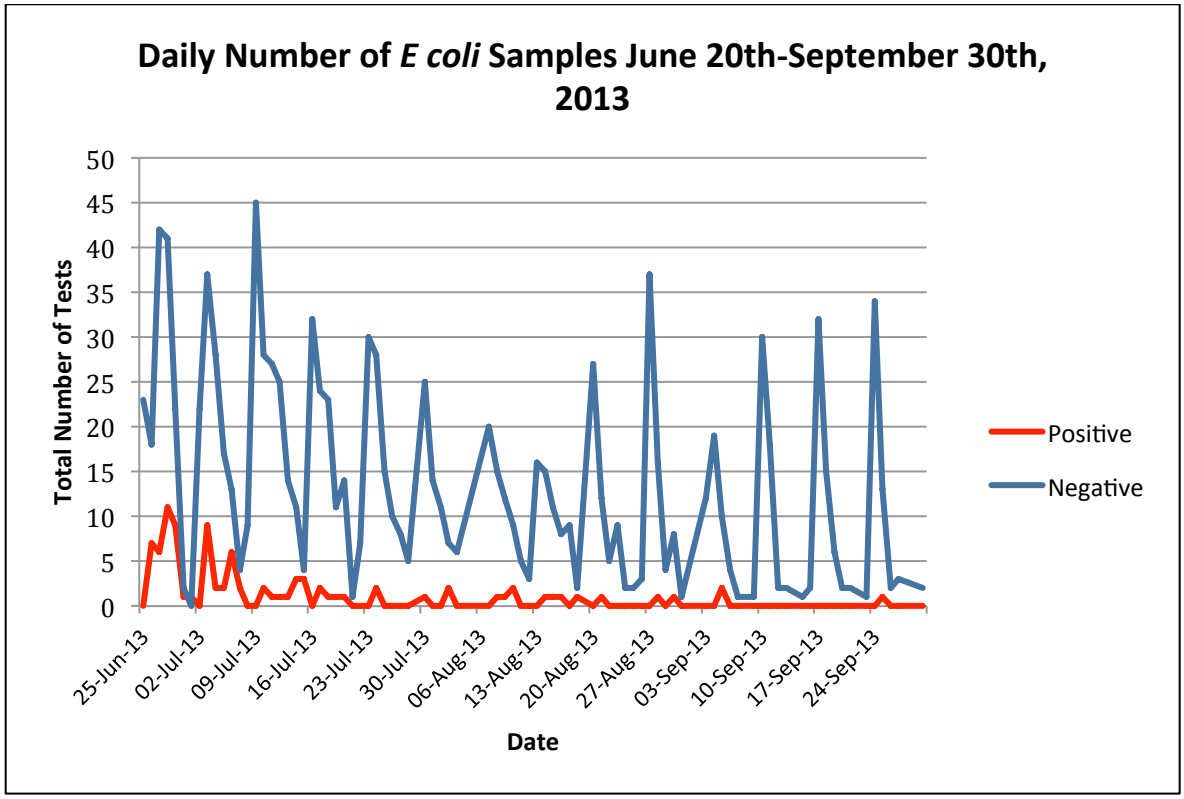

Figure 7. Total *E.coli* test result by date.

The nearest neighbour analysis, which assessed the distribution of all samples, indicated the observed mean distance was 4150.93 meters. This means that on average the sample locations are slightly more than 4 km apart. With a nearest neighbour ratio value of 0.43 and a z=-14.82, which was statistically significant (p=0.00), with confidence the null hypothesis can be rejected. Therefore, based on this analysis, it is concluded that the sample locations in this study are clustered. There is less than a one percent chance that the clustered pattern exhibited by the sample locations could be due to random chance.

Global Moran's I for total coliforms based on an euclidian fixed band distance of 428.83 km produced a non-statistically significant global Moran's I (index=0.000263, p=0.55). Similar results were produced for the *E. coli* test results. The global Moran's I, which was based on an euclidian fixed band distance of 428.83 km, similarly produced a non- statistically significant global Moran's I (index=0.0036, p=0.144). Obtaining non-

statistically significant results indicates that at the global level, there is no spatial autocorrelation Getis and Ord's G* was used to assess the local indicators of spatial association (LISA). The two maps below show the locations of the clusters, where the green and blue colours indicate clusters of negative test results, and the yellow, orange, and red indicate clusters of positive test results. The colour corresponds to the confidence interval the quarter section belongs to for each clustered group. The light grey indicates the statistically insignificant quarter sections. The first of the two maps, Figure 8, shows that there is a cluster of negative total coliforms sample results in the central northern region of Calgary Health Zone. As well, there is a cluster of positive total coliforms results in the southern central region of the Calgary Health Zone. For total coliforms negative the cluster is much larger than the positive cluster.

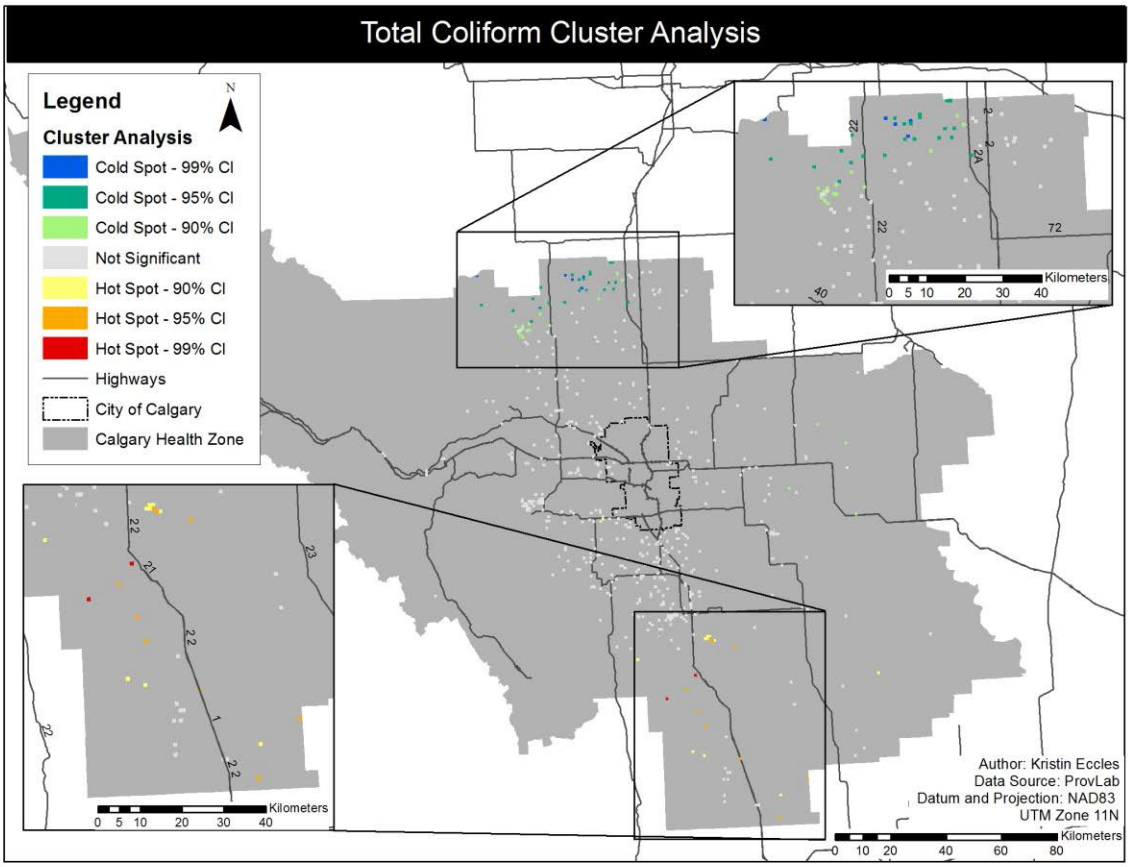

Figure 8. Getis and Ord's G* cluster analysis for total coliform.

Figure 9, is the second of the two cluster analysis maps depicting *E. coli* and it is notably very different from the total coliforms cluster map. In the map below, there are

only positive clusters of the *E. coli*. The first is located west of the Calgary city limits north of Highway 1, the TransCanada Highway. The cluster extends to a portion south of Highway 1; however, this part of the cluster is less statistically significant than the region north of the highway. There is also a small cluster of positive *E. coli* test results in the southeastern part of the Calgary Health Zone.

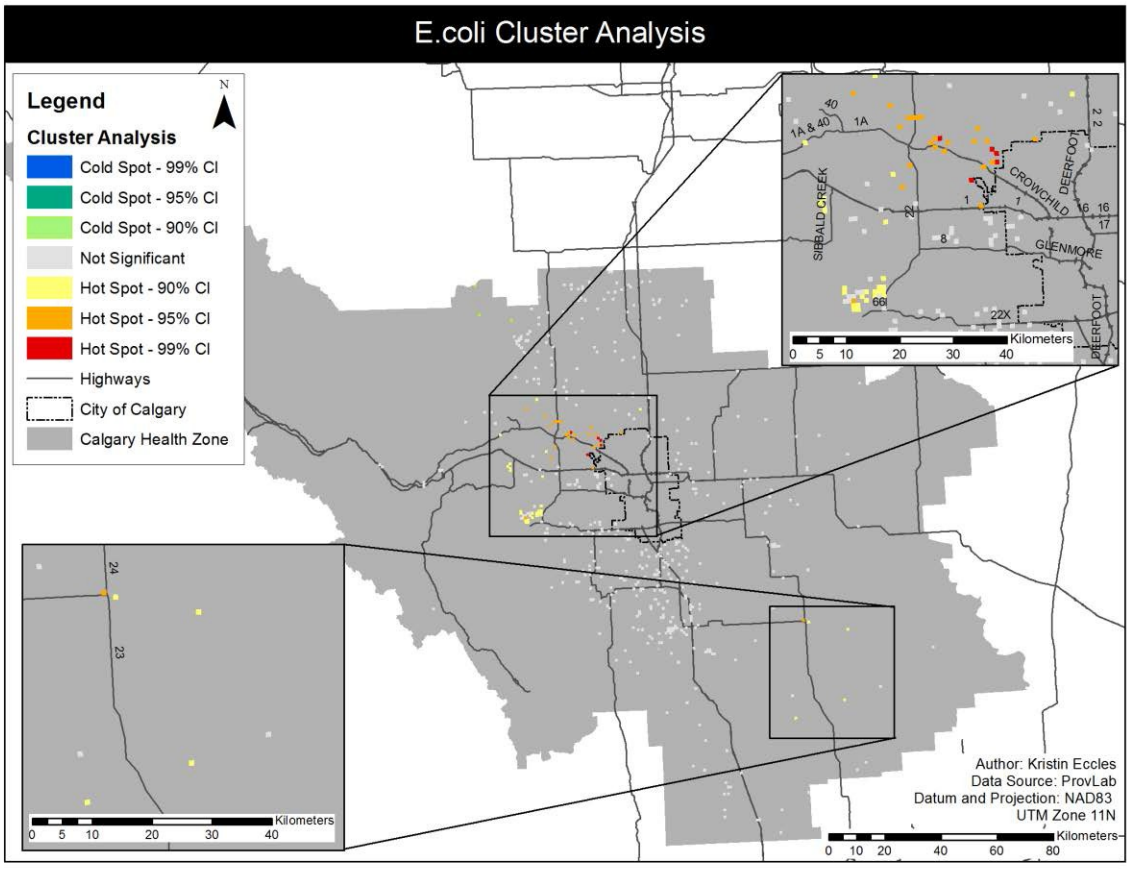

Figure 9. Getis and Ord's G* cluster analysis for *E.coli.*

Analytical Results

Environmental Variables

The environmental variables that were created are demonstrated in Figures 13-16. All of these created layers serve as input for the regression model. The modeled surface rainfall surface was completed using Kriging interpolation in Figure 10. During the month of June, 2013, the total amount of rainfall ranges from 29.9 mm of rain in the eastern part

of the health region to 265.2 mm of rain west of the city of Calgary extending into the Rocky Mountain Range.

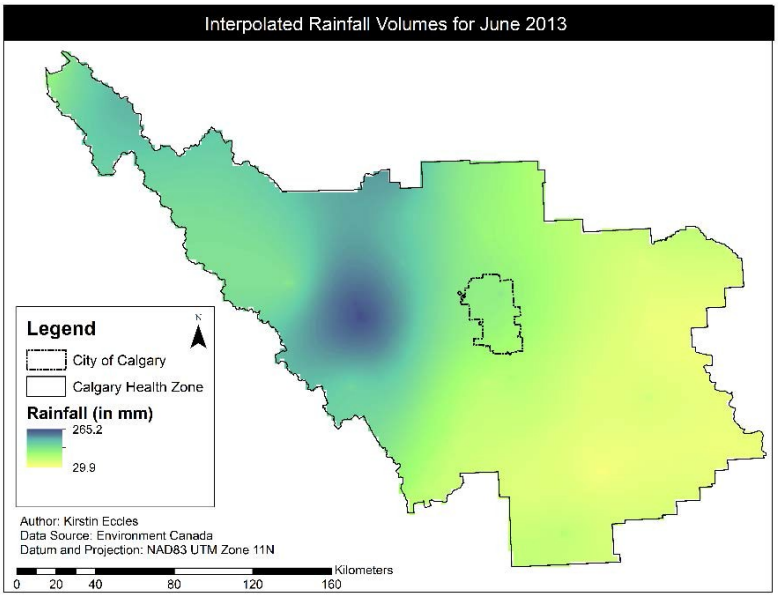

Figure 10. **Interpolated rainfall values for June 2013 using Kriging.**

The mosaic of the individual DEM tiles produces an elevation map of the Calgary Health Zone, which can be seen in Figure 11. In this area, the highest elevation is 3543

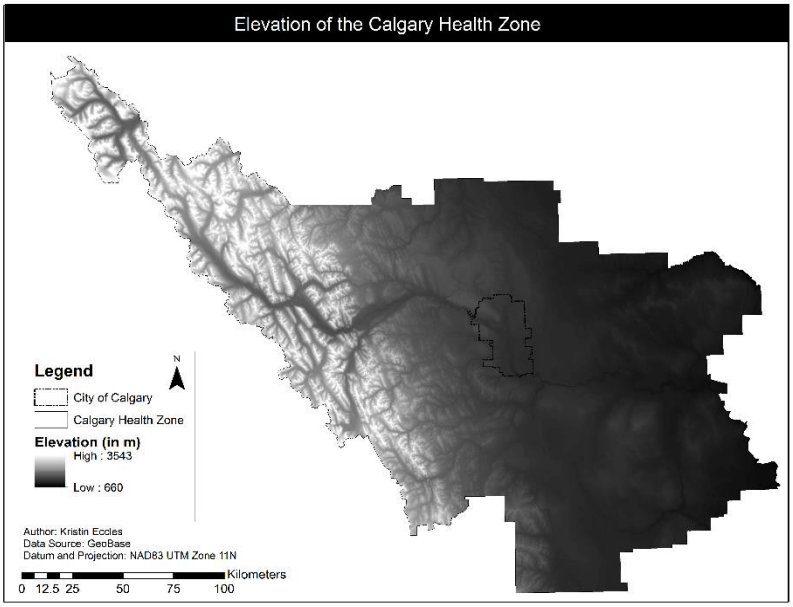

Figure 11. Elevation map extracted from the DEM

meters located in the Rocky Mountain Range and the lowest elevation is 660m located in the eastern part of the health zone.

Figure 12 is a map of the slope in the Calgary Health Zone. This was derived from the elevation map seen in Figure 11. The greatest slope is 86.8 degrees and is found within the Rocky Mountain Range. The flattest land, which has no slope, is predominantly found in the eastern region of the health zone. Also note it is possible to see the sloped riverbanks of the Bow and Elbow River.

