

UNCOVERING TRENDS OF *E. COLI* IN PRIVATE WELLS

UNCOVERING TRENDS OF *E. COLI* TRANSPORT IN PRIVATE DRINKING
WATER WELLS: AN ONTARIO CASE STUDY

By KATIE WHITE, B. Eng

A THESIS

SUBMITTED TO THE DEPARTMENT OF CIVIL ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

DOCTOR OF PHILOSOPHY

MCMASTER UNIVERSITY © COPYRIGHT BY KATIE WHITE

SEPTEMBER 2023

DOCTOR OF PHILOSOPHY (2023) McMASTER UNIVERSITY

(CIVIL ENGINEERING) HAMILTON, ONTARIO, CANADA

TITLE: UNCOVERING TRENDS OF *E. COLI* FATE AND TRANSPORT IN PRIVATE
DRINKING WATER WELLS: AN ONTARIO CASE-STUDY

AUTHOR: KATIE WHITE, B.ENG. (McMASTER UNIVERSITY)

SUPERVISORS: DR. SARAH DICKSON-ANDERSON

DR. CORINNE SCHUSTER-WALLACE

NUMBER OF PAGES: XXIII, 212

Lay Abstract

There are millions of people globally relying on private groundwater to access drinking water. Unfortunately, these wells come with many challenges including a lack of government regulations, and limited resources for maintenance, management, and protection. These challenges also result in an increased risk of illness in private well users.

Groundwater research is often limited by lack of numerical data making it extremely difficult to understand how groundwater and contaminants are transported. This research utilizes a large dataset with 795,023 contamination observations, 253,136 unique wells, and over 33 variables (i.e., well and aquifer characteristics, human behaviours, weather-related) across Ontario, Canada between 2010 and 2017.

The work in this thesis utilizes a data-driven approach, using various machine learning techniques combined with subject matter expertise, to uncover trends and insights into when and how contamination events occur in private wells, to both inform policy makers and empower well users.

Abstract

Millions of Canadians rely on private groundwater wells to access drinking water, which presents many challenges including lack of government regulations, and limited resources for maintenance, monitoring, management, and protection. These challenges result in an increased risk of acute gastrointestinal illness in private well users. The goal of this work is to improve the understanding of drivers of *E. coli* fate and transport in groundwater using a data-driven approach to better inform well owners and policy makers. Specifically, the objectives include: exploratory analysis of the physical and human drivers of private well contamination; advancing the understanding of the relationships between land use-land cover and *E. coli* presence in wells; assessment of rainfall intermittency patterns as a driver of contamination, as an alternative to standard lag times; and, the development of data-driven explanatory models for *E. coli* contamination in private wells that move towards a novel coupled-systems approach.

This research utilizes a large dataset with 795,023 contamination observations, 253,136 unique wells, and over 33 variables (i.e., microbiological, hydrogeological, well characteristic, meteorological, geographical, and testing behaviour) across Ontario, Canada between 2010 and 2017. Data used includes the Well Water Information Database, Well Water Information System, Daymet, Provincial Digital Elevation Model, Ontario Land Cover Compilation, Southern Ontario Land Resource Information System, and Roads Network. Data analysis methods range from univariate and bivariate analyses to supervised and unsupervised machine learning techniques, including regression, clustering, and classification.

This work has contributed important understandings of the relationships between *E. coli* contamination and well and aquifer characteristics, seasonality, weather, and human behaviour. Specifically, increased well depth reduced, but did not eliminate, likelihood of contamination; wells completed in consolidated material increased likelihood of contamination; the most significant driver of contamination was identified as land use - land cover, which was categorized into four classes of *E. coli* contamination potential for wells, ranging from very high to low; latitude was found to drive seasonality and consequent weather patterns, leading to the creation of geographically-based seasonal models; liquid water (i.e., rainfall, snow melt) was a key driver of contamination, where increased water generally increased presence of *E.coli* while causing decreasing prevalence; time of year, not habit, drove user testing, generally peaking in July; and, a surrogate measure of well user stewardship was identified as driving time to closest drop-off location. Further, this work has contributed methodological advancements in identifying drivers of groundwater contamination including: utilizing literature confidence ratings alongside regression analyses to supply strategic direction to policy makers; demonstrating the value of large datasets in combination with innovative machine learning techniques, and subject matter expertise, to identify improved physically-based understandings of the system; and highlighting the need for coupled-systems approaches as physical models alone do not capture human behaviour-based factors of contamination.

Acknowledgments

I would like to express my deepest gratitude to Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace for their guidance, support, mentorship, and patience over the course of my studies. This achievement would not have been possible without them.

I would also like to thank my supervisory committee, Dr. Wael El-Dakhakhni and Dr. Anna Majury, for their time and support.

I am also thankful for Dr. Ahmed Yosri, Christine Homuth, and Dr. Ahmed Siam for sharing their time, knowledge, and support when I needed it.

Finally, I want to express my gratefulness to my wife and family who loved, supported, and encouraged me throughout this endeavor.

Contents

| | |
|---|-------|
| Lay Abstract..... | iii |
| Abstract | iv |
| Acknowledgments..... | v |
| List of Figures | xii |
| List of Tables | xviii |
| Notation and Abbreviations | xxiii |
| Declaration of Academic Achievement | xxiv |
| CHAPTER 1 Introduction | 1 |
| 1.1 Private Wells Users and Groundwater Quality | 2 |
| 1.2 Fate and Transport Mechanisms Driving Groundwater Contamination | 3 |
| 1.2.1 Seasonality, Weather, and Climate | 5 |
| 1.2.2 <i>E. coli</i> Sources | 7 |
| 1.2.3 Hydrogeology | 8 |
| 1.2.4 Well Characteristics | 9 |
| 1.2.5 Human Behaviour | 10 |
| 1.3 Site Description – Ontario, Canada..... | 11 |
| 1.4 Research Objectives | 15 |

| | | |
|--|---|----|
| 1.4.1 | Thesis Outline | 16 |
| 1.5 | Works Cited..... | 18 |
| CHAPTER 2 Exploration of <i>E. coli</i> contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset | | |
| | | 28 |
| 2.1 | Abstract | 31 |
| 2.2 | Introduction | 32 |
| 2.3 | Methods | 36 |
| 2.3.1 | Dataset..... | 36 |
| 2.3.2 | Data Processing..... | 36 |
| 2.3.3 | Statistical Analysis..... | 38 |
| 2.4 | Results and Discussion..... | 44 |
| 2.4.1 | Seasonal Drivers (Driver 1 in Figure 2-1) | 46 |
| 2.4.2 | Hydrogeological Drivers (Driver 2 in Figure 2-1)..... | 49 |
| 2.4.3 | Well Characteristics (Driver 3 in Figure 2-1) | 51 |
| 2.4.4 | Informed Physical Model..... | 52 |
| 2.4.5 | Testing Practices (Driver 4 in Figure 2-1) | 53 |
| 2.5 | Study Limitation..... | 58 |
| 2.6 | Conclusion..... | 59 |

| | | |
|--|--|-----|
| 2.7 | Works Cited..... | 62 |
| CHAPTER 3 Converting Land Use – Land Cover to <i>E. coli</i> Contamination Potential | | |
| | Classes for Groundwater Wells: Utilizing a Large Ontario-based Dataset | 67 |
| 3.1 | Abstract | 70 |
| 3.2 | Introduction | 71 |
| 3.3 | Methods | 72 |
| 3.3.1 | Data Sources | 72 |
| 3.3.2 | LULC Categorization..... | 73 |
| 3.3.3 | <i>E. coli</i> Contamination Potential Estimation..... | 77 |
| 3.4 | Results and Discussion..... | 79 |
| 3.5 | Policy Implications..... | 84 |
| 3.6 | Conclusions | 85 |
| 3.7 | Works Cited..... | 87 |
| CHAPTER 4 Exploring the role of rainfall intermittency on <i>E. coli</i> contamination | | |
| | events in private wells: an Ontario case study | 92 |
| 4.1 | Abstract | 95 |
| 4.2 | Introduction | 95 |
| 4.3 | Methods..... | 100 |
| 4.3.1 | Datasets | 100 |

| | | |
|--|--|-----|
| 4.3.2 | Data Processing..... | 100 |
| 4.3.3 | Data Analytics Techniques | 103 |
| 4.4 | Results and Discussion..... | 113 |
| 4.4.1 | Improving on Traditional Lag Times..... | 113 |
| 4.4.2 | Identification of Rainfall Intermittency Patterns | 115 |
| 4.4.3 | Identification of Meaningful Rainfall Intermittency Clusters | 117 |
| 4.4.4 | Addition of Variables to Improve Accuracy..... | 125 |
| 4.4.5 | Regression Techniques | 126 |
| 4.5 | Study Limitations | 130 |
| 4.6 | Conclusion..... | 131 |
| 4.7 | Works Cited..... | 132 |
| CHAPTER 5 Towards a Coupled-Systems Approach for the Exploration of <i>E. coli</i> | | |
| | Contamination in Private Drinking Water Wells..... | 139 |
| 5.1 | Abstract | 141 |
| 5.2 | Introduction | 142 |
| 5.3 | Methods..... | 147 |
| 5.3.1 | Datasets | 147 |
| 5.3.2 | Data Processing..... | 148 |
| 5.3.3 | Statistical Analyses | 164 |