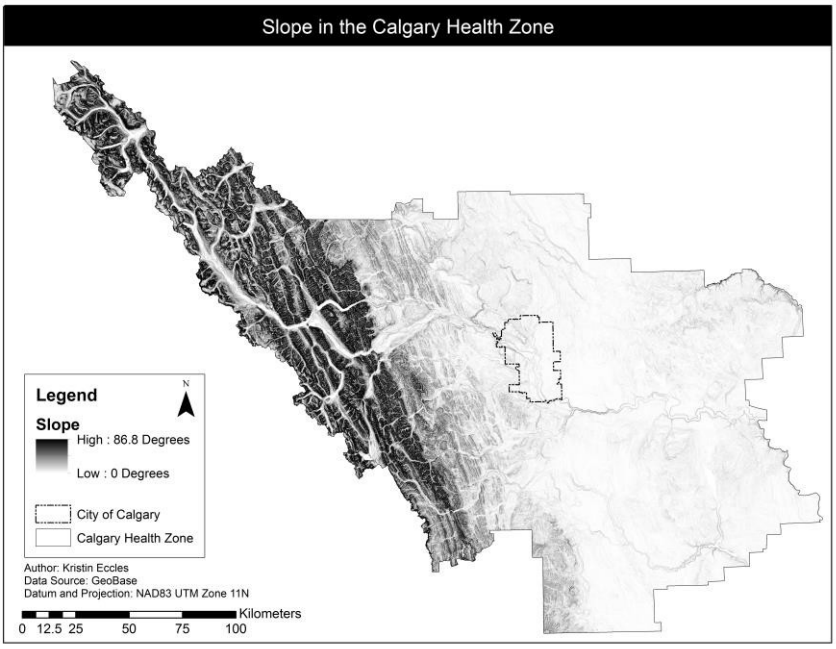

Figure 12. Varying slope present in the study area.

The aspect, also derived from the DEM, can be seen in Figure 13. Aspect refers to the direction that the slope is facing. This is a cyclical variable, which can be interpreted in degree that the slope is facing (0-360°) or by direction the slope is facing (North, Northeast, East, etc.)

From these maps of the study area, it is evident how geographically diverse the study region is with the mountains to the west and prairie to the east. Within the created elevation related values, there is a gradient from West to East.

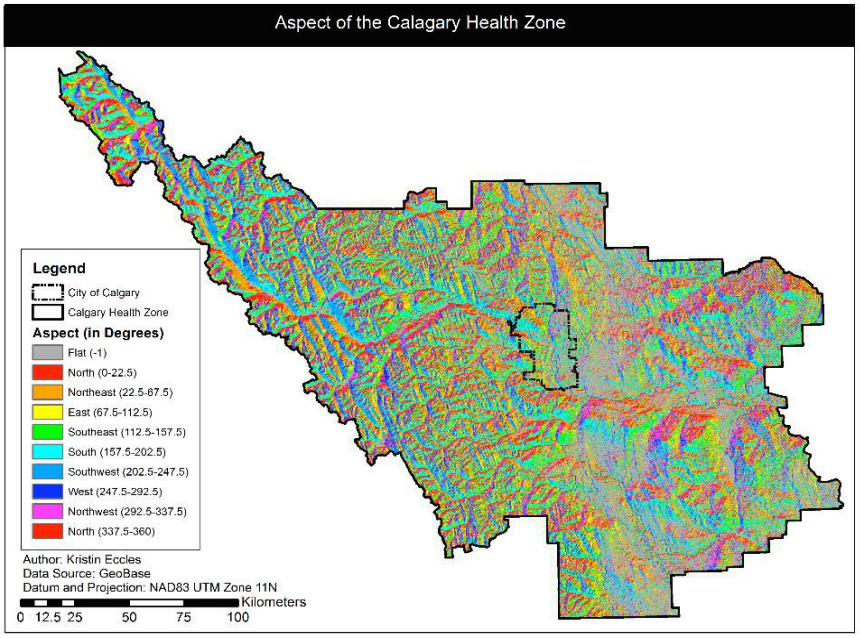

Figure 13. Aspect of the Calgary Health Zone derived from the DEM.

Regression Modeling

The variables selected for initial inclusion after the completion of Spearman's correlation analysis in each regression model can be seen below in Table 7.

Table 7. Independent Variables used in the Regression Modeling			
TC Independent Variables		**EC Independent Variables**	
Variable	**Spearman's Correlation**	**Variable**	**Spearman's Correlation**
Rainfall	0.012	**Rainfall**	0.057
Elevation	0.087	**Elevation**	0.056
Slope	0.019	**Slope**	-0.062
Aspect	-0.05	**Aspect**	0.014
Water Lines 6400M	0.08	**Water Lines 800M**	0.087
Overland Flooding 1600M	0.109	**Overland Flooding 1600M**	0.15
Minor Water 3200M	0.129	**Minor Water 800M**	0.100
Major Water 400M	0.145	**Major Water 400M**	0.313
Intermittent Water 3200M	-0.193	**Intermittent Water 3200M**	-0.139
Floodway 400M	0.229	**Floodway 800M**	0.25
Flood Fringe 400M	0.230	**Flood Fringe 400M**	0.23
Developed Land 400M	0.086	**Developed Land 800M**	0.099
Agricultural Land 800M	-0.148	**Agricultural Land 800M**	-0.131
Forested Land 400M	0.096	**Forested Land 400M**	0.156
Near Abandoned Wells	-0.074	**Near Abandoned Wells**	-0.072
Population Density 400M	0.081	**Population Density 400M**	0.081

Dwelling Density 400M	0.169	**Dwelling Density 400M**	0.169
HA of Farm Land	-0.015	**HA of Farm Land**	-0.06
Number of Farms	0.018	**Number of Farms**	0.03

Based on the variables in Table 7, the regression models for total coliforms and *E. coli* were developed. The script that was used for the regression analysis can be found in Appendix A.

Total Coliform Regression Model

First, the correlation of the coefficients was calculated in a global Poisson GLM. The results of this showed that while rainfall and elevation were highly correlated (-0.84) in the Spearman's correlation matrix, the correlation of the coefficients of these two variables were not highly correlated (-0.51), therefore none of the variables were removed from the regression. However, the correlation of the coefficients for farms and HA of harms were highly correlated (-0.85). Therefore, for the total coliforms regression HA of farms was removed, as number of farms had a slightly higher correlation with the dependent variable. Next, a dispersion test was utilized to ensure the assumption of equidispersion was satisfied. For the total coliforms outcome, this test revealed that the dependent variable was underdispersed. Therefore a modified version of the Poisson regression model was used; the Conway-Maxell Poisson distribution is used specifically for underdispersed data as addressed in the background section. The other assumption of a GLM noted in the background section, were also met as the correct link was chosen, the log link for a Poisson regression, and the variance function is automatically set to match the log link. The aspatial model can be seen in Table 8.

Table 8. Total Coliform Conway-Maxwell Poisson Distribution Regression				
	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**
Intercept	-5.134e-01	2.550e-01	-2.013	0.0441
Rainfall	-2.479e-03	1.528e-03	-1.623	0.1047
X800M_Agri	-2.367e-07	1.172e-07	-2.019	0.0435
Residual Deviance	555.03	**Pseudo R^2**	0.013	
AICc	561.09	**DoF**	467	

In this regression only rainfall received in the month of June and the area of agricultural land within 800M were found to be significant. However, agricultural land within 800M is only significant at the 89% CI. When this variable was removed rainfall also became insignificant (p=0.52). This occurrence in conjunction with the McFadden pseudo R^2 that was 0.0132, indicates that aspatially the environmental variables were not able to the model the occurrence of total coliform positive well water samples. As can be seen in Figure 14, the model predicts that most of the quarter sections will not test positive for total coliform. There are one a few a quarter sections west of Calgary at the start of the Elbow River, and south of Calgary that will have one or more positive tests as predicted by the aspatial Conway-Maxwell Poisson regression.

In this model, only rainfall received in the month of June and the area of agricultural land within 800M were found to be significant. However, agricultural land within 800M is only significant at the 89% CI. When this variable was removed rainfall also became insignificant (p=0.52). Additionally, in the aspatial regression, the McFadden pseudo R^2 (R^2=0.0132) is very low. As can be seen in Figure 14, the model predicts that most of the quarter sections will not test positive for total coliforms. There are only a few a quarter sections west of Calgary at the start of the Elbow River, and south of Calgary that will have one or more positive tests as predicted by the aspatial Conway-Maxwell Poisson regression.

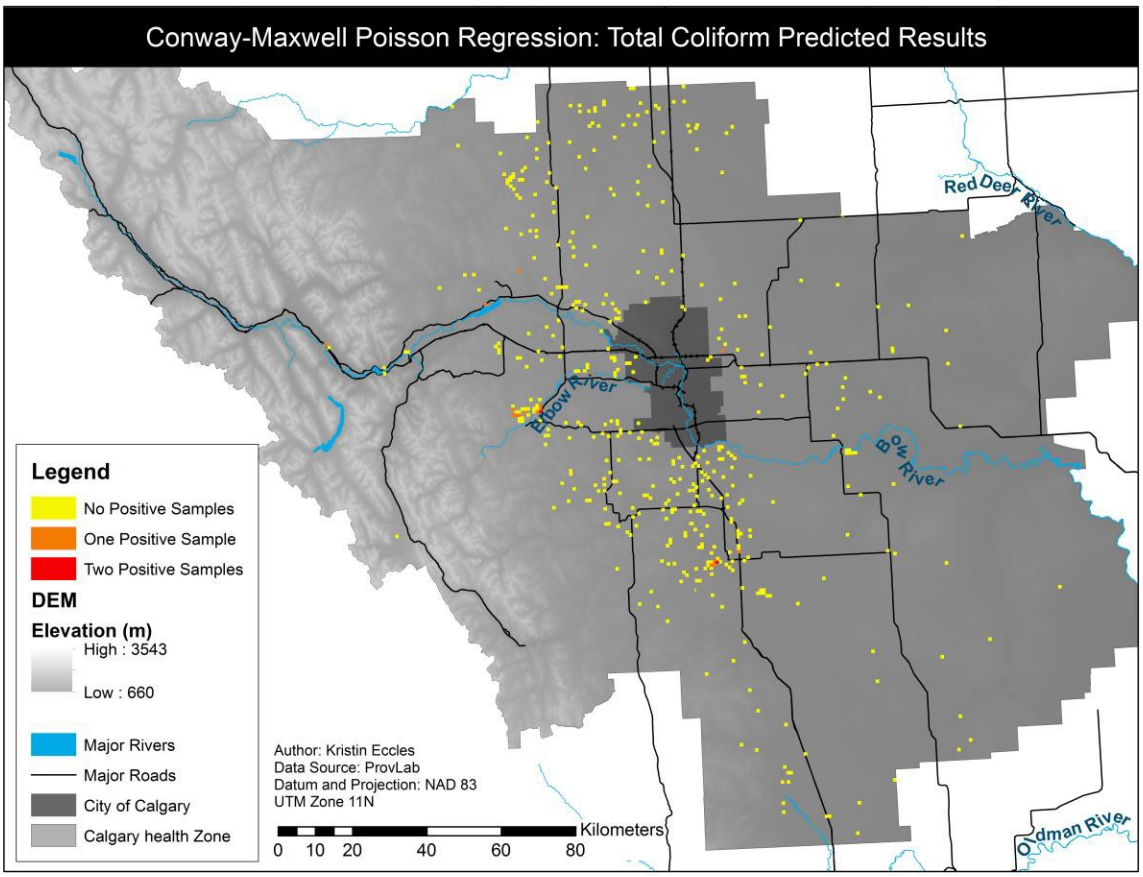

Figure 14. Predicted number of positive water samples for total coliform using a aspatial Poisson GLM.

Figure 15 visually demonstrates the residuals of the aspatial model. The red indicates where the model over predicts the most and the blue represents where the model under predicts the number of positive samples in each quarter section. Here it can be seen that the model over-predicts, more than the model under-predicts. In this map, a cluster of moderate under-predictions can be seen on the west end of the Elbow River. Other than this cluster, no distinct pattern can be seen. The residuals of this aspatial model are not spatially autocorrelated (I=0.003 p=0.516).

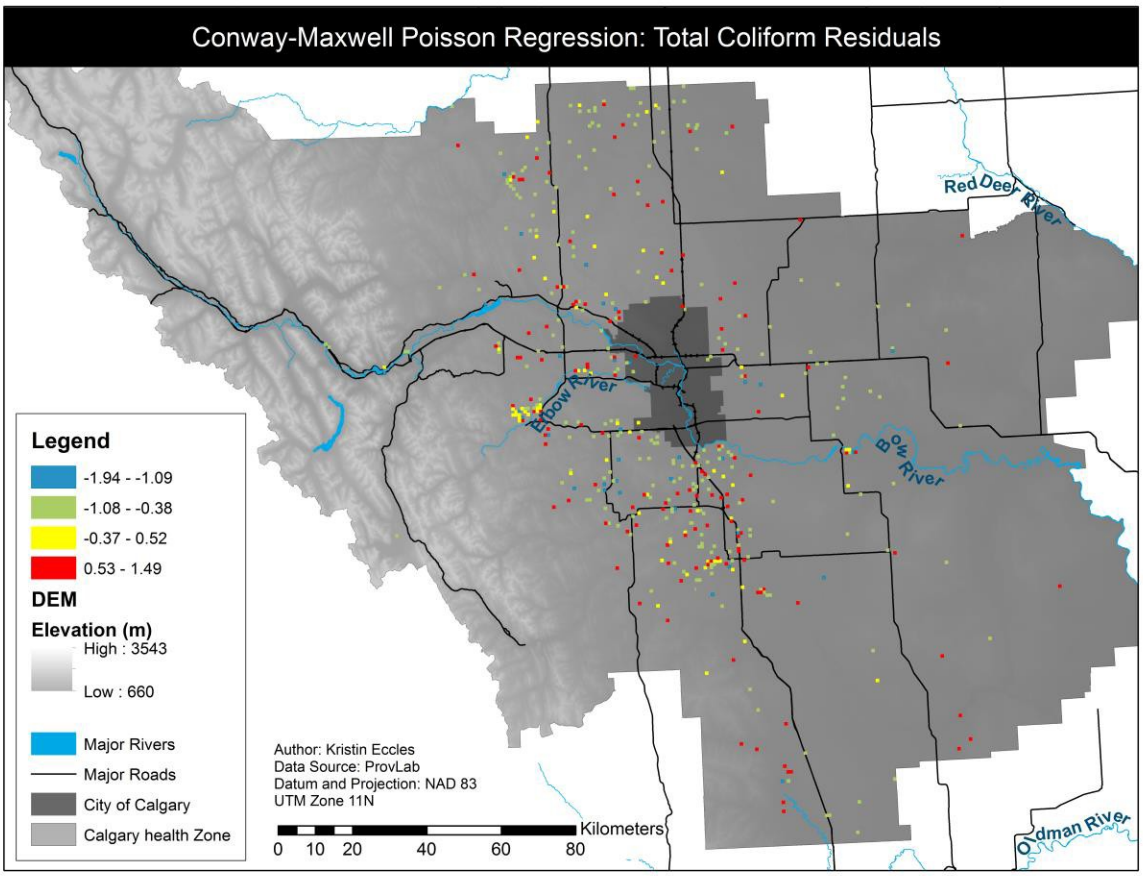

Figure 15. Aspatial GLM total coliform model residuals.

Upon testing the heteroscedasticity of the residuals in this regression, the test revealed that the data does not exhibit non-stationarity (bp=5.33, p=0.06). However, since the significance of the Breusch–Pagan test narrowly missed the 95% confidence interval of significance a geographically weighted regression was completed, this test was originally intended for linear models and was used only as a proxy for this non-linear model.

Table 9. Total Coliform: Spatial Poisson Regression with COM Distribution.						
	Min.	**1st Qu.**	**Median**	**3rd Qu.**	**Max.**	**Range**
Intercept	-7.96E-01	3.28E-01	5.98E-01	1.23E+00	2.87E+00	-3.67E+00
Rainfall	-2.34E-02	-2.65E-03	6.63E-04	3.21E-03	6.47E-03	-2.98E-02
X800M Agri	-7.82E-07	-4.23E-07	-3.05E-07	-1.19E-07	2.17E-07	-9.99E-07
Residual Deviance	337.68		**Quasi-Global R^2**	0.11		
AICc	344.68		**DoF**	467		

Using an optimized fixed Gaussian bandwidth of 24266.59 meters the GWR-Poisson model can be seen above in Table 9. The varying coefficients seen in Table 9 that show the relationship between the dependent and independent variables vary over space. The intercept of the model varies the most, followed by the amount of rain that fell within the month of June 2013, and finally, the amount of agricultural land within 800M. Interestingly, the geographically weighted regression also changed the significance of the agricultural variable in the regression. While in the aspatial global model, the agricultural variable was not significant at the 90% CI, in the geographically weighted model, this variable is significant at the 95% CI in 242 (of 470) of the regressions (quarter sections) and is significant at the 90% CI in all the local regressions. The smallest t-value was -1.78.

The predicted outcome of water well contamination by total coliforms produced by the geographically weighted regression can be seen below in Figure 16. In comparison with the map of the aspatial regression, there are a greater number of quarter sections that this model predicts will have one positive test results for total coliforms. However, this model does not predict that there will be greater than one positive test sample

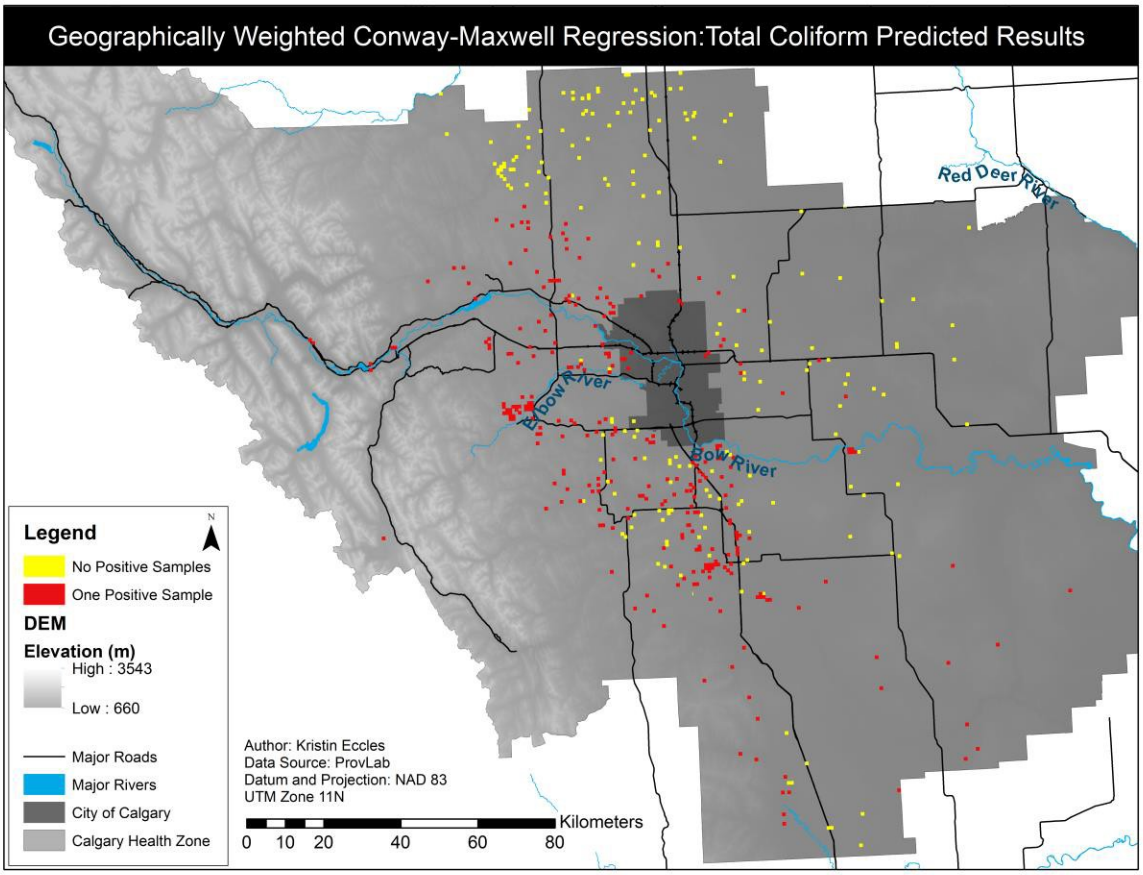

Figure 16. Predicted number of water well samples positive for total coliform using a geographically weighted regression.

The residuals for this model can be seen in Figure 17. Over most of the study region, the predicted values are close to the actual value of number of water samples positive for total coliforms in each quarter section. This is represented by the yellow and green coloured quarter sections. Yellow indicates where the model slightly over-predicted and the green indicates where the model slightly under-predicted. The model under-predicts more than it over-predicts.

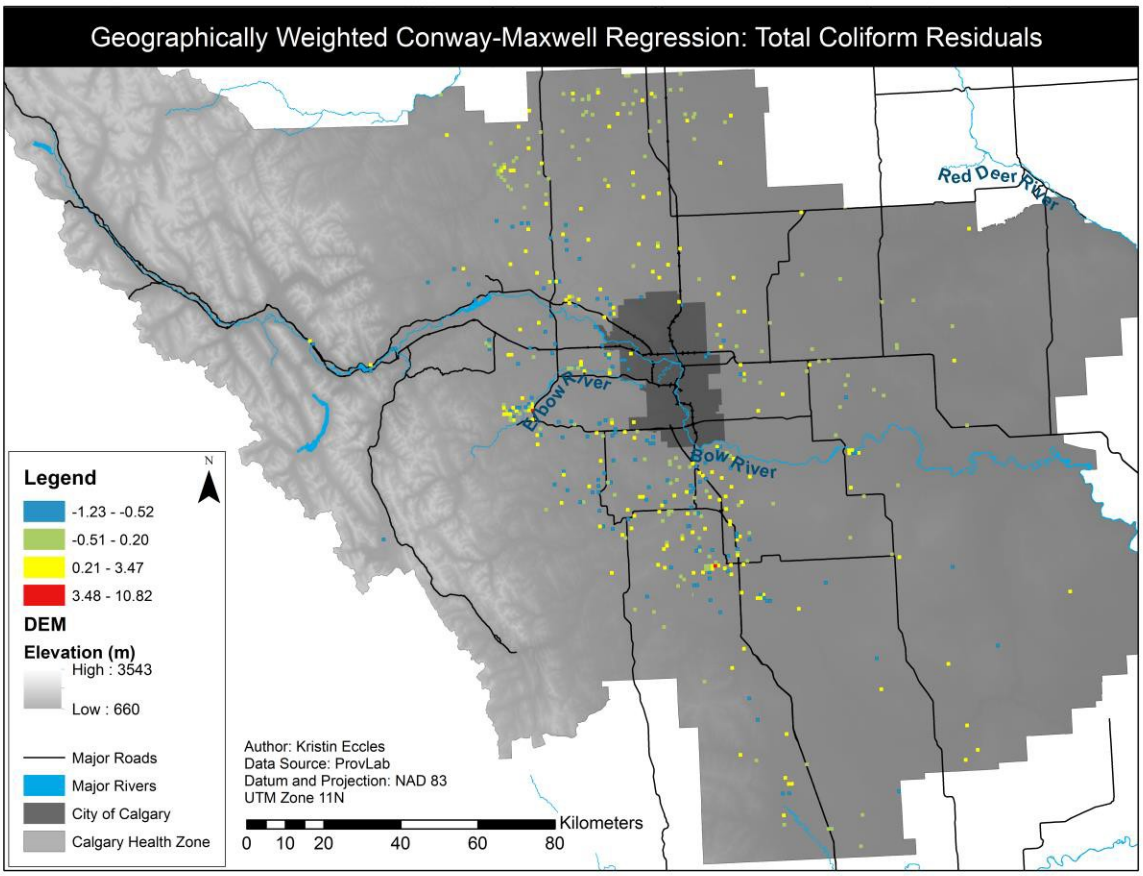

Figure 17. Residuals of the geographically weighted regression for total coliform.

Compared to the aspatial regression, the geographically weighted regression performed better as indicated by the AICc, the deviance of the residuals, and the quasi-global R^2. The deviance of the residuals decreased to 555.03 from 337.68. This smaller deviance in the geographically weighted model indicates that the predicted values deviate less from the measured values. This indicates the geographically weighted model that is more accurate in predicting the correct outcome. Additionally, the AICc also decreased from 561.09 to 344.68 indicating that the geographically weighted model is statistically a better fit than the aspatial model. Lastly, the quasi-global R^2 increases. As this measure is an average of all the local R^2 values, there are local R^2 values that are above and below this average. The lowest local R^2 is 0.018 and the highest value is 0.27.

The variation of local R^2 can be seen in Figure 18. The variation of the local R^2 demonstrates that the highest values occur in the southeast corner of the Calgary Health Zone. In this area the model can explain between 15% and 27% of the variance of the

dependent variable by the independent variables, rainfall and agriculture. The highest local R^2 is represented by quarter sections with the colour red. The central region of the health zone extending from the northern region to the city of Calgary is not able to explain the variance of the dependent variable well. This is represented by the colour yellow in the figure. East of the city of Calgary, the ability to explain the variance is in the mid- range between 0.07 and 0.14. This is indicated by the light and dark orange colour.

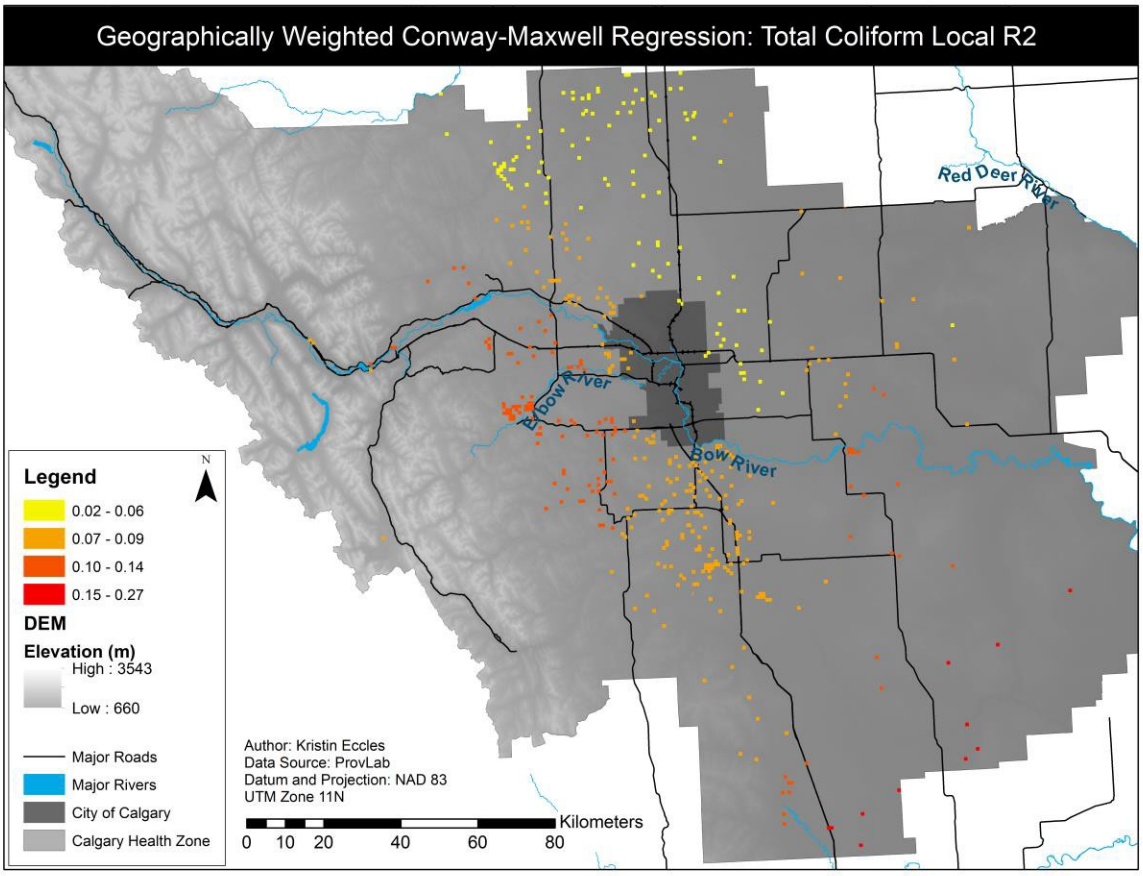

Figure 18. Local R^2 of the geographically weighted Conway-Maxwell regression for total coliform.

The smallest beta values for all variables are negative indicating in some regions, there is a negative relationship between the rainfall in June 2013, and agriculture within 800M and the occurrence of water well samples positive for total coliforms. However at the high end of the beta values, the maximum values are all positive, indicating a positive relationship between the independent variables and the occurrence of water well samples positive for total coliforms. This changing relationship can be seen below in Figure 19. In this figure, the red indicates where the amount of rain that fell in the month of June is

significant in the regression at the 95% CI. The orange indicates quarter sections where rainfall is significant at the 90% CI. Yellow is insignificant. Quarter sections that exhibit a negative relationship between rainfall and total coliforms contamination of wells can be

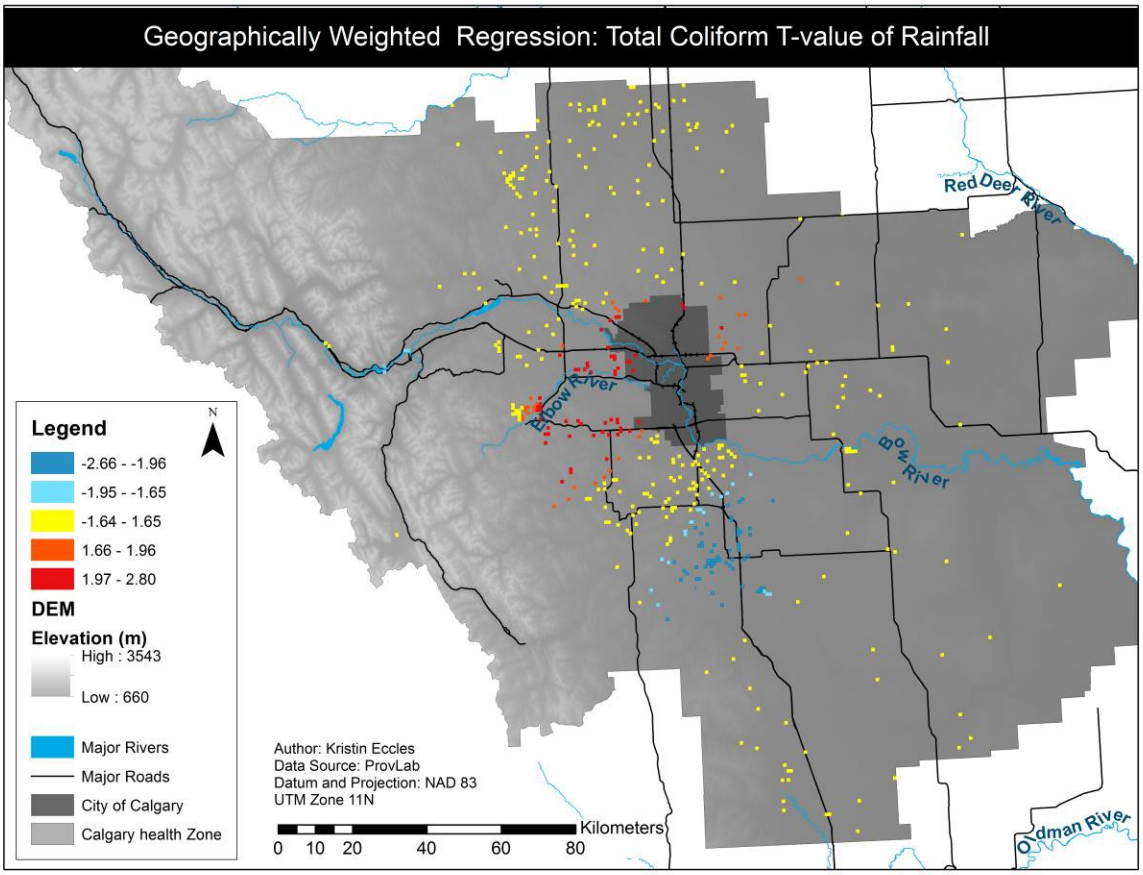

Figure 19. T-value and corresponding significance and relationship of rainfall in the geographically weighted total coliform regression model.

seen in light blue significant at the 90% CI and dark blue which are significant at the 95% CI. The same colour scheme and scale are used in Figure 20 where the t-values and corresponding significance are represented for the other significant variable in the regression, agricultural land within 800m.