| | | |
|-----------|---|-----|
| 5.4 | Results and Discussion..... | 168 |
| 5.4.1 | Exploratory Analyses..... | 168 |
| 5.4.2 | Informed Coupled-systems Models | 181 |
| 5.5 | Study Limitations | 191 |
| 5.6 | Conclusion..... | 191 |
| 5.7 | Works Cited..... | 194 |
| CHAPTER 6 | Conclusions and Recommendations..... | 202 |
| 6.1 | Conclusions | 203 |
| 6.1.1 | Exploration of private well dataset utilizing innovative machine learning approaches..... | 205 |
| 6.1.2 | Introduction of data-driven LULC <i>E. coli</i> contamination potential class mapping..... | 206 |
| 6.1.3 | Exploration of wet-dry patterning as a driver of <i>E. coli</i> contamination events, moving beyond current standard rainfall lag periods..... | 207 |
| 6.1.4 | Development of geographically-driven, seasonally-based models, moving toward a coupled-systems approach, to explain <i>E. coli</i> contamination in private wells | 207 |
| 6.2 | Study Limitations | 208 |
| 6.3 | Recommendations for Future Research | 210 |
| 6.4 | Works Cited..... | 212 |

| | | |
|-----------|----------------------------|-----|
| CHAPTER A | Appendices | 213 |
| A.2 | Chapter 2 Appendices | 213 |
| A.2.1 | Data Processing..... | 213 |
| A.2.2 | Table and Figures..... | 218 |
| A.2.3 | Works Cited | 243 |
| A.3 | Chapter 3 Appendices | 244 |
| A.3.1 | Tables and Figures | 244 |
| A.4 | Chapter 4 Appendices | 266 |
| A.4.1 | Figures and Tables | 266 |
| A.5 | Chapter 5 Appendices | 272 |
| A.5.1 | Figures and Tables | 272 |

List of Figures

| | |
|--|----|
| Figure 1-1: Transport mechanisms driving <i>E. coli</i> presence in private wells considering a coupled-systems approach explored in this work. | 4 |
| Figure 1-2: Ontario maps depicting: A – Major LULC categories in Ontario (White et al. 2022), B – Population Density (Statistics Canada, 2016), and C – Surficial Geology (Thompson et al. 2000). | 14 |
| Figure 2-1: Fate and transport mechanisms driving <i>E. coli</i> concentrations in private wells (i.e., contamination risk) considering a coupled-systems approach, adapted from (Di Pelino et al., 2019), where numbered text represent drivers used to develop explanatory models. | 34 |
| Figure 2-2: Summary of explanatory variables across “best” models for each driver (Figure 2-1). | 45 |
| Figure 2-3: Percentage of individual wells versus number of times tested over the eight-year period given an initial sample that was “no significant evidence” (73% of wells), “no result” (6%), “may be unsafe” (13%), or “unsafe to drink” (8%). Insets show a) under two times per year threshold (i.e., 1-15 tests), and b) at or over two times per year threshold tested tail (i.e., 16-446 tests) of this curve, respectively. | 56 |
| Figure 2-4: Occurrence of private well testing and adverse results over the study period, where non-detects are defined as 0 CFU/100mL, “All <i>E. coli</i> ” represents the summation of the three <i>E. coli</i> concentration categories. Vertical bars are extensions of peak adverse points each year. | 57 |

| | |
|---|-----|
| Figure 3-1: Map of Ontario based on newly defined major LULC categories (White et al., 2022). | 76 |
| Figure 3-2: Map of <i>E. coli</i> contamination potential classes for Ontario based on derived raster data (White et al., 2022). | 83 |
| Figure 4-1: Summary output of random forest analysis exploring adverse <i>E. coli</i> observations for conditions from day n to day n-60. Figure A depicts the ranking order of antecedent days, where lower ranks represent more important variables. The dotted line highlights the highest mean rank of the antecedent days considered (i.e., n to n-9, n-22, and n-45). Figure B depicts the change in accuracy after the addition of one further antecedent day (i.e., n to n-45 represents the model accuracy when days n to n-45 are included in the model). | 114 |
| Figure 4-2: Summary of the top 10% (based on frequency) of dry-wet rainfall patterns leading to <i>E. coli</i> contamination in wells. A - represents the prediction accuracy of each pattern in predicting <i>E. coli</i> contamination. B - represents the dry-wet pattern over the 36-days preceding a contamination event, where blue represents wet days and yellow represents dry days. C – represents the maximum and median total rainfall found within a specific pattern, across all pattern occurrences. Note: the x-axis represents the order of patterns from most to least frequent. | 116 |
| Figure 4-3: Box A depicts results from clustering showing rainfall pattern clusters resulting in <i>E. coli</i> contamination in wells. Box B depicts a summary of these 6 clusters based on the breakdown of LULC, Stratigraphy, and <i>E. coli</i> contamination severity, where the y-axis represents the percentage of the cluster. | 119 |

| | |
|--|-----|
| Figure 5-1: Summary of key elements of coupled-systems model explored in this work to explain <i>E. coli</i> contamination events in private wells..... | 146 |
| Figure 5-2: Summary of Ontario seasons (b) based on spatial distributions of Hardiness Zones (a; derived by McKenney et al. (2014)). | 158 |
| Figure 5-3: Overview of dependent and independent variable subsets explored in this work. | 167 |
| Figure 5-4: Summary of exploratory variable sets, identifying individual variables explored and variables that were deemed important in the informed seasonal coupled-systems models. | 169 |
| Figure 5-5: Summaries of trends between minimum driving times and testing frequency (a), testing hotspots (b), and testing coldspots (c)..... | 180 |
| Figure 6-1: Summary of work completed in this thesis through the exploration of the key drivers of <i>E. coli</i> contamination in private wells. Primary variables represent those with consensus across three regression models (Chapter 5) and secondary variables represent those with consensus across two regression models (Chapter 5)..... | 204 |
| Figure A.2-1: Sensitivity analysis of "most explanatory" seasonality variables considering seasons. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent one standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. | 231 |

| | |
|---|-----|
| Figure A.2-2: Statistical significance of all seasonality variables in the seasonal assessment model when compared to one another. Statistically significant is defined as $p \leq 0.1$ | 232 |
| Figure A.2-3: Sensitivity analysis of "most explanatory" seasonality variables considering months. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. Note the absence of April, this was removed as it was deemed insignificant by the models. | 233 |
| Figure A.2-4: Statistical significance of all seasonality variables in the monthly assessment model when compared to one another. Statistical significance is set to $p \leq 0.1$ | 234 |
| Figure A.2-5: Sensitivity analysis of "most explanatory" seasonality variables considering years. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. Note the absence of 2010, this was not considered explanatory in the models. | 235 |
| Figure A.2-6: Statistical significance of all seasonality variables found within the yearly assessment model when compared to one another. Statistical significance is set to $p \leq 0.1$ | 236 |
| Figure A.2-7: Sensitivity analysis of "most explanatory" hydrogeological variables. Red horizontal line represents a coefficient of zero, black points represent the mean of | |

| | |
|--|-----|
| the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. | 237 |
| Figure A.2-8: Statistical significance of all hydrogeological variables when compared to one another. Statistical significance is set to $p \leq 0.1$ | 238 |
| Figure A.2-9: Sensitivity analysis of "most explanatory" informed physical variables. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. | 239 |
| Figure A.2-10: Statistical significance of all informed physical variables when compared to one another. Statistical significance is set to $p \leq 0.1$ | 240 |
| Figure A.2-11: Sensitivity analysis of "most explanatory" well testing variables. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. | 241 |
| Figure A.2-12: Statistical significance of all testing practice variables when compared to one another. Statistical significance is set to $p \leq 0.1$ | 242 |
| Figure A.4-1: Summary of original 14 clusters based on spectral analysis. | 266 |
| Figure A.4-2: Fraction of explained variance of the model based on the number of clusters allowed. | 267 |

| | |
|--|-----|
| Figure A.5-1: Closest driving time bins of closest drop-off location versus first test message received..... | 272 |
| Figure A.5-2: Closest driving time bins of closest drop-off location versus year of test. | 273 |
| Figure A.5-3: Closest driving time bins of closest drop-off location versus month of test. | 273 |
| Figure A.5-4: Closest driving time bins of closest drop-off location versus weekday of test. | 274 |