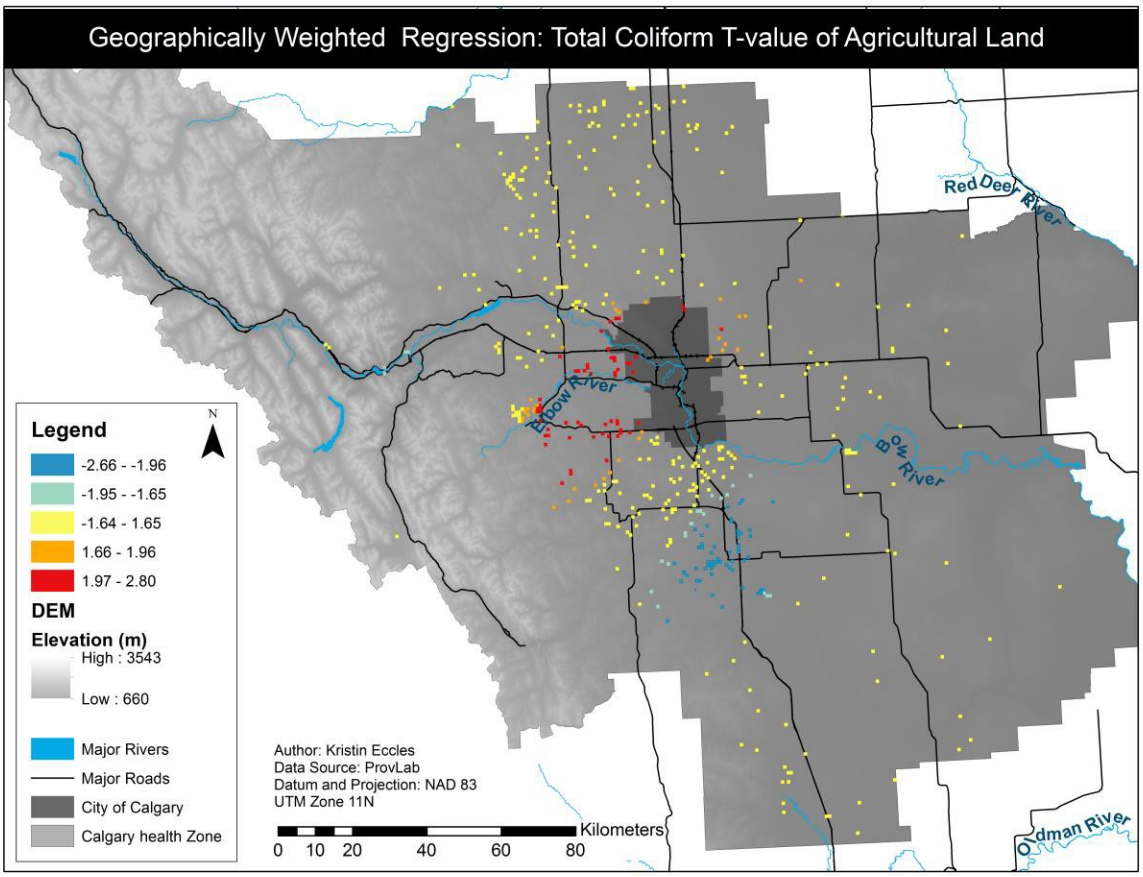

Figure 20. T-value and corresponding significance and relationship of agricultural land in the geographically weighted total coliform regression model.

In both of these figures representing the regression t-values and variable significance it is evidence that west of the city there is a positive relationship between the number of positive total coliforms well water results, the amount of rainfall that fell in June 2013, and the amount of agricultural land. However, south of the city of Calgary, city there is a negative relationship between the number of positive total coliforms well water results, the amount of rainfall that fell in June 2013, and the amount of agricultural land. As a result of the varying, and contradictory relationship as both directions (negative and positive relationship) are statistically significant at the 95% CI within the geographically weighted regression. Therefore, a risk map was not created.

E.coli Regression Model

First, the correlation of the coefficients were calculated. The results of this were the same as the total coliforms regression and showed that when using the Spearman's correlation coefficient, rainfall and elevation were highly correlated (-0.84). However, when the correlation of the coefficients was calculated, the two variables did not have a high enough correlation to warrant one to be removed (-0.51), therefore both the variables remained in the regression. Again, the correlation of the coefficients for farms and HA of harms were highly correlated (-.85), therefore for the total coliforms regression, HA of farms was removed, as number of farms had a slightly higher correlation with the dependent variable.

After the correlation assumption was satisfied, a dispersion test was utilized to ensure the assumption of equidispersion was satisfied. For the dependent variable, *E. coli*, this test revealed that this variable was neither underdispersed nor overdispersed. The dispersion of this data was calculated to be 1.0141, a normal Poisson distribution regression was used. As the link and the variance function are already pre-specified, all other assumptions for using a Poisson GLM were satisfied. The aspatial model for *E. coli* can be seen in Table 10.

Table 10. *E.coli* Poisson Distribution Regression

	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**
Intercept	-2.918e+00	2.852e-01	-10.231	< 2e-16
X3200M_INwater	-7.585e-07	3.810e-07	-1.991	0.04646
X800M_Floodway	6.375e-07	2.573e-07	2.478	0.01322
X400M_Flood_FR	5.629e-06	2.179e-06	2.584	0.00978
Farms	1.423e-03	8.177e-04	1.740	0.08185
Residual Deviance	272.37	**Pseudo R^2**	0.0991	
AICc	282.49			

In the aspatial *E. coli* regression, two flood variables, floodway within 800 meters, and flood fringe within 400 meters, were significant. Both of these variables had a positive correlation with the number of positive *E. coli* water samples. As the amount of flood and flood fringe land around a well increases, the more likely a well is to become

contaminated during a flood in the Calgary Health Zone. Intermittent water is another feature of water that appeared significant in the regression model. However, unlike the flood variables, the relationship between positive *E. coli* samples and intermittent water within 3200 meters is a negative relationship. As the area of intermittent water increases the number of wells that test positive for *E. coli* decreases. Lastly, the number of farms around each of the sampled well also was significant in the regression model, although this variable was only significant at the 91% CI and was retained due to significance in the literature discussed above. This farm variable also exhibits a positive relationship with the number of wells positive for *E. coli*.

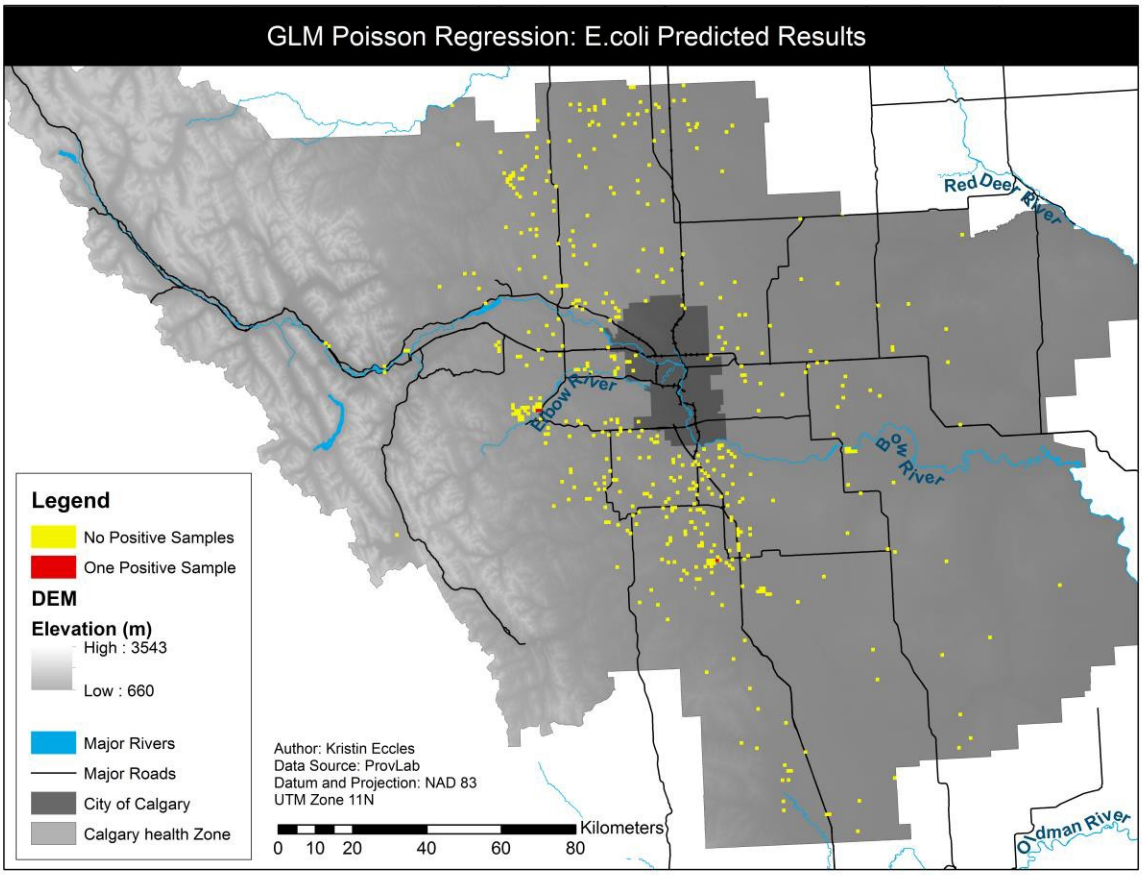

Figure 21. Aspatial Poisson GLM Regression: E.coli Predicted Results.

Figure 21 is a visual representation of the predicted number of positive samples in each quarter section. Of all quarter section included in the regression model, this model predicts that only a few quarter sections will have one positive test results for *E. coli*. This

is represented by the red quarter sections below. The predicted positive quarter sections are located on the Elbow River west of Calgary, as well as south of Calgary.

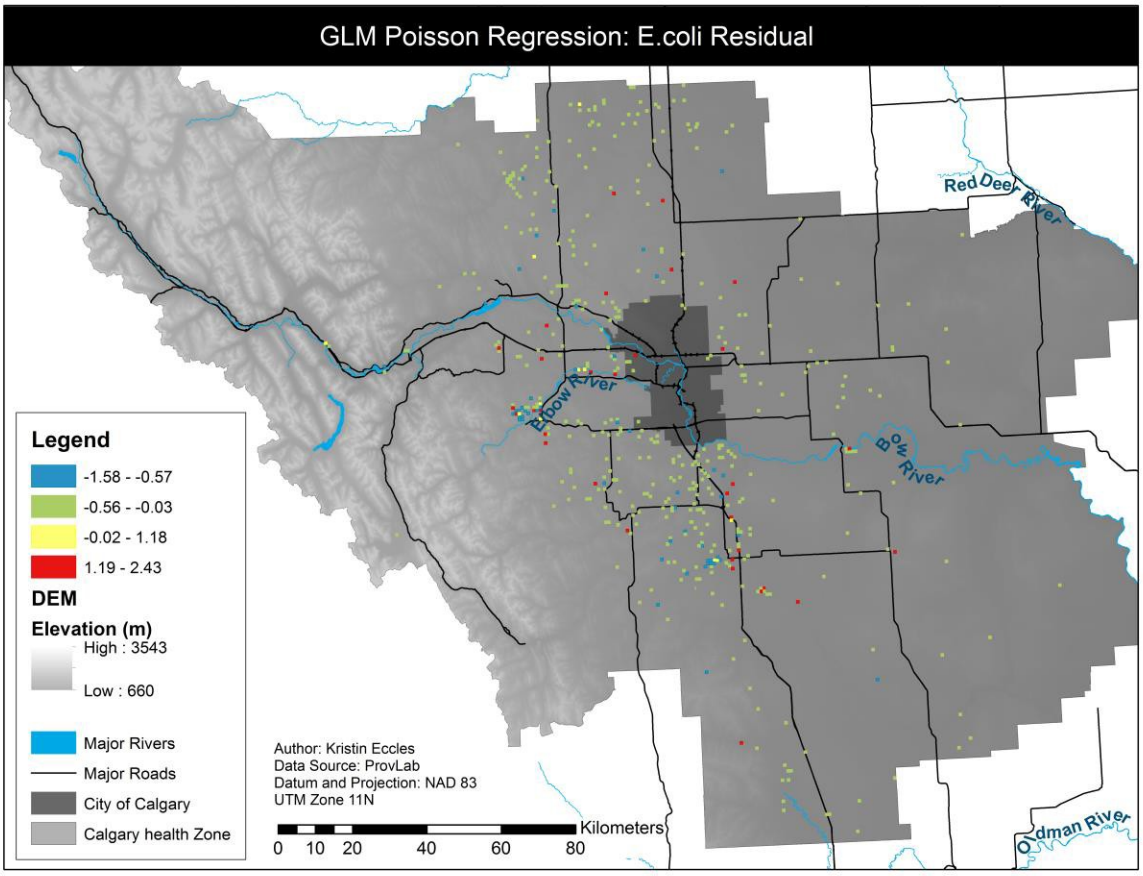

Figure 22. Map of Residuals for the aspatial Poisson GLM regression model.

Figure 22 shows the residuals of the aspatial *E. coli* regression model. The quarter sections that are red represent where the model over-predicts the number of positive *E. coli* water samples. Conversely, the blue indicates where the model under-predicts. This map demonstrates that most of the predictions in the model are slightly lower than the number of positive samples results in each quarter section. The residuals of the global model are not spatially autocorrelated (I=0.003, p=0.6499).

Upon testing the heteroscedasticity of the residuals of this regression model, the test revealed that the *E. coli* model does exhibit non-stationarity (bp=78.41, p=3.792e-16). As mentioned above, although the Breush-Pagan test is intended for linear data, and can only be used as a proxy for non-stationarity, the significance of the test indicates that a

geographically weighted regression would be an important improvement to the model. Using an optimized fixed Gaussian bandwidth of 39805.55 meters the GWR-Poisson model can be seen below in Table 11.

Table 11. *E.coli* Spatial Poisson Regression						
	Min.	**1st Qu.**	**Median**	**3rd Qu.**	**Max.**	**Range**
Intercept	8.16E-03	7.10E-02	8.09E-02	8.84E-02	1.11E-01	1.03E-01
3200M Intermittent Water	-1.29E-07	-5.65E-08	-3.97E-08	-3.12E-08	-9.88E-09	1.19E-07
800M Floodway	1.93E-07	2.39E-07	2.86E-07	3.38E-07	3.95E-07	2.02E-07
400M Flood Fringe	2.74E-07	5.06E-06	6.83E-06	8.60E-06	1.02E-05	9.92E-06
Farms	-5.33E-05	6.90E-05	1.20E-04	2.02E-04	6.11E-04	6.64E-04
Residual Deviance	188.82		**Quasi-Global R^2**	0.155		
AICc	199.07					

All variables in the geographically weighted Poisson regression vary over space, with the intercept having the greatest range over space. This is followed by the number of farms that surround the well. The water related variables vary less than the other aforementioned variables. All variables retain the original sign, and therefore the original relationship between the dependent and independent variable. All variables have a positive relationship with the number of *E. coli* positive well samples except for intermittent water that retains a negative relationship.

The geographically weighted regression also changed the significance of the farm variable in the regression. In the aspatial global model, this variable was not significant at the 90% CI, however in the geographically weighted model, this variable is significant at the 95% CI in 27 (of 470) of the regressions and is significant at the 90% CI in 32 more regression models. The significance of the remaining variables are below the 90% CI (411 of 470).

Compared to the aspatial regression, the geographically weighted regression performed better as indicated by the AICc decrease from 282.49 to 199.07, the deviance of the residuals, and the quasi-global R^2. The deviance of the residuals decreased from 272.37 to 188.82. This smaller deviance in the geographically weighted model indicates that the predicted values deviate less from the measured values than in the aspatial

model, indicating the more accurate predictive model. Additionally, the decrease in the AICc indicates that the geographically weighted model is statistically better fit than the aspatial model. Lastly, the quasi-global R^2 increases. While the R^2 is reported to be 0.15, this is an average of all of the local R^2. Therefore, the local R^2 range from 0.065 to 0.198 indicates that in this model, between 6.5% and 19.8% of the variance in the dependent variable can be explained by the independent variables. Below in Figure 23 are the results from the geographically weighted regression.

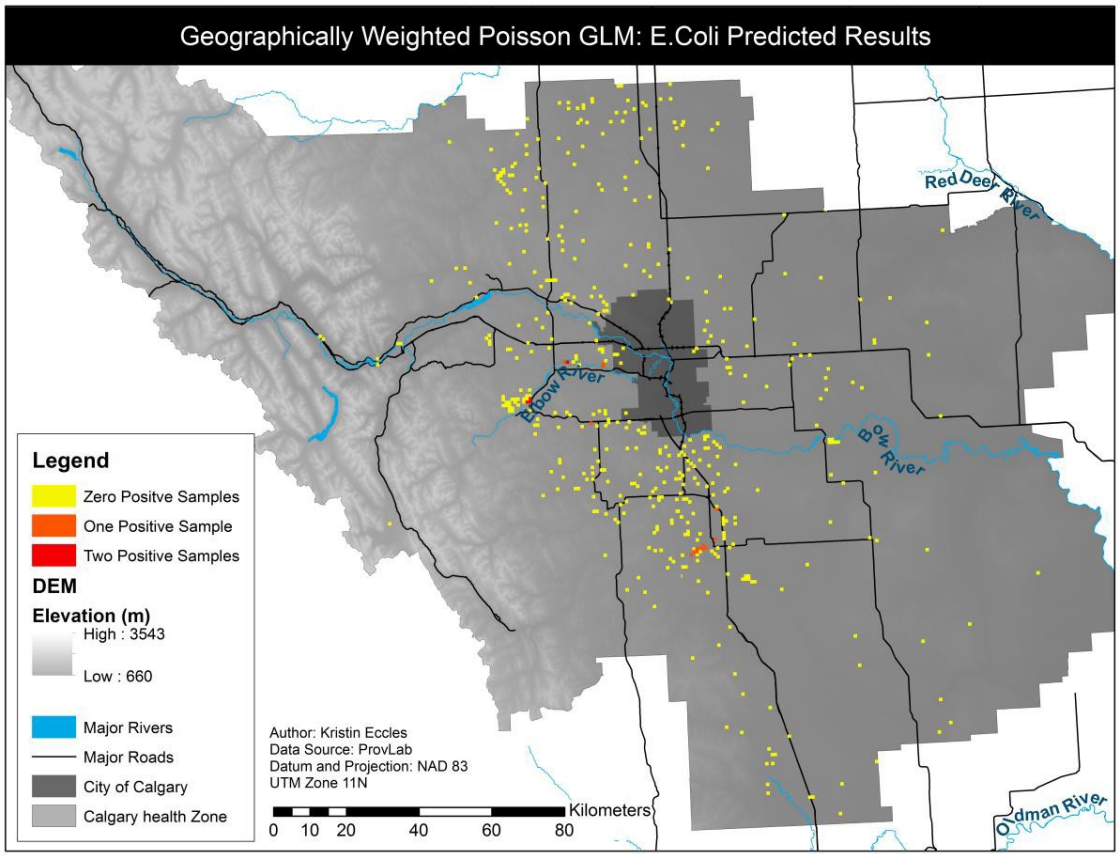

Figure 23. Predicted results for the geographically weighted regression for *E.*

Compared to the aspatial regression, the geographically weighted regression predicts additional quarter sections that will have at least one water sample positive for *E. coli*. In addition, unlike the aspatial regression model, the geographically weighted regression model also predicts quarter sections that two positive water results that will be positive for *E. coli*.

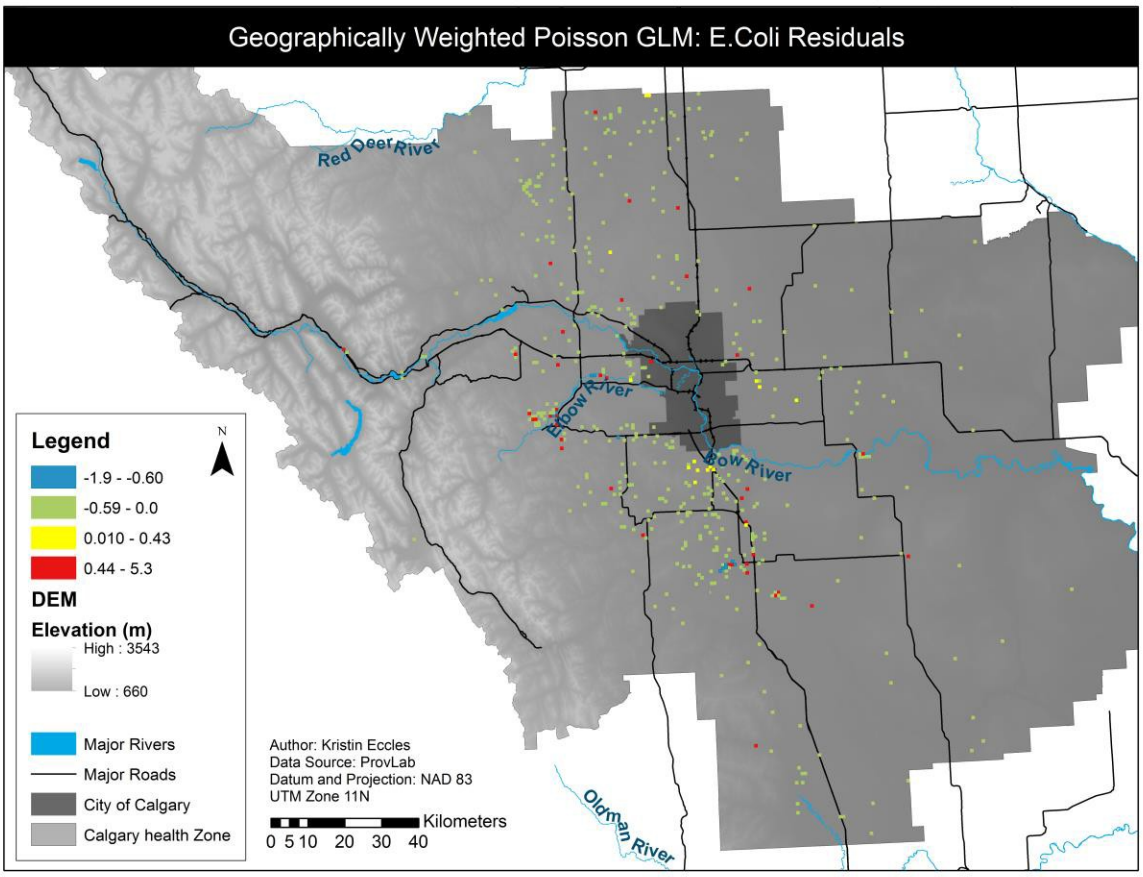

Figure 24. Residuals of the geographically weighted *E.coli* regression model.

The residuals of this model indicate that for the most part, the model slightly under predicts the actual value, which can be seen in Figure 24, as most of the quarter sections are a light green colour. There are also a few quarter sections that over predict the number of positive *E. coli* positive water well results. After the geographically weighted regression was performed, the residuals exhibited homoscedasticity, as well, had an insignificant global Moran's I=-0.03 (p=0.0012). This indicates that using the geographically weighted regression was able to reduce the heteroscedasticity of the residuals, while not inducing spatial autocorrelation.

Figure 25 shows the local R^2 of the geographically weighted *E. coli* model. In this figure, as the colour progresses from yellow to red as the more the variance in the dependent variables can be explained by the independent variables. More variance can be explained by the independent variables west of the city of Calgary and less of the variance

can be explained decreasing concentrically the further south, north and east of the city of Calgary the quarter section is.

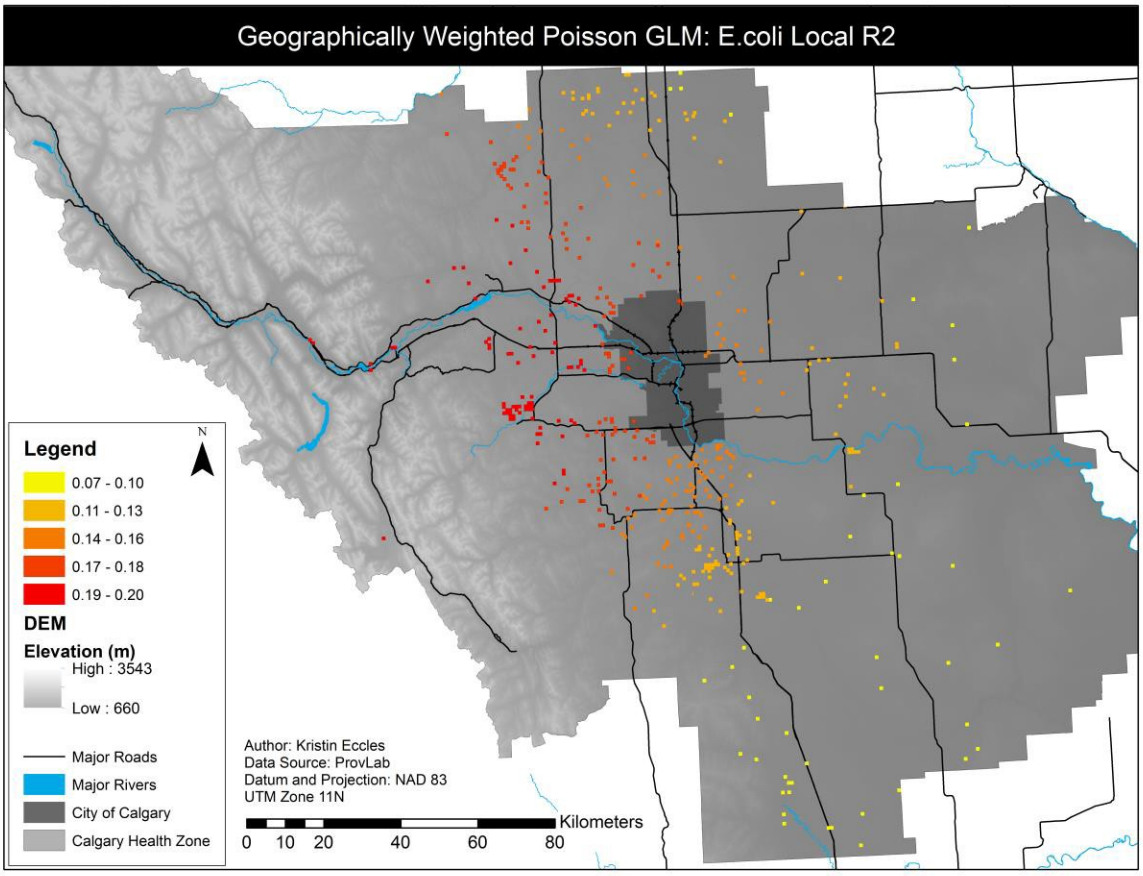

Figure 25. Local R2 of the geographically weighted E.coli regression model.

In the figures below it is possible to see how each of the variables contribute to the regression model by assessing the direction and the strength of the t-value. Additionally, as this is a geographically weighted regression it is possible to assess how the relationship changes over space. In Figure 26, the t-values of the intermittent water variables can be seen. Over the entire study areas there are no quarter sections that have a positive relationship between the area of intermittent water and number of positive *E. coli* samples. The region where this variable is most significant is around the Elbow and Bow Rivers west of the city of Calgary. This is the only variable that has a negative relationship with the dependent variable.

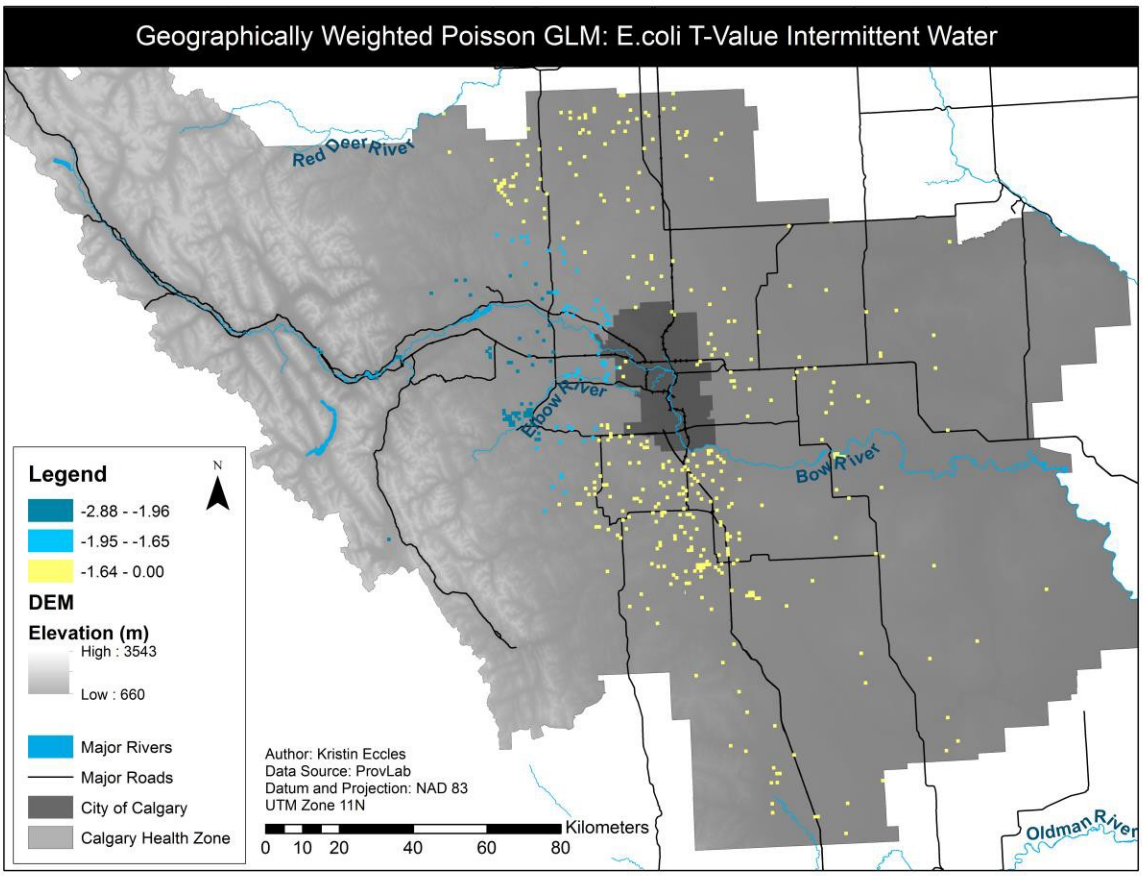

Figure 26. Intermittent water t-values of quarter sections in the geographically weighted regression model for *E.coli*.

The remaining variables in the regression all exhibit a positive relationship between the dependent variable and independent variable. The second variable, flood fringe can be seen below in Figure 27. The flood fringe has the most significance of all variables in the regression as demonstrated by the red quarter sections in the figure. The red represents quarter sections where the flood fringe variable is significant at the 95% CI. There are only a few quarter sections that are not significant in the regression.

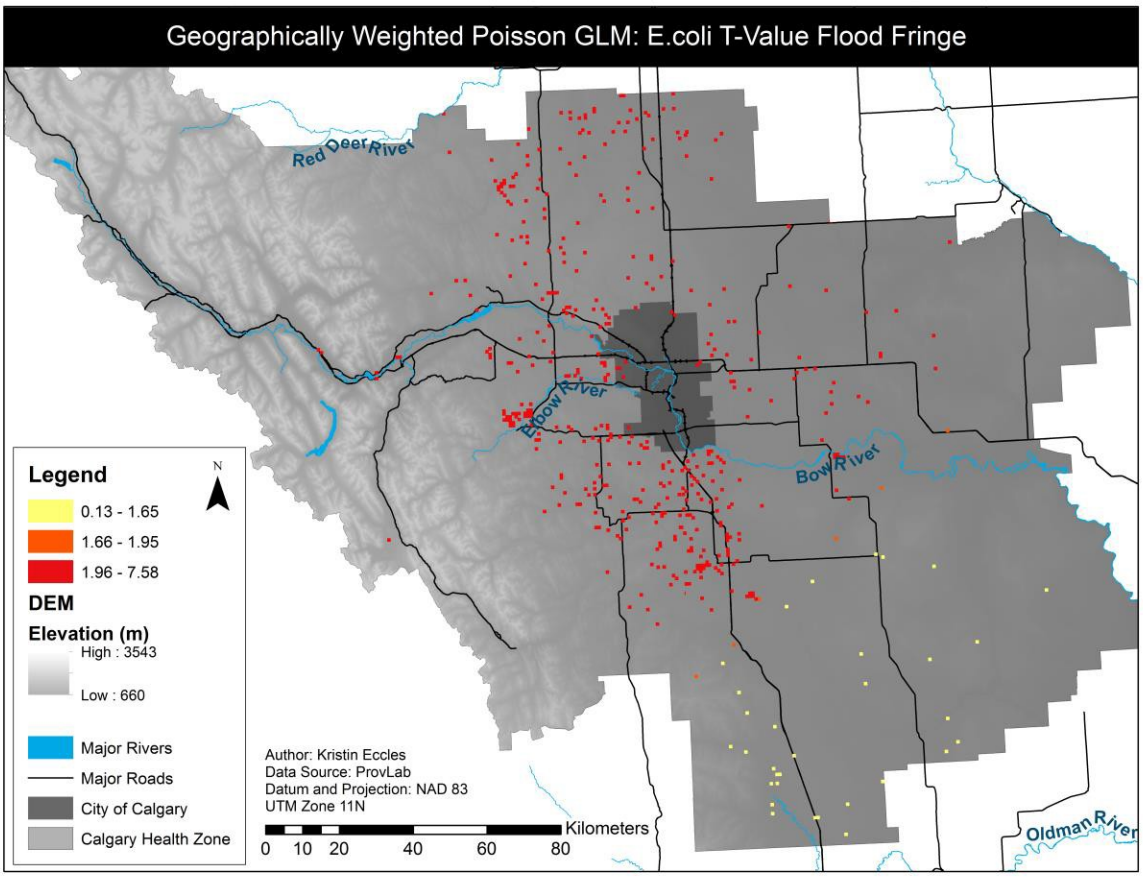

Figure 27. Flood fringe t-values of quarter sections in the geographically weighted regression model for *E.coli*.