List of Tables

| | |
|--|-----|
| Table 2-1: Description of variables contained within the merged WWIS and WWTD dataset, including sub-classifications derived for the purpose of these analyses.. | 37 |
| Table 3-1: Summary of major LULC categories. | 74 |
| Table 3-2: Summary of regression models exploring varying subsets of data including and excluding <i>E. coli</i> non-detect water sample results. Ordering of LULC categories (i.e., 1 -10) is based on the averaged coefficient of 10 iterations of cross validated regression analyses, where one results in the highest adverse <i>E. coli</i> contamination potential and 10 the least adverse potential. | 79 |
| Table 3-3: Summary of LULC <i>E. coli</i> contamination potential and respective level of confidence with respect to literature. Where ★★★★★ represents high confidence, and ★ represents low confidence. | 81 |
| Table 4-1: Summary of literature exploring the relationship between rainfall and its respective outcome (i.e., water quality or human health). | 97 |
| Table 4-2: Summary of datasets and variables used in this study. | 100 |
| Table 4-3: Summary of all regression analysis variables, based on classification model finding of n-36 days representing the defined antecedent time period for <i>E. coli</i> presence prediction. | 106 |
| Table 4-4: Summary of clusters based on adverse and non-adverse <i>E. coli</i> observations. | 121 |

| | |
|---|-----|
| Table 4-5: Summary of regression technique outputs for all <i>E. coli</i> and adverse <i>E. coli</i> observations. | 129 |
| Table 5-1: Summary of datasets and variable used in this study. | 147 |
| Table 5-2: Summary of all variables. | 150 |
| Table 5-3: Comparison of well density categories and LULC <i>E. coli</i> contamination classes. | 175 |
| Table 5-4: Summary of key trends identified from the informed coupled-systems regression analyses in the summer months. | 181 |
| Table 5-5: Summary of key trends identified from the informed coupled-systems regression analyses in the shoulder months. | 185 |
| Table 5-6: Summary of key trends identified from the informed coupled-systems regression analyses in the winter months. | 188 |
| Table 5-7: Summary of key trends identified from the informed coupled-systems regression analyses in all months. | 190 |
| Table A.2-1: Classification of bedrock types. | 218 |
| Table A.2-2: Classification of permeability. | 219 |
| Table A.2-3: NA sample status explanations. | 220 |
| Table A.2-4: Variable descriptions for regression analyses. | 221 |
| Table A.2-5: Statistical summary of seasonality, hydrogeological, well characteristics, and testing practices model variables used in regression analyses. | 223 |

| | |
|---|-----|
| Table A.2-6: Statistical significance testing for bottom stratigraphy analysis comparing unconsolidated and consolidated categorization. Statistical significance is set to $p \leq 0.05$ | 228 |
| Table A.2-7: Top ten “most interesting” rules based on an association rule analysis including adverse <i>E. coli</i> concentrations and rock type. Where Category 1 is 1-10 CFU/100mL, Category 2 is 11-50 CFU/100mL, Category 3 is 51+ CFU/100mL, and water is “Unsafe to Drink (target)”. | 228 |
| Table A.2-8: Statistical significance testing for well depth analysis comparing shallow and deep wells. Statistical significance is set to $p \leq 0.05$. Where non-detects are defined as 0 CFU/100mL, category 1 as 1-10 CFU/100mL, category 2 as 11-50 CFU/100mL, and category 3 as 51+ CFU/100mL..... | 229 |
| Table A.2-9: Summary of user testing decay curves based on first test message received. Two decay equations are summarized per first test message received, curve characteristics for total tests 1 through 15, and curve characteristics for total tests 16+. | 229 |
| Table A.2-10: Exhaustive list of categorical bottom stratigraphy variable levels. Green represents present, red represents absent. | 230 |
| Table A.3-1: Tables adapted from Caldeira et al. (2019) and Mastrandrea et al. (2010). A – agreement is the qualitative measure of the level of concurrence in the literature, B – evidence is the amount of work supporting a finding, C and D– level of confidence is determined by a summary of the evidence and agreement levels, | |

| | |
|--|-----|
| where darker blocks (HA&RE) represent a higher confidence than lighter blocks (LA&LE) (Caldeira et al., 2019; Mastrandrea et al., 2010). | 244 |
| Table A.3-2: Summary table comparing ordered LULCs based on impact to <i>E. coli</i> contamination potential utilizing regression analyses and corresponding literature. Icons in “Key Process Notes” column are defined in Table A.3.1 to depict likelihood, evidence, and level of confidence of findings based on regression analyses and literature. | 245 |
| Table A.3-3: Summary of intercepts and coefficients of regression models exploring LULC (independent variable) and <i>E. coli</i> contamination potential (dependent variable). | 265 |
| Table A.4-1: Summary of clusters based on adverse and non-adverse <i>E. coli</i> observations from n to n-5 days. | 268 |
| Table A.4-2: Summary of clusters based on adverse and non-adverse <i>E. coli</i> observations from n to n-16 days. | 270 |
| Table A.5-1: Summary of SWE inclusion criteria based on Ontario Plant Hardiness Zones. | 275 |
| Table A.5-2: Summary of impact to <i>E. coli</i> in “summer” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable. | 276 |
| Table A.5-3: Summary of impact to <i>E. coli</i> in “shoulder” models if one variable was increased at a time, only model deemed significant variables record an impact. | |

“Cat” represents the use of categorical variables, “Cont” represents the use of continuous variable. 279

Table A.5-4: Summary of impact to *E. coli* in “winter” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable..... 281

Table A.5-5: Summary of impact to *E. coli* in “all seasons” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable..... 284

Notation and Abbreviations

AGI – Acute Gastrointestinal Illness

ARA - Association Rule Analysis

E. coli – *Escherichia Coli*

GAMLSS – Generalized Additive Model for Location, Space, Scale

LASSO – Least Absolute Shrinkage and Selection Operator

LULC – Land Use-Land Cover

MARS – Multivariate Adaptive Regression Splines

OLCC - Ontario Land Cover Compilation

PDEM - Provincial Digital Elevation Model

SOLRIS - Southern Ontario Land Resource Information System

WWIS - Well Water Information System

WWTD - Well Water Testing Database

Declaration of Academic Achievement

This thesis has been prepared in accordance with the regulations for a sandwich thesis format as outlined by McMaster University's School of Graduate Studies. As such, manuscripts that comprise Chapters 2, 3, 4, and 5 have been co-authored. The work reported in this thesis was undertaken from January 2018 to August 2023.

Chapter 2: Exploration of E. coli contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset, by K. White, S. Dickson-Anderson, A. Majury, K. McDermott, P. Hynds, S.R. Brown, and C. Schuster-Wallace, Journal of Water Research, 197, doi: 10.1016/j.watres.2021.117089, 2021. (With permission from publisher)

Conceptualization of this work was conducted by Katie White, Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace. Data curation and validation was conducted by Katie White and Kevin McDermott. Data analysis was conducted by Katie White. Editing of the work was conducted by Dr. Sarah Dickson-Anderson, Dr. Corinne Schuster-Wallace, Dr. Anna Majury, Dr. Paul Hynds, and Dr. Stephen R. Brown.

Chapter 3: Converting Land Use – Land Cover to E. coli Contamination Potential Classes for Groundwater Wells: Utilizing a Large Ontario-based Dataset, by K. White, C. Schuster-

Wallace, and S. Dickson-Anderson, Journal of Applied Spatial Analysis and Policy, Submitted on February 23, 2023, Under Review, ASAP-D-23-00035.

Maps of Major Land Use - Land Cover Categories and Respective E. coli Contamination Potential Classes for Groundwater Wells in Ontario, by K. White, S. Dickson-Anderson, and C. Schuster-Wallace, Federated Research Data Repository, doi: 10.20383/102.0530. 2023.

Conceptualization of this work was conducted by Katie White, Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace. Data curation, validation, and analysis was conducted by Katie White. Editing of the work was conducted by Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace.

Chapter 4: Exploring the role of rainfall intermittency on E. coli contamination events in private wells: an Ontario case study, by K. White, A. Yosri, S. Dickson-Anderson, and C. Schuster-Wallace, To be submitted to Journal of Weather and Climate Extremes by September 2023.

Conceptualization of this work was conducted by Katie White, Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace. Data curation and validation was conducted by Katie White. Data analysis was conducted by Katie White and Dr. Ahmed Yosri. Editing of the work was conducted by Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace.

Chapter 5: Towards a Coupled-Systems Approach for the Exploration of E. coli Contamination in Private Drinking Water Wells, by K. White, C. Schuster-Wallace, and S. Dickson-Anderson, To be submitted to Water Research by October 2023.

Conceptualization of this work was conducted by Katie White, Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace. Data curation, validation, and analysis was conducted by Katie White. Editing of the work was conducted by Dr. Sarah Dickson-Anderson and Dr. Corinne Schuster-Wallace.

CHAPTER 1 INTRODUCTION

1.1 PRIVATE WELLS USERS AND GROUNDWATER QUALITY

It is widely believed that raw water obtained from a groundwater source is safer than surface water sources due to the natural filtration that occurs, which can attenuate particulate contaminants. While these protective attenuation mechanisms are often effective, groundwater may still become contaminated and therefore pose health risks (Charrois, 2010; Murphy et al., 2017). Although not all microbiological constituents found in groundwater are hazardous to human health, those that are, known as waterborne pathogens, are typically fecal based, originating from livestock, wild animals, and humans (Medema et al., 2004).

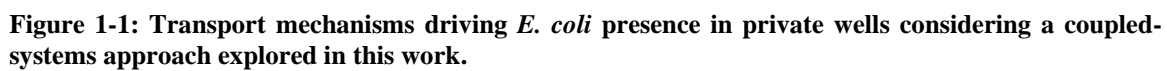
When infected by a waterborne pathogen, a person may experience a variety of symptoms that are collectively referred to as acute gastrointestinal illness (AGI). In general, *Norovirus*, *Cryptosporidium* spp., *Giardia lamblia*, *Campylobacter jejuni*, *Salmonella enterica*, and pathogenic *E. coli* contribute the largest portion of all waterborne illnesses in North America (Murphy et al., 2017; Soller et al., 2010). When assessing the impact of groundwater on public health, *E. coli* has been deemed the primary indicator of drinking water quality and safety due to its presence in the faeces of all mammals, inexpensive testing methods, and its survivability in a range of drinking water conditions (Edberg et al., 2000). While not all *E. coli* is pathogenic, the strains of *E. coli* that are can cause gastrointestinal upset and, in more serious cases, bloody diarrhea and haemolytic uremic syndrome (HUS) (Soller et al., 2010). In North America, children under the age of ten are most likely to contract HUS after being infected with *E. coli*, which can cause serious

complications such as kidney failure, stroke, inflammatory colitis, or heart problems (Bhandari & Sedhai, 2022).