The independent variable floodway, seen in **Figure 28** also has a large proportion of the quarter sections exhibiting statistical significance at the 95% CI. It is evident that the significance of the variable decreases towards the northwest section of the health zone. The variable is most significant around the city of Calgary and to the south and southeast of the city of Calgary.

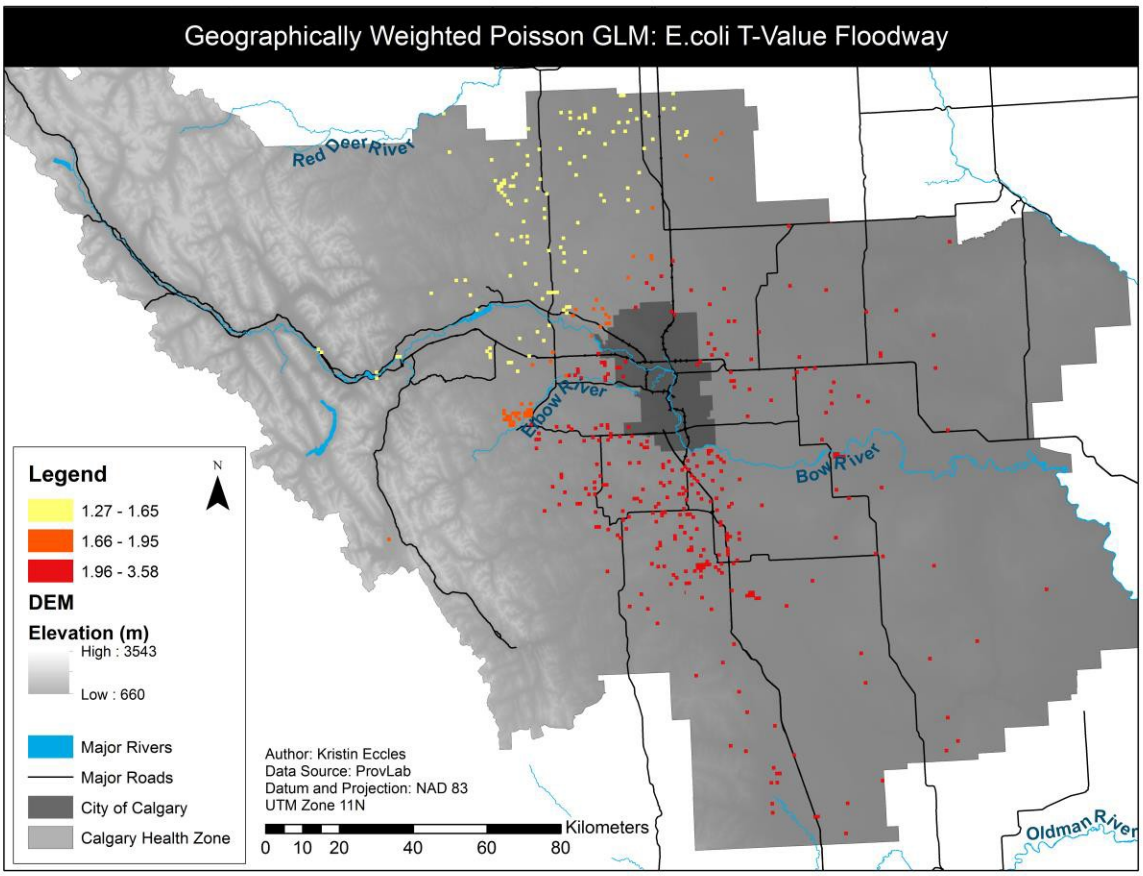

Figure 28. Intermittent water t-values of quarter sections in the geographically weighted regression model for *E.coli.*

The final independent variable in the regression is the number of farms. The varying significance can be seen in Figure 29. This variable is most significant around the western section of the Bow River and in the western end of the Elbow River. While the geographically weighted regression was able to demonstrate a region where the farm variable was statistically significant at the 95% CI, there are still many quarter sections that remain insignificant in this regression model.

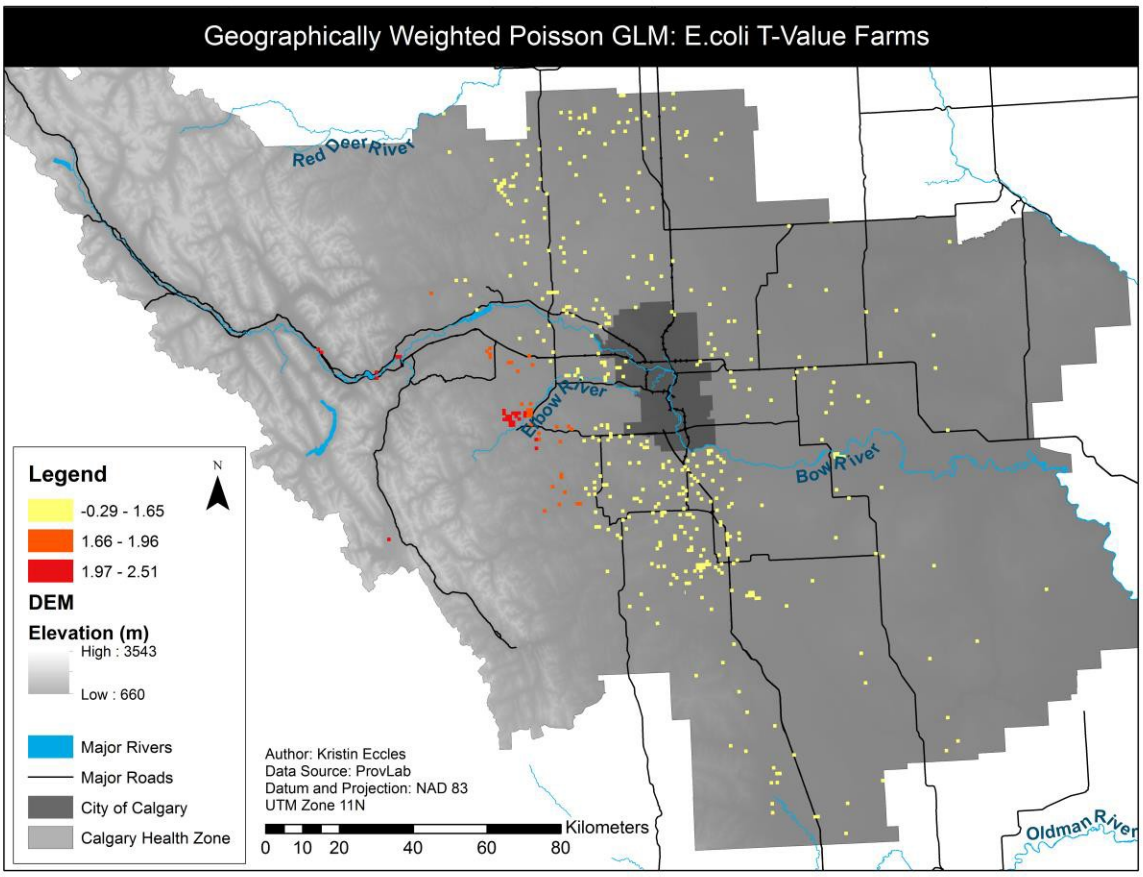

Figure 29. Farm t-values of quarter sections in the geographically weighted regression model for *E.coli*.

Finally the interpolated beta values were used to create a risk surface for each quarter section in the Calgary Health Zone. The result was the environmental risk map below in Figure 30. The yellow indicates areas where the environmental risk is low, as model by the geographically weighted generalized linear model. The light orange colour that covers most of the Calgary Health Zone indicates a moderate environmental risk of drinking water well contamination during a flood, and finally the dark orange colour indicates a higher environmental risk of contamination during a flood. In this map there is also an overlay of the relief to demonstrate how the risk changes as the elevation also changes. The model predicts the risk of contamination to be highest in the center of the Calgary Health Zone. The lower risk is predominately seen in eastern region of the health zone.

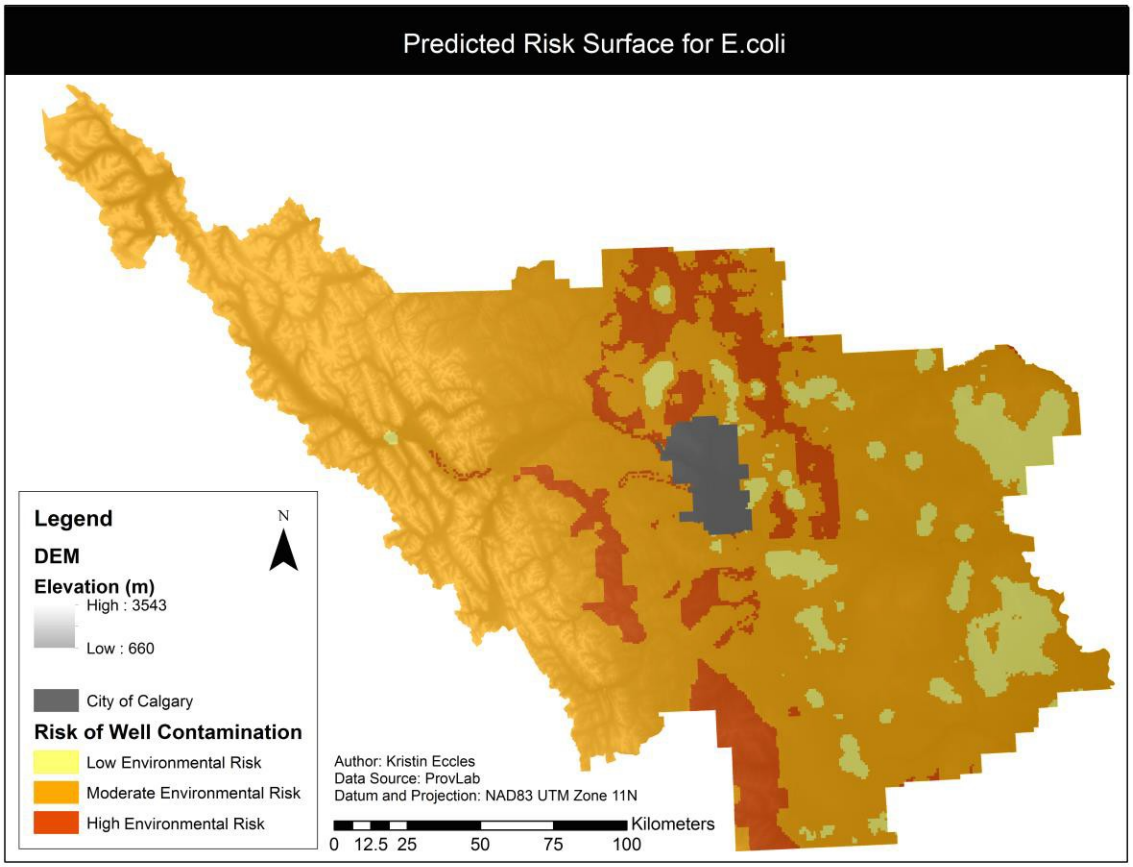

Figure 30. Risk surface produced from the geographically weighted Poisson generalized linear model.

Chapter Five: Discussion

Historical Comparison

The results of the Wilcoxon Rank Sum Test (z-test) indicate that the percentage of positive total coliforms and *E. coli* positive water samples seen in the previous seven years is statistically different from the percentage of positive results seen in 2013. From these results it is likely that there was an event that occurred in 2013, which had not occurred in the previous years, causing the percentage of positive samples to be statistically different. It is likely that this event was the June 2013 flood. The results of this test indicates that it is important to look at what factors could have contributed to this increased percentage of positive results, which was explored through environmental factors.

Interestingly, the percentage of positive results seen in 2013 was not statistically larger than the results seen in 2005. The total coliforms the results seen in 2005 and 20013 were statistically the same indicating that there was no difference between the percentages of positive total coliforms test results between the two years. However, for *E. coli* the percentage of positive samples seen in 2005 was statistically larger than the results seen in 2013. In June of 2005, there was also the occurrence of a flood. Similar to the June 2013 flood experienced in southern Alberta, the flood that occurred in June of 2005 also extreme earning a place in Environment Canada's top weather story for 2005 (Environment Canada, 2013a; Environment Canada, 2013b). Comparing the two flood years, following the June 2005 flood more samples were received than there were following the June 2013 flood.

The winter preceding June 2005, much like the winter preceding the June 2013 flood, had seen an abnormally large volume of precipitation (Environment Canada, 2013b). Although, in 2005 the spring was dry, June is known to be the wettest month. As the snowpack from the mountains was melting, in conjunction with three major storms that passed through southern Alberta in the same week, caused water levels to rise drastically. During this time the Bow, Old, and Red Deer Rivers had a flow rate between 10 to 30 times greater than normal. The flooding that occurred in June 2005 was a one in 200 year flood. Similar to the June 2013 flood, the June 2005 flood resulted in a state of

emergency being declared in 14 different towns, evacuations, and damage reported to be about $400 million (Environment Canada, 2013b). Comparing the percent of positive samples seen in 2005 to 2013 suggest that the 2005 flood affected the groundwater sourced by private households more than the 2013 flood.

There were more samples taken in 2013 than in the preceding years 2006-2012, however using a proportion (percent of positive test results) for the analysis controls for the number of samples. Interestingly, though statistically there were more samples taken in 2005, than in the same time period in 2013, the percentage of positive total coliforms results are not statistically different between the two flood years. The higher number of samples that were taken by homeowners during flood years could also indicate that when there is a potential threat to drinking water quality, homeowners are more likely to sample their drinking well water.

Descriptive Results

Spatially, while the results of the nearest neighbour dispersion analysis indicated that the location of all samples are clustered, the results (positive or negative) of these samples were not spatially autocorrelated at the global level as indicated by the Moran's I analysis. However, locally, the Getis and Ord's G* cluster analysis indicated that there was spatial association at a local level occurring in both the total coliforms and *E. coli* positive sample results. This clustering can be seen in both maps (Figure 8 and Figure 9). Interestingly, there were only cold spot clusters of negative test results for total coliforms. Although total coliforms did have a hot spot cluster of positive test results, *E. coli* only had clusters of positive results. Another interesting aspect was the location where these clusters occur. Between total coliforms and *E. coli* there were similar hot spot clusters in the southern region of the Calgary Health Zone. The other clusters occurred in the northern region of the Calgary Health Zone, however, the cold cluster of negative total coliforms results occurred in the north region of the Calgary Health Zone and the hot cluster of positive *E. coli* results occurred further south from the total coliforms cold spot cluster, northwest of the city of Calgary.

The different locations and different types of clusters seen in the study area indicates that likely, different processes were affecting the contamination of private drinking water wells with total coliforms and with *E. coli*; regression modeling can provide some insight into the processes that either do or do not cause the patterns seen. As a result, it was appropriate to model total coliforms and *E. coli* separately. This clustering also indicated that likely, a geographically weighted regression would have to be utilized as this non-stationarity violates the regression assumptions, which would cause the variance to be inflated.

Quantitative Results

Total Coliform Model

Although it appeared as though the geographically weighted COM Poisson regression was a better fit statistically, based on the comparison between the spatial and the aspatial deviance of residuals values, AICc, and pseudo-R^2/Quasi R^2, the interpretability of the produced model was troubling. The high variability, exemplified by the changes between negative and positive relationships represented by the beta values, made the models difficult to interpret. While it appeared that the model improved with the use of the geographically weighted regression, when applied to the real world, in actuality it was not a better fit. This suggests that the variables included in the regression were not able to accurately model the relationship between the environmental variables and the occurrence of samples positive for total coliforms. It is possible that the correct environmental variable was not included. It could also indicate that the correct environmental variable was included, but the resolution of the GIS layer was used was too coarse for a fine scale analysis of a quarter section. However, since none of the flood variables were significant in the regression model, it is likely that the June 2013 flooding is not associated with the number of positive total coliforms samples.

As noted in the background section, there can be many different sources of total coliforms that are not classified as faecal. Total coliforms are used as an indicator of water quality as a part of a multi-barrier approach to ensuring the safety of drinking water. Although there are no health-based risk assessments for total coliforms in the absence of

faecal coliforms, as most total coliforms are not considered a risk to human health, a positive total coliforms test does warrants further investigation. When a water sample tests positive for total coliforms but negative for *E. coli*, this indicates that the source is not faecal. As protected groundwater system should contain zero total coliforms, when this test result occurs, this indicates that there is likely contamination from the surrounding environment. This contamination of well water can occur after the construction of a new well, or during maintenance of an old well. Additionally, total coliforms can become desensitized to the treatment of drinking water, which can result in the regrowth of bacteria in the water system (Health Canada, 2013).

As indicated by Health Canada the presence of total coliforms in the absence of *E. coli* is usually caused by maintenance issues associated with the individual well. Although the environment is associated with total coliforms contamination, it appears the construction and maintenance may have a greater association with contamination. It is likely that the total coliforms regression is lacking variables that could have been able to explain more of the variance in the dependent variable; these variables are the individual characteristics of the wells. Therefore, the total coliforms model would benefit from an individual scale analysis.

Although variables that can be associated with groundwater contamination, as indicated by the DRASTIC methodology, were included in the regression variables, none of the variable remained significant in the regression. Additionally, none of the flood variables were statistically significant in the regression. This is another indication that the flood and other water related variables did not have an impact on the contamination of private drinking water wells with total coliforms that could be detected in this regression model.

E.coli Regression Model

As there are two flood variables, area of land designated as flood fringe and floodway were significant in the *E. coli* regression model, this indicated that the flooding that occurred in June 2013 is associated with the risk of *E. coli* contamination. As none of the surficial features were significant, such as the hydraulic connectivity, it suggest that

this flood that occurred within a short period of time affected mostly surface water, although more research is needed to confirm this. Therefore, this type of contamination is more likely caused from contamination through the well head.

In the geographically weighted regression, the highest R^2 value was 0.19 with a mean of 0.15, suggesting that there is uncertainty not accounted for in this model. Aside from the limitations pertaining to the layers included in the regression, it is likely that some of this variance could be accounted for by variables representing the characteristics of the individual wells. Research demonstrates that well characteristics such as construction type and depth, as well as maintenance are associated with the quality of water obtaining from the well (Health Canada, 2013). Therefore, when the wells that had prior vulnerability due to poor well maintenance and construction, were impacted by the flood, the combination of these two occurrences may have predisposed certain wells to drinking water contamination. If the characteristics of the individual wells are able to explain more of the variance, this would be most advantageous to the home owner as characteristics of the individual well are easier to modify than factors associated with the environment. Further research is needed to confirm this theory.

Areas where the residual deviance was high indicate a quarter section where the variables included in the model do not as accurately model the contamination of drinking well water. There was no spatial autocorrelation in the residuals. Thus, the areas of high deviance could indicate where the contamination is more likely due to the condition and maintenance of the individual well. The condition of individual wells is assumed to be independent of neighbouring wells. If there were clustered results, then this could indicate that there was a process, either environmental or non-environmental, not accounted for with the model acting upon those well. As this was not the case, it is more likely that the individual characteristics of the well, such as a cracked well head, in conjunction with the flood acted multiplicatively causing the contamination of private drinking water wells. This information can be of importance in policy as individuals who have submitted samples in quarter sections of high deviance should have an inspection of their well for damage or defects that could more easily allow for contamination.

One of the most interesting relationships in this regression is that with intermittent water or sloughs. Intermittent water bodies can dry up during certain times of the year (DTMI, 2009). These regions have characteristics of wetlands, with shallow bodies of water that are predominately covered in vegetation suitable to saturated soil conditions. Although when using a geographically weighted regression is not uncommon for the direction of the relationship to change between the global and local model, and even within the local model (Kupfer and Calvin, 2007), this did not occur between the aspatial and spatial regression for the intermittent water variance. This variable remained negative even after being geographically weighted. Vegetation commonly found in Alberta's sloughs are duckweed, bulrushes and cattail (Smith et al., 2007). Sloughs/ wetlands are a vital part of the ecosystem. They serve functions such as peat production, carbon storage, and water purification (Zedler and Kercher, 2005). Wetlands have demonstrated the ability to reduce high concentration of nutrient in water over long periods of time, improving water quality (Verhoeven et al., 2006). Due to this ability, wetlands have been constructed as methods of water treatment for municipal, domestic, and animal wastewater treatment (Kadlec and Wallace, 2008). Natural wetlands are efficient in reducing the nutrient load of nitrates and phosphorus and are more effective at removing *E. coli* from surface water through biogeochemical processes. These processes and abilities of the wetlands could explain the negative relationship between drinking water wells contaminated with *E. coli* and intermittent water bodies. When properties have larger volumes of intermittent water nearby, the regression revealed the wells on such properties were less likely to become contaminated with *E. coli* during a flood. More research would be necessary to assess whether this relationship holds true when there was no flooding.

Limitations

The most prominent limitation that this study faces is the availability of GIS layers. A model built using a GIS is only as good as the input information (Longley et al., 2005). The coarse resolution of the geologic variables (1:100,000), and simple lack of information

available (distance to groundwater) had the largest effect on the modeling. Limitations of specific layer inputs and variables will be discussed below.

Farm Variables

The lack of a reliable GIS layer containing information pertaining to farms greatly hampered this model. Although farm variables were included in two different ways, the data was not optimal. The 2006 agricultural census data was reported by Soil Land Survey of Canada Polygons for both hectares of farm land, as well as number of farms. These Soil Land Survey polygons are an irregular shape and are not optimal for the method of intersecting buffers and environmental features that was used in this research. Additionally, due to the sensitivity of the data, farm information recorded in the 2011 Agricultural Census was too coarse to be used. There were approximately four polygons that were completely within the study area, and two polygons that were only partially in the study area. The 2006 Agricultural Census data was used, rather than the 2011 Agricultural Census data, as this older data is less sensitive and could be obtained at a finer resolution. Due to the tradeoff between resolution of currency of data, as well as non-optimal data format, this layer can only be used as a proxy for farms. The second category of data that had information pertaining to farms was obtained through remotely sensed data. Based on the accuracy assessment of the land cover classification, the five general groups were has accuracy of 88%. This means that 12% of the pixels over the province were classified incorrectly. This error adds to the uncertainty of the modeling. Additionally, the overlap between the two farming variables is unknown. However, since only the farming proxy variable (not the land cover classification data) was included in the final regression models, it is unlikely that this would effect the final result.

Previous studies clearly demonstrate a link between agriculture and *E. coli* contamination of drinking water (Richardson et al., 2009; Wallender et al., 2013). To better this research and produce a more accurate model, having accurate information pertaining to number of farms, types of farming, and manure practices is imperative. Farming is a major industry in Alberta that provides a livelihood for many of the residents. Alberta represents 21.1% of Canada' agricultural production and is one of the most

productive agricultural economies in the world. In the province there is over 50 million acres of land devoted to crop and livestock production. This agricultural production added $10.5 billion to the economy of Alberta in 2011. Additionally, secondary industries of farming such as food processing also provide job opportunities for Albertans and contributed more than $11 billion to the economy from the food processing industry. Of this $11.3 billion, meat processing accounted for more than half (AARD, 2013).

It is clear that the agricultural industry is important to Alberta's economy providing sources of income, revenue, and jobs. It is also clear that the unintentional ingestion of *E. coli* via drinking water has the potential to make many individuals very ill. To have the health of all individuals, including the farm operators, and farming operations co-exist in a way that is not only beneficial to human health but also to the productivity and overall health of the economy, data sharing and transparency is necessary. Being able to understand the relationship between farming operations and human health is important, as it is through this type of research that proactive measures can be put in place to avoid the need for reactive measures and remediation.

Hydraulic Input

This model also lacks the input of another important system, which was the amount of water that area around each of the wells received. In both models, the amount of rain that had fallen in the month of June was included as increased amounts of rainfall can be associated with the contamination of drinking water (Dorners et al., 2007; Hofstra, 2011). However, this did not take into account the larger system that was at play, which was the overburdening of the river systems. During June of 2013, Alberta received higher than normal volumes of water due to rainfall. This was taken into consideration in the model. What was not taken into consideration was how the river system came to be overburdened. During the winter of 2012-2013, a large volume of snow had fallen in the mountain. During the spring of 2013, the large volume of snow began to melt and flow into the Elbow River. This river flows from the Rocky Mountains towards Calgary. In the City of Calgary, the Elbow River merges into the Bow River. It was the taxation from the

higher than normal volume of spring melt water, in addition to the heavy rain that was received June 2013 that caused the flooding.

Having an environmental input into the model that could capture the process described above would have very beneficial for this research. Although the model's inputs included information on the flood fringe area, floodway, and area of overland flooding, based on satellite imagery from the June 2013 flood, having variables with additional information such as water height, volume and/or speed would also have benefited this research.

Chapter Six: Conclusions

To conclude, this research demonstrated that the number of private drinking water samples positive for total coliforms and *E. coli* were significantly higher in 2013 than previous non-flood years. This warrants the further investigations of how the 2013 June flood impacted private drinking water in the Calgary Health Zone. Descriptive and analytical methodologies were undertaken to further investigate.

The descriptive analyses indicated that the dependent variable, total coliforms and *E. coli* contaminated water wells were clustered. As a result, a geographically weighted regression was used for the quantitative analysis. Additionally, the clusters of negative test results and the cluster of positive test results occurred in different locations indicating that different processes were influencing water well contamination. As a result, total coliforms and *E. coli* were modeled separately to determine the underlying processed that caused the differing pattern seen in the cluster analysis.

Although a statistically significant GLM Poisson regression model for total coliforms was produced, this methodology was not able to accurately model the relationship between total coliforms and environmental variables. As a result, it is likely that characteristics of the individual wells are associated with well contamination by total coliforms. From this research, it is also likely that the flood was not a large influencing factor on the contamination of private drinking water wells with total coliforms.

The *E. coli* model produced a statistically significant, and interpretable model. Less common spatial methods of a geographically weighted Poisson regression were used. This aided in satisfying the assumptions of the GLM Poisson regression. This methodology demonstrated an improved model over the aspatial model, as indicated by model indicators such as AICc, and residual deviance.

In this model, flood variables proved to be statistically significant showing a positive relationship with the number of positive well samples positive for *E. coli*. This demonstrated that the flooding is associated with the contamination of private drinking water wells. Additionally, a proxy for the number of farms around a well was found to be statistically significant, demonstrating that proximity to farms can impact well water,

especially during a flood. Interestingly, intermittent water/ sloughs was also significant however, the relationship between the two, unlike the other variables, was negative. This indicates that having intermittent water, also known as sloughs or wetlands, around the water well helps to decrease the *E. coli* contamination of private drinking water wells during a flood. This is due to biogeochemical processes within the wetlands that naturally purifies water by removing nutrients and bacteria such as *E. coli*.

It is recommended that in the future, research be conducted at the individual well level. This would likely improve the modeling of total coliforms as discussed above. Additional information such as characteristics of the individual wells such as age, depth, and construction type, should be included in the model. Having this information may be able to explain more of the variance seen in the contamination variables, total coliforms and *E. coli*. This research would also benefit from more accurate GIS layers as identified above.

References

Alberta Agriculture and Rural Development (AARD).(2005). Agricultural Land Resource Atlas of Alberta - Manure Production Index for the Agricultural Area of Alberta. Retrieved from: http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/agdex10335

Alberta Agriculture and Rural Development (AARD).(2010). Groundwater Vulnerability in Alberta. Retrieved from: http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/wat6416

Alberta Agriculture and Rural Development (AARD). (2013). Agriculture Statistics Factsheet 2012. Retrieved from: http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/sdd12807

Alberta Environment and Sustainable Resource Development (AESRD). (2012). Groundwater. Retrieved from: http://environment.alberta.ca/03137.html

Alberta Environment and Sustainable Resource Development (AESRD). (2013). Flood Hazard Mapping. Retrieved from: http://environment.alberta.ca/01655.html

Alberta Biodiversity Monitoring Institute (ABMI). 2010. Data Request Overview. Retrieved from: http://www.abmi.ca/abmi/rawdata/rawdataselection.jsp

Alberta Government. (2013). Updated provincial flood statistics. Retrieved from: http://alberta.ca/release.cfm?xID=3504558578E5F-0AE3-E7A3-F26352B04E9C94B0

Alberta Health Services (AHS). (2009). Environmental Public Health: Frequently Asked Questions About Well Water Testing Frequently. Retrieved from: http://www.albertahealthservices.ca/Advisories/ne-pha-2011-07-14-faq-well-testing.pdf

Alberta Health and Wellness. (2004). Environmental Public Health Field Manual For Private, Public and Communal Drinking Water Systems in Alberta (Second Addition). Retrieved from: www.health.alberta.ca/documents/Drinking-Water-Systems-2004.pdf

Alderman, K., Turner, L. R., & Tong, S. (2012). Floods and human health: a systematic review. Environment international, 47, 37-47.

Aller, L., Bennett, T., Lehr, J. H., Petty, R.J., and Hackett G. (1987).DRASTIC: A standardized system for evaluating ground water pollution potential using hydrogeologic settings: NWWA/EPA Series, EPA-600/2-87-035.

Auld, H., MacIver, D., & Klaassen, J. (2004). Heavy rainfall and waterborne disease outbreaks: the Walkerton example. Journal of Toxicology and Environmental Health, Part A, 67(20-22), 1879-1887.

Babiker, I. S., Mohamed, M. A., Hiyama, T., & Kato, K. (2005). A GIS-based DRASTIC model for assessing aquifer vulnerability in Kakamigahara Heights, Gifu Prefecture, central Japan. Science of the Total Environment, 345(1), 127-140.

Bear, J. (2014). Hydraulics of groundwater. Courier Dover Publications.

Borchardt, M. A., Haas, N. L., & Hunt, R. J. (2004). Vulnerability of drinking-water wells in La Crosse, Wisconsin, to enteric-virus contamination from surface water contributions. Applied and Environmental Microbiology, 70(10), 5937-5946.

Bouwer, L. M. (2011). Have disaster losses increased due to anthropogenic climate change?. Bulletin of the American Meteorological Society, 92(1).

Breslow, N. E. (1996). Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata*, *8*, 23-41.

Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environment and planning A, 30, 1905-1927.

Brunsdon, C., Fotheringham, A., & Charlton, M.(1996). Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis, 28(4), 281-298.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician), 47(3), 431-443.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1999). Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science, 39(3), 497-524.

Charlton, M., Fotheringham, S., & Brunsdon, C. (2009). Geographically weighted regression. White paper. National Centre for Geocomputation. National University of Ireland Maynooth.

Chatterjee, S., & Hadi, A. S. (2006). Regression analysis by example (Vol. 607). Wiley-Interscience.

Charrois, J. W. (2010). Private drinking water supplies: challenges for public health. Canadian Medical Association Journal, 182(10), 1061-1064.

City of Calgary. (2014b). Flood recovery. Retrieved from: http://www.calgary.ca/General/flood-recovery/Pages/home.aspx?

Dobson, A. J. (2001). *An introduction to generalized linear models*. CRC press.

Dorner, S., Anderson, W., Huck, P., Gaulin, T., Candon, H., Slawson, R., & Payment, P. (2007). Pathogen and indicator variability in a heavily impacted watershed. Journal of water and health, 5(2), 241-257.