Private well users rely on groundwater for their domestic water supply and these private wells are unregulated by the government in Canada. This leaves well stewardship to the well's user, meaning that they face a greater health risk than those who rely on municipally treated water. An estimated 4.1 million people living in Canada rely on privately owned and maintained groundwater supplies (Murphy et al., 2016) that are outside government regulation and oversight (Government of Ontario, 2018). It is estimated that three million people rely on private wells for domestic water in Ontario alone (Kreutzwiser et al., 2010). Historically, approximately 66% of Canadian waterborne disease outbreaks originated from privately owned wells or small municipal systems (Schuster et al., 2005), which may be due to limited resources for maintenance, management, and protection strategies (Rivera, 2017).

1.2 FATE AND TRANSPORT MECHANISMS DRIVING GROUNDWATER CONTAMINATION

While there are many fate and transport mechanisms that drive groundwater contamination, the main factors explored in this work include weather and climate, *E. coli* sources and concentrations, hydrogeological conditions, well characteristics, and human behaviour (Figure 1-1).



1.2.1 Seasonality, Weather, and Climate

Seasonality is the cyclic and predictable fluctuations in the environment driven by the earth's orbit around the sun making it subject to large-scale geographic factors (Lisovski et al., 2017). Local weather and climate (e.g., temperature and precipitation) driven by seasonality is an important consideration of contaminant fate and transport. Some ways in which seasonality drives the fate and transport of groundwater contaminants include variation in precipitation leading to seasonal changes in groundwater flow as well as surface runoff and variation in temperature leading to seasonal changes in precipitation form and overburden permeability (Hayashi et al., 2003; Pejman et al., 2009).

Due to the large spatial extent of Ontario, Canada, the study area of this work, seasonality exists in four distinct and predictable seasons (winter, spring, summer, fall/autumn) with each offering their own nuances when considering *E. coli* fate and transport in groundwater. Ontario winters are cold, with precipitation often falling as snow. Springs are often more unpredictable with temperatures beginning to increase above zero resulting in frozen liquid, and mixed precipitation occurring in conjunction with the thaw of snowpacks and river and lake ice. The large areal extent also means that there is significant variation in the timing of freeze and breakup when comparing latitudes (OMAFRA, 2020). Summers are often hot and humid, particularly in southern Ontario, with the potential for significant summer storms generating substantial runoff events. Fall/autumn seasons are often cool with precipitation falling as rain or snow. This uncertainty and complexity results in an increased risk to private well users, with heavy precipitation events, snow melt, and wet-dry patterning tied to acute gastrointestinal illness

(AGI) in humans (de Roos et al., 2020; Kraay et al., 2020; Namugize et al., 2018; Whitman et al., 2008). Hereafter, spring and autumn will be considered together as the shoulder seasons experiencing mixed precipitation. Again, due to the large spatial extent of the study area, there is a need to identify the onset and cessation of the four distinct seasons based on the geographic location within the province.

Rainfall is a key transport mechanism of groundwater contamination that mainly occurs in Ontario's summer and shoulder seasons. Rainfall events have been tied to an increased likelihood of drinking water source contamination, which is associated with the occurrence of diarrhea in consumers (Carlton et al., 2014; de Roos et al., 2020; Gleason & Fagliano, 2017; Jagai et al., 2015; Kraay et al., 2020; Levy et al., 2016). More specifically, rainfall after a dry period has been linked to an increase in diarrhea occurrence, while rainfall after a wet period was found to be protective against diarrhea (Levy et al., 2016). The higher likelihood post dry period has been associated with the "concentration effect", a flushing of pathogens that have accumulated on the land (Kraay et al., 2020), remobilizing them through the surface and groundwater systems (Levy et al., 2016). Conversely, the lower likelihood caused by rainfall after a wet period is attributed to the "dilution effect", in which pathogens on the surface and in the groundwater have been mobilized and are diluted by recent, prior rainfalls (Levy et al., 2016). These relationships demonstrate that intermittency, a core characteristic of rainfall patterns (Trenberth et al., 2017), plays an important role in introducing waterborne pathogens to drinking water sources, including wells.

In Ontario's winter season in particular, precipitation often falls as snow and ice and is more likely to be present as temperatures fall below 0°C. Frozen ground and snow presence can be protective against groundwater contamination as reduced infiltration and groundwater flow, due to saturated or partially saturated soil freezing and creating ice filled pores, resulting in decreased *E. coli* occurrence and transport (Charron et al., 2004; Hayashi et al., 2003). Conversely, during snowmelt events (typical in shoulder seasons) *E. coli* becomes remobilized and groundwater transport rates rapidly increase due to increased soil saturation (Jonsson & Agerberg, 2015; Khaleel et al., 1978). These relationships demonstrate the importance of understanding the change in snow presence, whether a melt or accumulation event is occurring, which will either protect against or promote *E. coli* transport.

1.2.2 *E. coli* Sources

E. coli contamination of groundwater requires the presence of an *E. coli* source. Some typical sources of *E. coli* include leaking septic tanks, manure-based fertilizers, and domestic and wild animal waste. Land Use – Land Covers (LULC) classes are useful proxies because they can be used as a proxy for the human and animal activity in an area, which are indicative of *E. coli* sources, as well as represent potential transport characteristics based on the typical physical environment (Dusek et al., 2018; Gregory et al., 2019; Jabbar et al., 2019). Pastoral and agricultural LULCs are associated with a high prevalence of *E. coli* in groundwater as a result of fertilizing fields with manure and the presence of livestock (Dusek et al., 2018; Jabbar et al., 2019; Petersen & Hubbart, 2020). Urban LULCs are associated with a high prevalence of *E. coli* in groundwater due to the

higher occurrence of entities like domestic animals, leaking septic tanks, and presence of wastewater treatment plants (Hua, 2017; Jabbar et al., 2019; Paule-Mercado et al., 2016). Conversely, some LULCs are believed to effectively remove *E. coli*, such as wetlands, due to their vegetation with antimicrobial properties and soils with natural filtration properties (Dordio et al., 2008). However, consideration and assessment of other LULCs, such as scrubland, disturbance, and bedrock, also typically associated with lower *E. coli* prevalence, are not present in the literature, making it difficult to achieve a strong consensus.

1.2.3 Hydrogeology

One of the primary drivers of pathogen transport into a well is the local hydrogeology. More specifically, stratigraphy can be used to explain and better understand the mechanisms or pathways of contaminant transport in groundwater (Hynds et al., 2014). At the highest level, stratigraphy, the types of rock from the various layers that make up the overall geology of a region, can be categorized as consolidated or unconsolidated material. Generally, the presence of unconsolidated material is considered protective against *E. coli* contamination in groundwater wells due to their slower transport rates and the opportunity for natural attenuation to occur (Wiebe et al., 2021). Bedrock (consolidated) wells with minimal overburden (unconsolidated material near the surface) are more likely to become contaminated due to the lack of soil available to filter pathogens before they reach potential fractures or channels often associated with bedrock (Conboy & Goss, 2000; Latchmore et al., 2020). Considering consolidated stratigraphy more closely, wells located in sedimentary rock (e.g., limestone or dolostone) were at a higher risk of contamination due

to the increased likelihood of fractures and solution channels to be present, versus their metamorphic (e.g., granite or gneiss) counterpart (Conboy & Goss, 2000). Relating stratigraphy with *E. coli* contamination events has been difficult in the past due to a lack of data, since constructing monitoring wells is often expensive and time consuming. Thus, there is opportunity to improve the understanding of this relationship.

Beyond geology, specific capacity - a measure of the well's productivity - can be used as a proxy to better understand the transport rate of groundwater around a well. Wells with higher specific capacity are typically in an aquifer with higher transmissivity, representing an area with increased transport rates, which could result in increased contamination potential due to lower opportunity of natural attenuation (Ehirim & Nwankwo, 2010). Conversely, a well with lower specific capacity may result in limitations to the well user due to decreased groundwater flow but may decrease risk of *E. coli* contamination. While specific capacity has been used in groundwater potential mapping (Kadam et al., 2020; Muleta & Abate, 2020), its use in contaminant transport related applications has been limited.

1.2.4 Well Characteristics

Well characteristics impact the physical integrity of the well and thus influence *E. coli* ingress (di Pelino et al., 2019). Some characteristics include well depth, casing diameter, and casing material, often driven in part by the drilling method required for a location or the standards or best practices when a well was established (Harris et al., n.d.). Three main well construction methods include dug wells, driven wells, and drilled wells, each with their advantages and disadvantages. Dug wells typically have large casing diameters and a

shallow depth, tapping into shallow aquifers. While these wells can produce from less permeable materials, they are more susceptible to contamination as they are more likely to be under the influence of surface water and more likely to run dry (USGS, 2018). Driven wells are smaller in diameter, but also very shallow, meaning that, like dug wells, they are also more susceptible to contamination. Finally, drilled wells are the most expensive to install but have a smaller diameter and can be very deep. These wells relieve some concern for contamination as they are unlikely to be under the direct influence of surface waters (Harris et al., n.d.; USGS, 2018). Due to this reduced concern, well users typically assume that deep wells are protected from contamination (Kreutzwiser et al., 2010).

1.2.5 Human Behaviour

Understanding human behaviour towards well stewardship is a crucial component in helping to reduce the presence of *E. coli* in wells. While this relationship is not necessarily a direct driver of contamination, it can inform and explain certain physical variables like well condition, well characteristics, well protection and intervention strategies, and can be important in linking well contamination to illness in consumers. Water quality testing is critical as it is the only way to characterize well water quality, which provides important information for both well stewardship practices and human health protection. Private wells have historically been undertested, often not even being tested once per year (Kreutzwiser et al., 2010; Maier et al., 2014). This limited testing may be attributable to complacency (e.g., history of non-adverse sample results or no concerning colour or odour), no experience of adverse health effects, or inconvenience (e.g., limited hours at sample drop off locations) (Invik, 2015). For example, Qayyum et

al. (2020) found that well users that received an initial negative index test (not containing *E. coli* or total coliforms) retested 64% of the time, compared to a 74% retesting rate when the initial index test was positive (containing *E. coli* or total coliforms). There are many factors impacting the testing practices of private well users and it is imperative that rationale is better understood to increase well user safety and increase data acquisition rates for further research to be done. Once testing trends are better understood, increased well user outreach will improve knowledge, attitudes, and practices with respect to well sampling and testing to move well user sampling practices closer to the “temporal truth” (Latchmore et al., 2020) of *E. coli* contamination events. Not only can improved well stewardship work to better protect the well from *E. coli* contamination, but it can also improve researchers' knowledge of well responses to both human and environmental variables through increased testing (i.e., increasing data points). For this reason, moving towards a coupled-systems approach for well contamination can have a significant impact on overall well health.

1.3 SITE DESCRIPTION – ONTARIO, CANADA

The geographic scope of this work is Ontario, Canada's second-largest province, covering more than one million square kilometers. It extends from approximately 42°N to 57°N latitude and from 75°W to 95°W longitude (Thompson et al., 2000). The greatest LULC categories by areal extent in Ontario are agricultural/pastoral, urban, and forests, followed by wetlands and scrublands (Figure 1-2A).

Ontario has a population of over 14.5 million people (and a 1.3% growth rate in 2021), with more than 85% living in urban centres, predominantly in cities surrounding the

Great Lakes (Figure 1-2B) (Douglas and Pearson, 2022). Key population hubs include cities in central Ontario such as Toronto and Hamilton; southwestern Ontario including London and Sarnia; eastern Ontario including Ottawa and Kingston; and northern Ontario including Thunder Bay and Sault Ste. Marie (Government of Ontario, 2023). The median age of Ontarians is approximately 40 years, and life expectancies are 79 years for men and 84 years for women (Government of Ontario, 2023).

Geologically, Ontario exhibits diverse landscapes, ranging from the Canadian Shield, comprising most of northern Ontario and housing large mineral deposits (generally unsuitable for agriculture), to the sedimentary limestone, shale, and sandstone underlying southern Ontario (Figure 1-2C) (Hillmer et al., 2023). Ontario experiences four distinct seasons representing significant climate variations throughout the year with notable regional variations. For example, the most northern town in Ontario, Fort Severn, can expect a mean winter temperature of -20°C and a mean summer temperature of 13°C versus the most southern city in Ontario, Kingsville, which can expect a mean winter temperature of -3°C and a mean summer temperature of 21°C . Similarly, snowfall amounts also vary regionally, where significant amounts of snowfall can be expected within the Snow Belt of Ontario (e.g., Owen Sound, Parry Sound, Sault Ste. Marie), receiving in excess of 250 cm annually, whereas more southern areas (e.g., Toronto, Hamilton) receive only 150 cm annually (Hillmer et al., 2023).

In addition to its diverse climate, Ontario is a province known for its diversity in other areas as well including, but not limited to, a range of socio-economic status, land use-land covers, and population densities. This diversity gives rise to a complex coupled-

system that affects the over 4 million Ontarians relying on private well water differently based on their geographic location and the associated characteristics.

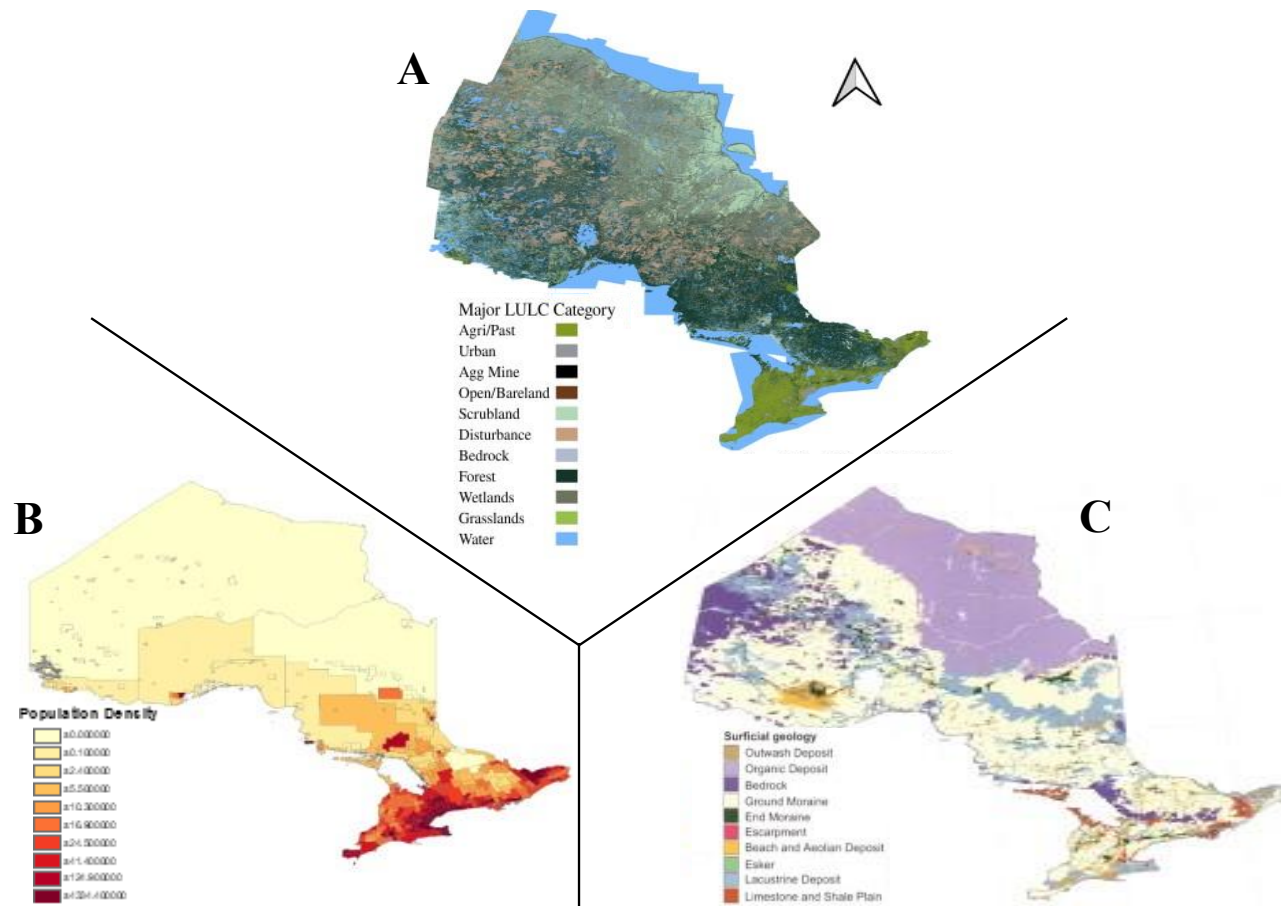


Figure 1-2: Ontario maps depicting: A – Major LULC categories in Ontario (White et al. 2022), B – Population Density (Statistics Canada, 2016), and C – Surficial Geology (Thompson et al. 2000).

1.4 RESEARCH OBJECTIVES

The understanding of the fate and transport mechanisms of pathogens driving groundwater contamination is often lacking due to data limitation caused by high data acquisition costs. The goal of this research is to improve the understanding of the drivers of private well contamination, specifically exploring *E. coli* in an Ontario, Canada context. To achieve this, analyses move towards a coupled-systems approach utilizing innovative data-driven machine learning techniques. Key objectives in this work include:

- Better understand the drivers of, and independent relationships between, the physical environment (i.e., climate, well and aquifer characteristics, location), *E. coli* presence, and human behaviour (i.e., sampling behaviour) in wells (Chapter 2);
- Expand and improve upon existing knowledge of drivers of *E. coli* presence and concentration in wells, and identify surrogate measures for these drivers that improve access to information (Chapters 3 and 4); and
- Move towards a holistic, seasonally-based coupled-systems model by including both physical (i.e., weather and climate, well and aquifer characteristics, LULC) and human behaviour (i.e., testing behaviour, impacts of driving time to, and hours of operation of, testing drop-off locations) variables in an explanatory model exploring *E. coli* presence in private wells (Chapter 5).