Environment Canada. (2013a). Canada's Top Ten Weather Stories For 2005. Retireved from: http://ec.gc.ca/meteo-weather/default.asp?lang=En&n=A4DD5AB5-1

Environment Canada. (2013b). Canada's top 10 weather stories for 2013. Retrieved from: http://ec.gc.ca/meteo-weather/default.asp?lang=En&n=5BA5EAFC-1&offset=2&toc=show

Environmental Protection Agency (EPA). (2011). Groundwater Contamination. Retrieved from: http://www.epa.gov/superfund/students/wastsite/grndwatr.htm

Erener, A., & Düzgün, H. S. B. (2010). Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway). Landslides, 7(1), 55-68.

Few, R. (2003). Flooding, vulnerability and coping strategies: local responses to a global threat. Progress in Development Studies, 3(1), 43-58.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically weighted regression. Chichester: Wiley.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically weighted regression: the analysis of spatially varying relationships. Wiley.

Fotheringham, A. S., & Rogerson, P. A. (Eds.). (2008). The SAGE handbook of spatial analysis. SAGE Publications Limited.

Getis, A., & Aldstadt, J. (2010). Constructing the spatial weights matrix using a local statistic. In Perspectives on Spatial Data Analysis (pp. 147-163). Springer Berlin Heidelberg.

Goodchild, M. F. (2004). GIScience, geography, form, and process. *Annals of the Association of American Geographers*, *94*(4), 709-714.

Gray, S. (2008). Long-term health effects of flooding. Journal of public health, 30(4), 353-354.

Haines-Young, R. H., & Petch, J. R. (1986). *Physical geography: its nature and methods*. Harper & Row.

He, F., Zhou, J., & Zhu, H. (2003). Autologistic regression model for the distribution of vegetation. Journal of agricultural, biological, and environmental statistics, 8(2), 205-222.

Health Canada. (2013). Guideline Technical Document Total Coliforms. Retrieved from: http://www.hc-sc.gc.ca/ewh-semt/pubs/water-eau/coliforms-coliformes/index-eng.php#a8.0

Hiscock, K. M., & Grischek, T. (2002). Attenuation of groundwater pollution by bank filtration. Journal of Hydrology, 266(3), 139-144.

Hofstra, N. (2011). Quantifying the impact of climate change on enteric waterborne pathogen concentrations in surface water. Current Opinion in Environmental Sustainability, 3(6), 471-479.

Hrudey, S. E., Payment, P., Huck, P. M., Gillham, R. W., & Hrudey, E. J. (2003). A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. Water science & technology, 47(3), 7-14.

Hu, X., Waller, L. A., Al-Hamdan, M. Z., Crosson, W. L., Estes Jr, M. G., Estes, S. M., ... & Liu, Y. (2012). Estimating ground-level PM 2.5 concentrations in the southeastern US using geographically weighted regression. Environmental research.

Jamrah, A., Al-futaisi, A., Rajmohan, N., & Al-yaroubi, S. (2008). Assessment of groundwater vulnerability in the coastal region of oman using DRASTIC index method in GIS environment. Environmental Monitoring and Assessment, 147(1-3), 125-38.

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches*. Sage.

Kadlec, R. H., & Wallace, S. (2008). *Treatment wetlands*. CRC press.

Knox, A. K., Dahlgren, R. A., Tate, K. W., & Atwill, E. R. (2008). Efficacy of natural wetlands to retain nutrient, sediment and microbial pollutants. *Journal of Environmental Quality*, *37*(5), 1837-1846.

Kupfer, J. A., & Farris, C. A. (2007). Incorporating spatial non-stationarity of regression coefficients into predictive vegetation models. *Landscape Ecology*, *22*(6), 837-852.

Longley, P., Goodchild, M., Maguire, D., and Rhind, D. (2005). Geographic Information Systems and Science, 2nd Edition. Wiley.

Maloy, S. R., Cronan, J. E., & Freifelder, D. (1994). Microbial genetics. Jones & Bartlett Learning.

McCallum, J. E., Ryan, M. C., Mayer, B., & Rodvang, S. J. (2008). Mixing-induced groundwater denitrification beneath a manured field in southern Alberta, Canada. Applied Geochemistry, 23(8), 2146-2155.

McMichael, A. J., Woodruff, R. E., & Hales, S. (2006). Climate change and human health: present and future risks. The Lancet, 367(9513), 859-869.

Mennis, J. L., & Jordan, L. (2005). The distribution of environmental equity: Exploring spatial nonstationarity in multivariate models of air toxic releases. Annals of the Association of American Geographers, 95(2), 249-268.

Milly, P. C. D., Wetherald, R., Dunne, K. A., & Delworth, T. L. (2002). Increasing risk of great floods in a changing climate. Nature, 415(6871), 514-517.

Neshat, A., Pradhan, B., Pirasteh, S., & Shafri, H. Z. M. (2013). Estimating groundwater vulnerability to pollution using a modified DRASTIC model in the Kerman agricultural area, Iran. Environmental Earth Sciences, 1-13.

Olson, B. M., Miller, J. J., Rodvang, S. J., & Yanke, L. J. (2005). Soil and groundwater quality under a cattle feedlot in southern Alberta. Water quality research journal of Canada, 40(2), 131-144.

Panagopoulos, G. P., Antonakos, A. K., & Lambrakis, N. J. (2006). Optimization of the DRASTIC method for groundwater vulnerability assessment via the use of simple statistical methods and GIS. Hydrogeology Journal, 14(6), 894-911.

Pitman, A. J. (2005). On the role of geography in earth system science. *Geoforum*, *36*(2), 137-148.

Plant, R. E. (2012). Spatial data analysis in ecology and agriculture using R. CRC Press.

Pitt, M. (2008). Learning lessons from the 2007 floods. London: Cabinet Office.

Rahman, A. (2008). A GIS based DRASTIC model for assessing groundwater vulnerability in shallow aquifer in Aligarh, India. Applied Geography, 28(1), 32-53.

Richardson, H. Y., Nichols, G., Lane, C., Lake, I. R., & Hunter, P. R. (2009). Microbiological surveillance of private water supplies in England–The impact of environmental and climate factors on water quality. Water research, 43(8), 2159-2168.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(1), 127-142.

Smith, K. B., Smith, C. E., Forest, S. F., & Richard, A. J. (2007). A field guide to the wetlands of the boreal plains ecozone of Canada. *Ducks Unlimited Canada, Western Boreal Office: Edmonton, Alberta*.

Strahler, A. N. (1980). Systems theory in physical geography. *Physical Geography*, *1*(1), 1-27

Sliva, L., & Dudley Williams, D. (2001). Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water Research*, *35*(14), 3462-3472.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 46, 234-240.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. *Ecology*, *88*(11), 2766-2772.

Verhoeven, J. T., Arheimer, B., Yin, C., & Hefting, M. M. (2006). Regional and global concerns over wetlands and water quality. *Trends in ecology & evolution*, *21*(2), 96-103.

Wallender, E. K., Ailes, E. C., Yoder, J. S., Roberts, V. A., & Brunkard, J. M. (2013). Contributing Factors to Disease Outbreaks Associated with Untreated Groundwater. Groundwater.

Weiner, E. R. (2012). Applications of environmental aquatic chemistry: a practical guide. CRC Press.

Wheeler, A., Smith-Doiron, M., Xu, X., Gilbert, N., & Brook, J. (2008). Intra-urban variability of air pollution in Windsor, Ontario -- Measurement and modeling for human exposure assessment. Environmental Research, 106, 7 - 16.

Wilson, M. W., & Poore, B. S. (2009). Theory, practice, and history in critical GIS: Reports on an AAG panel session. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *44*(1), 5-16.

World Health Organization (Ed.). (2004). Guidelines for drinking-water quality: recommendations (Vol. 1). World Health Organization.

World Health Organization (WHO). (2014). Global Health Observatory (GHO): Urban population growth. Retrieved from: http://www.who.int/gho/urban_health/situation_trends/urban_population_growth_text/en/

Wu, W., & Zhang, L. (2013). Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. Applied Geography, 37, 52-62.

Zedler, J. B., & Kercher, S. (2005). Wetland resources: status, trends, ecosystem services, and restorability. *Annu. Rev. Environ. Resour.*, *30*, 39-74.

Appendix A: R Script for Regression Models

```
#Set up Working Directory
getwd()
setwd("C:/Users/Kristin/Desktop/Reg")
#Install Packages
library(car)
library("ctv")
library(maptools)
library(rgdal)
library(spdep)
owd <-getwd()
setwd(system.file("etc/shapes", package = "spdep"))
library(spgwr)
library(sp)
library(compoisson)
library(COMPoissonReg)
library(AER)
#Choose Variables for Regression
#Spearmans Correlation Coefficients
COR<-read.csv("C:/Users/Kristin/Desktop/Reg/TCEC_Master.csv", header=TRUE)
cor(COR, use="complete.obs", method="spearman")
#Refine Variables based on highest Correlation Coefficient
#TOTAL COLIFORM MODEL
#TC Aspatial Poisson GLM
TC<-read.csv("C:/Users/Kristin/Desktop/Reg/TC_Count_Master-April14.csv", header=TRUE)
attach(TC)
summary(TC)
#setup regression
colnames(TC)
modTC<-glm(formula = No_Pos ~ offset(log(Total_Sam)) + Rainfall + Elevation +
   Slope + Aspect + KSAT + X6400M_WatLines + X1600M_OL_Flood +
   X3200M_Mnwat + X400M_MJwater + X3200M_INwater + X400M_Floodway +
   X400M_Flood_FR + X400M_Developed + X800M_Agri2 + X400M_Forest +
   Near_AW + Farms+HA_Farms+Pop_Den+Dwell_Den, family = poisson(link = "log"), data
   = TC)
#Correlation of Coefficients
summary.glm(modTC,dispersion=NULL,correlation=T)
#Number of Farms and HA of farms highly correlated (0.85)- removed No of Farms bc less
correlated
#Test Dispersion
library(AER)
dispersiontest(modTC, alternative=c("greater"))
#Not over dispersed
dispersiontest(modTC, alternative=c("less"))
#Underdispered p= 2.38e-16, therefore must use COM
#Backwards Variable Selection Based on 95% CI
modTC<-com(formula = No_Pos ~ offset(log(Total_Sam)) + Rainfall + Elevation +
```

```
   Slope + Aspect + KSAT + X6400M_WatLines + X1600M_OL_Flood +
   X3200M_Mnwat + X400M_MJwater + X3200M_INwater + X400M_Floodway +
   X400M_Flood_FR + X400M_Developed + X800M_Agri2 + X400M_Forest +
   Near_AW +Farms+Pop_Den + Dwell_Den, family=poisson, na.action=na.exclude, data =
   TC)
#Final Model
modTC<-com(formula = No_Pos ~ offset(log(Total_Sam)) + Rainfall +
    X800M_Agri2, family = poisson(link = "log"), data = TC)
summary(modTC)
PID<-PID
TCpredict<-predict(modTC)
TCresid<-resid(modTC)
TCaspatial<-cbind(PID,TCpredict,TCresid)
write.csv(TCaspatial, "C:/Users/Kristin/Desktop/Reg/TCaspatial.csv")
#McFadden Test
mfTC<-1-LLTC/LLNullTC
mfTC
mfaTC<-1-((LLTC-8)/LLNullTC)
mfaTC
#Breusch Pagan Test for Heterscedascity
#lmtest packaged needed will add zoo
library(lmtest)
#bptest(formula, varformula = NULL, studentize = TRUE, data = list())
bp <- bptest(modTC)
bp
#BP = 5.3327, df = 2, p-value = 0.06951
#TC Spatial GWR Poisson GLM
TC_bw_aic <- gwr.sel(modTC, data = TC, coords = cbind(TC$X, TC$Y), method="aic")
TC_gauss <- gwr(modTC, data= TC, coords = cbind(TC$X, TC$Y), bandwidth = TC_bw_aic,
          hatmatrix = TRUE)
TC_gauss
#Getting T-Values
beta.rainfall<-TC_gauss$SDF$Rainfall
SE_rainfall<-TC_gauss$SDF$Rainfall_se
T_rainfall<-beta.rainfall/SE_rainfall
beta.Agri<-TC_gauss$SDF$X800M_Agri2
SE_Agri<-TC_gauss$SDF$X800M_Agri2_se
T_Agri<-beta.rainfall/SE_rainfall
TCcoeff<-cbind(beta.rainfall, SE_rainfall, T_rainfall, beta.Agri, SE_Agri, T_Agri)
write.csv(TCcoeff, "C:/Users/Kristin/Desktop/Reg/TC-GWRcoeff.csv")
results_TC_gauss<- as.data.frame(TC_gauss$SDF)
head(results_TC_gauss)
write.csv(results_TC_gauss, "C:/Users/Kristin/Desktop/Reg/TC_gauss_Results.csv")

#ECOLI MODEL
EC<-read.csv("C:/Users/Kristin/Desktop/Reg/EC_Count_Master-Apri14.csv", header=TRUE)
attach(EC)
summary(EC)
```

```
modEC<-glm(formula = EC_NO_POS ~ offset(log(EC_RESULT_COUNT)) + Rainfall +
   Elevation + Slope + KSAT + X1600M_OL_Flood + X800M_Mnwat +
   X400M_MJwater + X3200M_INwater + X800M_Floodway + X400M_Flood_FR +
   X800M_Developed + X800M_Agri2 + X400M_Forest + Near_AW +
   Farms + X1600M_PopDen + X1600M_DwellDen, family = poisson(link = "log"), data = EC)
cor(EC, use="complete.obs", method="spearman")
#Correlation of Coefficients
summary.glm(modEC,dispersion=NULL,correlation=T)
#Test Dispersion
dispersiontest(modEC, alternative=c("greater"))
#Not over dispersed
dispersiontest(modEC, alternative=c("less"))
#Not over dispersed, therefore Poisson Model is adequate
#Backwards Variable Selection Based on 95% CI
#Final TC Model
modEC<-glm(formula = EC_NO_POS ~ offset(log(EC_RESULT_COUNT)) +
    X3200M_INwater + X800M_Floodway + X400M_Flood_FR +
   Farms, family = poisson(link = "log"), data = EC)
summary(modEC)
PID<-PID
ECpredict<-predict(modEC)
ECresid<-resid(modEC)
ECaspatial<-cbind(PID,ECpredict,ECresid)
write.csv(ECaspatial, "C:/Users/Kristin/Desktop/Reg/ECaspatial.csv")
#McFadden Test
mfEC<-1-LLEC/LLNullEC
mfEC
mfaEC<-1-((LLEC-8)/LLNullEC)
mfaEC
#Bresush-Pagan Test for Heteroscedasticity
bp <- bptest(modEC)
bp
#BP = 78.4048, df = 4, p-value = 3.792e-16
#Spatial GWR Poisson GLM
EC_bw_aic<- gwr.sel(modEC, data = EC, coords = cbind(EC$X, EC$Y), method="aic")
EC_gauss<- gwr(modEC, data= EC, coords = cbind(EC$X, EC$Y), bandwidth = EC_bw_aic,
          hatmatrix = TRUE, predictions=TRUE)
EC_gauss
#Getting T-Vaules
beta.INwater<-EC_gauss$SDF$X3200M_INwater
SE_INwater<-EC_gauss$SDF$X3200M_INwater_se
T_INwater<-beta.INwater/SE_INwater
beta.Floodway<-EC_gauss$SDF$X800M_Floodway
SE_Floodway<-EC_gauss$SDF$X800M_Floodway_se
T_Floodway<-beta.Floodway/SE_Flooway
beta.Flood_FR<-EC_gauss$SDF$X400M_Flood_FR
SE_Flood_FR<-EC_gauss$SDF$X400M_Flood_FR_se
T_Flood_FR<-beta.Flood_FR/SE_Flood_FR
```

```
beta.Farms <-EC_gauss$SDF$Farms
SE_Farms<-EC_gauss$SDF$Farms_se
T_Farms<-beta.Farms/SE_Farms
ECcoeff<-cbind( beta.INwater,SE_INwater,T_INwater, beta.Flood_FR, SE_Flood_FR, T_Flood_FR,
          beta.Floodway, SE_Floodway, T_Floodway, beta.Farms, SE_Farms, T_Farms)
write.csv(ECcoeff, "C:/Users/Kristin/Desktop/Reg/EC-GWRcoeff.csv")
#Getting Coefficients
results_EC_gauss<- as.data.frame(EC_gauss$SDF)
head(results_EC_gaus)
write.csv(results_EC_gauss, "C:/Users/Kristin/Desktop/Reg/EC_log_Results.csv")
```