This work explores the human and physical drivers of *E. coli* contamination in private wells, utilizing innovative machine learning techniques, moving towards the development of a coupled-systems model. Additionally, a method was developed to integrate regression and literature-based findings to assign a confidence ranking to the variable “*E. coli*

contamination potential classes for private wells based on LULC”, resulting in an informative map (raster file) accessible to the public. Weather variables (i.e., snow, rain, snow and rain) in explanatory regression models were assigned to seasonal models utilizing a novel approach to seasonal delineations, moving past standard seasonal delineations (e.g., summer runs from June – August) towards a latitudinal, climate data-driven approach allowing for varying seasonal onset and cessation based on snow melt patterns. The innovative methods outlined in this work are geographically transferable and can be used to uncover more invaluable trends across vastly different climates, well practices, and social demographics. The trends and key variables identified in this work will inform policy and groundwater management strategies and empower well users to improve their own well stewardship.

1.4.1 Thesis Outline

This thesis includes six chapters. Chapter 1 introduces the current reliance on private drinking water wells and state-of-knowledge of key well contamination drivers. It provides a general context for the research and outlines the research objectives.

Chapter 2 presents a study to introduce an Ontario-based private well dataset and better understand drivers of, and relationships between, climate (seasonality), well and aquifer characteristics (geology, well depth), sampling behaviour (frequency, timing), and *E. coli* (presence, concentration). This chapter also introduces a novel application of supervised machine learning techniques, including Generalized Additive Models of Location, Scale, and Shape (GAMLSS) and Association Rule Analysis, to improve the state of data-driven models. This work is published in the journal of Water Research.

Chapter 3 contributes a new raster dataset depicting a spatial distribution and relationship of *E. coli* contamination potential in groundwater wells based on Land Use-Land Cover (LULC) classes in Ontario. This work utilizes LULC as a surrogate measure of *E. coli* loading and transport in the subsurface. Contamination potential classes were determined by evaluating the agreement between existing literature and data-driven models. Datasets are published in the Federated Research Data Repository, the associated manuscript has been submitted to Applied Spatial Analyses and Policy.

Chapter 4 presents a study exploring the possibility of moving beyond fixed rainfall lag times toward rainfall intermittency patterning to predict *E. coli* presence in private Ontario wells. This work uses an array of supervised and unsupervised machine learning techniques to improve the understanding of rainfall as a potential driver of well contamination. This work will be submitted to Journal of Weather and Climate Extremes by August 2023.

Chapter 5 is an extension of Chapters 2 through 4 combining relevant previously explored contamination drivers and new potential drivers, moving towards a coupled-systems approach. Explanatory models were split into geographically unique seasons (i.e., winter, summer, shoulder) to explore both physical (land characteristics, well density, flow accumulation and well and aquifer characteristics, weather) and human variables (testing hotspots, drive time to closest sample drop-off location) in one holistic model. The purpose of this culmination is to account for all major processes and represent both physical and human systems in order to supply key take-aways for private well users, policy and

decision makers to improve well stewardship. This work will be submitted to Water Research by August 2023.

Chapter 6 is a summary of the main conclusions as well as recommendations for future research.

1.5 WORKS CITED

Bhandari, J., & Sedhai, Y. R. (2022). Hemolytic Uremic Syndrome. *StatPearls*.
<https://www.ncbi.nlm.nih.gov/books/NBK556038/>

Carlton, E. J., Eisenberg, J. N. S., Goldstick, J., Cevallos, W., Trostle, J., & Levy, K. (2014). Heavy rainfall events and diarrhea incidence: The role of social and environmental factors. *American Journal of Epidemiology*, 179(3), 344–352.
<https://doi.org/10.1093/aje/kwt279>

Charrois, J. W. A. (2010). Private drinking water supplies: challenges for public health. *CMAJ: Canadian Medical Association Journal*, 182(10), 1061.
<https://doi.org/10.1503/CMAJ.090956>

Charron, D. F., Thomas, M. K., Waltner-Toews, D., Aramini, J. J., Edge, T., Kent, R. A., Maarouf, A. R., & Wilson, J. (2004). Vulnerability of waterborne diseases to climate change in Canada: A review. *Journal of Toxicology and Environmental Health - Part A*, 67(20–22), 1667–1677. <https://doi.org/10.1080/15287390490492313>

- Conboy, M. J., & Goss, M. J. (2000). Natural protection of groundwater against bacteria of fecal origin. *Journal of Contaminant Hydrology*, 43(1), 1–24. [https://doi.org/10.1016/S0169-7722\(99\)00100-X](https://doi.org/10.1016/S0169-7722(99)00100-X)
- de Roos, A. J., Kondo, M. C., Robinson, L. F., Rai, A., Ryan, M., Haas, C. N., Lojo, J., & Fagliano, J. A. (2020). Heavy precipitation, drinking water source, and acute gastrointestinal illness in Philadelphia, 2015-2017. *PLOS ONE*, 15(2), e0229258. <https://doi.org/10.1371/journal.pone.0229258>
- di Pelino, S., Schuster-Wallace, C., Hynds, P. D., Dickson-Anderson, S. E., & Majury, A. (2019). A coupled-systems framework for reducing health risks associated with private drinking water wells. *Canadian Water Resources Journal*, 44(3), 280–290. <https://doi.org/10.1080/07011784.2019.1581663>
- Dordio, A., Carvalho, A. J. P., & Pinto, A. P. (2008). Wetlands: Water “Living Filters”? In *Wetlands: Ecology, Conservation and Restoration* (pp. 15–71).
- Douglas, A.G. and Pearson, D. (2022). Ontario; Chapter 4 in *Canada in a Changing Climate: Regional Perspectives Report*, (ed.) F.J. Warren, N. Lulham, D.L. Dupuis and D.S. Lemmen; Government of Canada, Ottawa, Ontario.
- Dusek, N., Hewitt, A. J., Schmidt, K. N., & Bergholz, P. W. (2018). Landscape-scale factors affecting the prevalence of *Escherichia coli* in surface soil include land cover type, edge interactions, and soil pH. *Applied and Environmental Microbiology*, 84(10). <https://doi.org/10.1128/AEM.02714-17>

- Edberg, S. C., Rice, E. W., Karlin, R. J., & Allen, M. J. (2000). *Escherichia coli*: the best biological drinking water indicator for public health protection. Symposium Series (Society for Applied Microbiology), 88(29). <https://doi.org/10.1111/J.1365-2672.2000.TB05338.X>
- Ehirim, C. N., & Nwankwo, C. N. (2010). Evaluation of aquifer characteristics and groundwater quality using geoelectric method in Choba, Port Harcourt. *Archives of Applied Science Research*, 2(2), 396–403. www.scholarsresearchlibrary.com
- Gleason, J. A., & Fagliano, J. A. (2017). Effect of drinking water source on associations between gastrointestinal illness and heavy rainfall in New Jersey. *PLoS ONE*, 12(3), e0173794. <https://doi.org/10.1371/journal.pone.0173794>
- Government of Ontario. (2018). O. Reg. 319/08: SMALL DRINKING WATER SYSTEMS.
- Government of Ontario (2023). About Ontario. Retrieved from <https://www.ontario.ca/page/about-ontario>
- Gregory, L. F., Harmel, R. D., Karthikeyan, R., Wagner, K. L., Gentry, T. J., & Aitkenhead-Peterson, J. A. (2019). Elucidating the Effects of Land Cover and Usage on Background *Escherichia coli* Sources in Edge-of-Field Runoff. <https://doi.org/10.2134/jeq2019.02.0051>
- Harris, B. L., Hoffman, D. W., & Mazac, F. J. (n.d.). Reducing the Risk of Ground Water Contamination by Improving Wellhead Management and Conditions. Retrieved

August 16, 2022, from
http://publications.tamu.edu/WATER/PUB_water-Reducing%20the%20Risk%20of%20Groundwater%20Contamination.pdf

Hayashi, M., van der Kamp, G., & Schmidt, R. (2003). Focused infiltration of snowmelt water in partially frozen soil under small depressions. *Journal of Hydrology*, 270(3–4), 214–229. [https://doi.org/10.1016/S0022-1694\(02\)00287-1](https://doi.org/10.1016/S0022-1694(02)00287-1)

Hillmer, N., & Bothwell, R. (2023). Geography of Ontario. In *The Canadian Encyclopedia*. Retrieved from <https://www.thecanadianencyclopedia.ca/en/article/geography-of-ontario>

Hua, A. K. (2017). Land Use Land Cover Changes in Detection of Water Quality: A Study Based on Remote Sensing and Multivariate Statistics. *Journal of Environmental and Public Health*, 2017. <https://doi.org/10.1155/2017/7515130>

Hynds, P. D., Thomas, M. K., & Pintar, K. D. M. (2014). Contamination of groundwater systems in the US and Canada by enteric pathogens, 1990-2013: A review and pooled-analysis. PLoS ONE, 9(5). <https://doi.org/10.1371/journal.pone.0093301>

Invik, J. (2015). Total Coliform and Escherichia Coli Contamination in Rural Well Water in Alberta, Canada: Spatiotemporal Analysis and Risk Factor Assessment [University of Calgray]. In ProQuest Dissertations and Theses. <https://doi.org/10.11575/PRISM/28466>

- Jabbar, F. K., Grote, K., & Tucker, R. E. (2019). A novel approach for assessing watershed susceptibility using weighted overlay and analytical hierarchy process (AHP) methodology: a case study in Eagle Creek Watershed, USA. *Environmental Science and Pollution Research*, 26(31), 31981–31997. <https://doi.org/10.1007/s11356-019-06355-9>
- Jagai, J. S., Li, Q., Wang, S., Messier, K. P., Wade, T. J., & Hilborn, E. D. (2015). Extreme precipitation and emergency room visits for gastrointestinal illness in areas with and without combined sewer systems: An analysis of Massachusetts data, 2003–2007. *Environmental Health Perspectives*, 123(9), 873–879. <https://doi.org/10.1289/ehp.1408971>
- Jonsson, A., & Agerberg, S. (2015). Modelling of *E. coli* transport in an oligotrophic river in northern Scandinavia. *Ecological Modelling*, 306, 145–151. <https://doi.org/10.1016/j.ecolmodel.2014.10.021>
- Kadam, A., Wagh, V., Umrikar, B., & Sankhua, R. (2020). An implication of boron and fluoride contamination and its exposure risk in groundwater resources in semi-arid region, Western India. *Environment, Development and Sustainability*, 22(7), 7033–7056. <https://doi.org/10.1007/S10668-019-00527-W/FIGURES/7>
- Khaleel, R., Reddy, K. R., Overcash, M. R., & Westerman, P. W. (1978). TRANSPORT OF POTENTIAL POLLUTANTS IN RUNOFF WATER FROM LAND AREAS RECEIVING ANIMAL WASTES: A REVIEW. Paper - American Society of Agricultural Engineers, 14, 421–436.

- Kraay, A. N. M., Man, O., Levy, M. C., Levy, K., Ionides, E., & Eisenberg, J. N. S. (2020). Understanding the impact of rainfall on diarrhea: Testing the concentration-dilution hypothesis using a systematic review and meta-analysis. *Environmental Health Perspectives*, 128(12), 126001-1-126001–126016. <https://doi.org/10.1289/EHP6181>
- Kreutzweiser, R., de Loë, R. C., & Imgrund, K. (2010). Out of Sight, Out of Mind: Private Water Well Stewardship in Ontario. Report on the Findings of the Ontario Household Water Well Owner Survey 2008. In Water Policy and Governance Group, University of Waterloo, Waterloo, ON. www.wpgg.ca
- Latchmore, T., Hynds, P., Brown, S., Schuster-Wallace, C., Dickson-Anderson, Sarah McDermott, K., & Majury, A. (2020). Analysis of a Large Spatiotemporal Groundwater Quality Dataset, Ontario 2010 - 2017: Informing Human Health Risk Assessment and Testing Guidance for Private Drinking Water Wells.
- Levy, K., Woster, A. P., Goldstein, R. S., & Carlton, E. J. (2016). Untangling the Impacts of Climate Change on Waterborne Diseases: A Systematic Review of Relationships between Diarrheal Diseases and Temperature, Rainfall, Flooding, and Drought. In *Environmental Science and Technology* (Vol. 50, Issue 10, pp. 4905–4922). <https://doi.org/10.1021/acs.est.5b06186>
- Lisovski, S., Ramenofsky, M., & Wingfield, J. C. (2017). Defining the degree of seasonality and its significance for future research. *Integrative and Comparative Biology*, 57(5), 934–942. <https://doi.org/10.1093/icb/icx040>

- Maier, A., Krolik, J., Randhawa, K., & Majury, A. (2014). Bacteriological testing of private well water: A trends and guidelines assessment using five years of submissions data from southeastern Ontario. *Canadian Journal of Public Health*, 105(3), 203–208. <https://doi.org/10.17269/cjph.105.4282>
- Medema, G. J., Shaw, S., Waite, M., Snozzi, M., Morreau, A., & Grabow, W. (2004). Assessing Microbial Safety of Drinking Water Improving Approaches and Methods WHO Drinking Water Quality Series. *Science of The Total Environment*, 111–158. [https://doi.org/10.1016/S0048-9697\(04\)00275-X](https://doi.org/10.1016/S0048-9697(04)00275-X)
- Muleta, D., & Abate, B. (2020). Groundwater for Sustainable Development 12 (2021) 100485 Groundwater hydrodynamics and sustainability of Addis Ababa city aquifer. <https://doi.org/10.1016/j.gsd.2020.100485>
- Murphy, H. M., Prioleau, M. D., Borchardt, M. A., & Hynds, P. D. (2017). Review: Epidemiological evidence of groundwater contribution to global enteric disease, 1948–2015. *Hydrogeology Journal*, 25(4), 981–1001. <https://doi.org/10.1007/s10040-017-1543-y>
- Murphy, H. M., Thomas, M. K., Schmidt, P. J., Medeiros, D. T., McFadyen, S., & Pintar, K. D. M. (2016). Estimating the burden of acute gastrointestinal illness due to *Giardia*, *Cryptosporidium*, *Campylobacter*, *E. coli* O157 and norovirus associated with private wells and small water systems in Canada. *Epidemiology and Infection*, 144, 1355–1370. <https://doi.org/10.1017/S0950268815002071>

- Namugize, J. N., Jewitt, G., & Graham, M. (2018). Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. *Physics and Chemistry of the Earth*, 105, 247–264. <https://doi.org/10.1016/j.pce.2018.03.013>
- OMAFRA. (2020). Climate Zones and Planting Dates for Vegetables in Ontario. *Vegetable Crops*. <http://www.omafr.gov.on.ca/english/crops/facts/climzoneveg.htm>
- Paule-Mercado, M. A., Ventura, J. S., Memon, S. A., Jahng, D., Kang, J. H., & Lee, C. H. (2016). Monitoring and predicting the fecal indicator bacteria concentrations from agricultural, mixed land use and urban stormwater runoff. *Science of the Total Environment*, 550, 1171–1181. <https://doi.org/10.1016/j.scitotenv.2016.01.026>
- Pejman, A. H., Nabi Bidhendi, ; G R, Karbassi, ; A R, Mehrdadi, ; N, & Bidhendi, ; M Esmaeili. (2009). Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques. *Int. J. Environ. Sci. Tech*, 6(3), 467–476.
- Petersen, F., & Hubbart, J. A. (2020). Physical Factors Impacting the Survival and Occurrence of *Escherichia coli* in Secondary Habitats. *Water*, 12(6), 1796. <https://doi.org/10.3390/w12061796>
- Rivera, A. (2017). The state of ground water in Canada. *Ground Water Canada*. <https://www.groundwatercanada.com/the-state-of-ground-water-in-canada-3584/>
- Schuster, C. J., Ellis, A. G., Robertson, W. J., Dominique, F., Aramini, J. J., Marshall, B. J., & Medeiros, D. T. (2005). Infectious Disease Outbreaks Related to Drinking Water in Canada, 1974-2001. 96(4), 254–258.

- Soller, J. A., Schoen, M. E., Bartrand, T., Ravenscroft, J. E., & Ashbolt, N. J. (2010). Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water Research*, 44(16), 4674–4691. <https://doi.org/10.1016/j.watres.2010.06.049>
- Statistics Canada (2016). Boundary Files, 2016 Census. Statistics Canada Catalogue no. 92-160G.
- Thompson, I. D., Perera, A., Euler, D., & Ontario. Ministry of Natural Resources. (2000). Ecology of a managed terrestrial landscape: patterns and processes of forest landscapes in Ontario. Published by UBC Press in cooperation with the Ontario Ministry of Natural Resources.
- Trenberth, K. E., Zhang, Y., & Gehne, M. (2017). Intermittency in precipitation: Duration, frequency, intensity, and amounts using hourly data. *Journal of Hydrometeorology*, 18(5), 1393–1412. <https://doi.org/10.1175/JHM-D-16-0263.1>
- USGS. (2018, June). Groundwater Wells | U.S. Geological Survey. <https://www.usgs.gov/special-topics/water-science-school/science/groundwater-wells>
- White, K., Dickson-Anderson, S., Schuster-Wallace, C., 2022. Land Use Land Cover in Ontario and Respective Impact to *E. coli* Contamination in Private Wells. <https://doi.org/10.20383/102.0530>

- Whitman, R. L., Przybyla-Kelly, K., Shively, D. A., Nevers, M. B., & Byappanahalli, M. N. (2008). Sunlight, season, snowmelt, storm, and source affect *E. coli* populations in an artificially ponded stream. *Science of the Total Environment*, 390(2–3), 448–455. <https://doi.org/10.1016/j.scitotenv.2007.10.014>
- Wiebe, A. J., Rudolph, D. L., Pasha, E., Brook, J. M., Christie, M., & Menkveld, P. G. (2021). Impacts of event-based recharge on the vulnerability of public supply wells. *Sustainability (Switzerland)*, 13(14), 7695. <https://doi.org/10.3390/SU13147695/S1>

CHAPTER 2 EXPLORATION OF *E. COLI*

CONTAMINATION DRIVERS IN PRIVATE DRINKING

WATER WELLS: AN APPLICATION OF MACHINE

LEARNING TO A LARGE, MULTIVARIABLE, GEO-

SPATIO-TEMPORAL DATASET

Summary of Paper 1: Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset, by K.White, S. Dickson-Anderson, A. Majury, K. McDermott, P. Hynds, S.R. Brown, and C. Schuster-Wallace, Journal of Water Research, 197, doi: 10.1016/j.watres.2021.117089, 2021.

Summary:

This research sets the stage for the rest of the thesis. This work uses supervised machine learning approaches, Generalized Additive Models for Location, Shape, and Scale (GAMLSS) and Association Rule Analysis (ARA), on a large Ontario-based private well dataset to begin the process of finding connections between physical factors, well testing practices, well characteristics, and *E. coli* contamination private wells. The results of this demonstrated:

- Consensus with existing literature confirms the validity of the novel approach of utilizing supervised machine learning algorithms, like GAMLSS and ARA, to explain contamination events in private wells, while still highlighting the need to couple findings with disciplinary expertise.
- Some key independent variables for the explanation of *E. coli* contamination events included wells completed in consolidated material experiencing greater likelihood of contamination, while increased well depth decreases likelihood of contamination it does not eliminate likelihood which represents a requirement for well users to remain vigilant, human-based testing patterns are not aligning with

periods of increased contamination occurrences.

- The need for further exploration of seasonal delineations, meteorological-based variables, and impacts of human behaviour trends on well contamination events.

2.1 ABSTRACT

Groundwater resources are under increasing threats from contamination and overuse, posing direct threats to human and environmental health. The purpose of this study is to better understand drivers of, and relationships between, well and aquifer characteristics, sampling frequencies, and microbiological contamination indicators (specifically *E. coli*) as a precursor for improving knowledge and tools to assess aquifer vulnerability and well contamination within Ontario, Canada.

A dataset with 795,023 microbiological testing observations over an eight-year period (2010 to 2017) from 253,136 unique wells across Ontario was employed. Variables in this dataset include date and location of test, test results (*E. coli* concentration), well characteristics (well depth, location), and hydrogeological characteristics (bottom of well stratigraphy, specific capacity). Association rule analysis, univariate and bivariate analyses, regression analyses, and variable discretization techniques were utilized to identify relationships between *E. coli* concentration and the other variables in the dataset.

These relationships can be used to identify drivers of contamination, their relative importance, and therefore potential public health risks associated with the use of private wells in Ontario. Key findings are that: *i*) bedrock wells completed in sedimentary or igneous rock are more susceptible to contamination events; *ii*) while shallow wells pose a greater risk to consumers, deep wells are also subject to contamination events and pose a potentially unanticipated risk to health of well users; and, *iii*) well testing practices are influenced by results of previous tests. Further, while there is a general correlation between months with the greatest testing frequencies and concentrations of *E. coli* occurring in

samples, an offset in this timing is observed in recent years. Testing remains highest in July while peaks in adverse results occur up to three months later. The realization of these trends prompts a need to further explore the bases for such occurrences.

Keywords: Private Drinking Water, Groundwater, *E. coli*, Testing Trends, Large Dataset, Machine Learning

2.2 INTRODUCTION

Globally, groundwater resources are in high demand for agricultural, domestic, and industrial purposes. Over 50% of the world's population uses groundwater as a source of drinking water, while 35% rely solely on groundwater for all domestic use. Groundwater resources have become a casualty of these competing demands, resulting in an estimated 20% of aquifers being over-exploited (UN Water, 2015). Over-exploitation creates additional challenges beyond the loss of water supplies, including saltwater intrusion, loss of wetlands and springs, and land subsidence. Poor aquifer, waste, and wastewater management pose additional threats to groundwater through contamination by chemicals, radionuclides, and microorganisms. Once contaminated, remediation is particularly challenging due to large water volumes, long residence times, and physical inaccessibility of aquifers (Foster and Chilton, 2003). Where groundwater is still available for use, this ongoing over-exploitation and contamination introduces a cause for urgency in managing groundwater supplies more effectively, particularly for human health.

An estimated 22% of Canadians rely on groundwater for their domestic water supply (Murphy et al., 2017; Rivera, 2017), with 12% (~ 4.5 million) relying on privately

owned and maintained groundwater supplies (Murphy et al., 2016) that are outside governmental regulation and oversight. In the Great Lakes system, groundwater is considered the sixth great lake (Fong et al., 2007); however, ongoing microbiological groundwater contamination within the Great Lakes system (Fong et al., 2007) threatens public health. The combination of heavy reliance on this ‘sixth great lake’ as a drinking water source, ever-increasing contamination, and lack of government-imposed regulation for private systems present significant public health challenges. Historically, approximately 189 of 288 reported Canadian waterborne disease outbreaks occurred in privately owned wells or small drinking water systems (Schuster et al., 2005), this leaves approximately 2.9 million Canadians at risk due to reliance on these systems (Murphy et al., 2016). Challenges facing private and small systems include limited resources for maintenance, management, and protection (Rivera, 2017), and lack of regulation.

Microbiological groundwater contamination events occur periodically across space and time resulting in sporadic patterns of acute gastrointestinal illness (AGI) caused by consumption of contaminated water. These cases of AGI are difficult to track even in high income countries, not only due to their sporadic nature, but also significant under-reporting as individuals rarely seek medical attention (Murphy et al., 2016), and difficulties in confirming the exposure pathway (Schuster et al., 2005). As such, the number of actual groundwater-related cases of AGI is generally assumed to be significantly higher than reported (Murphy et al., 2016). To effectively mitigate these events and reduce risk, it is crucial to determine how and when pathogens are entering and travelling through the groundwater system. The four main factors impacting the fate and transport of

microbiological contaminants in aquifers are weather patterns, hydrogeologic conditions, presence of a source of microbiological contamination, and well conditions (location, construction, and maintenance) (O'Dwyer et al., 2018). *Escherichia coli* (*E. coli*) is used as a standard indicator for faecal contamination. Any contamination risk can be mitigated or exacerbated through human behaviours and practices, including well maintenance, water quality testing, water treatment, and water consumption patterns (Figure 2-1) (Di Pelino et al., 2019).

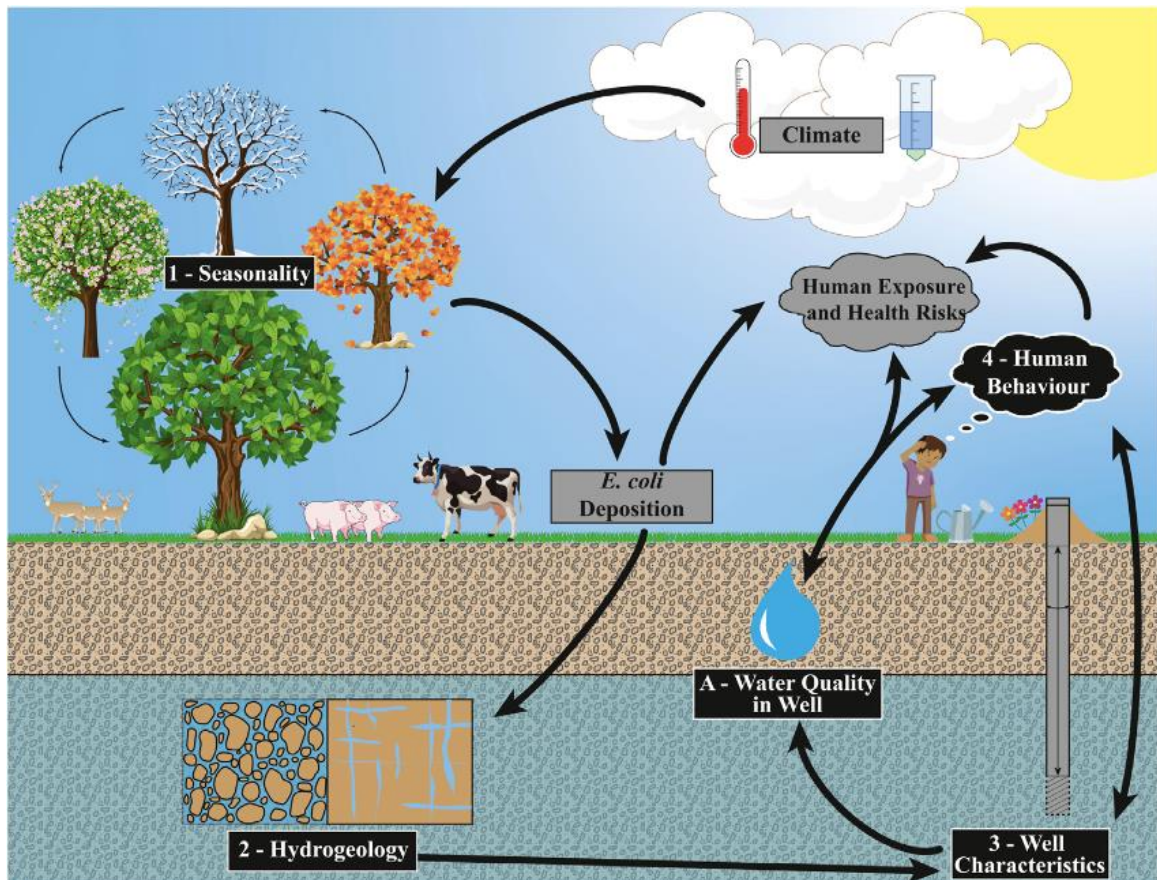


Figure 2-1: Fate and transport mechanisms driving *E. coli* concentrations in private wells (i.e., contamination risk) considering a coupled-systems approach, adapted from (Di Pelino et al., 2019), where numbered text represent drivers used to develop explanatory models.

The health risks associated with dependence on drinking water wells, combined with the increasing potential for groundwater to become contaminated, present a risk that most private well users are unaware of, and unable to access information on, beyond individual well sample results (Di Pelino et al., 2019; Kreutzwiser et al., 2010). As such, a need exists to improve our understanding of groundwater susceptibility and human health risk models.

This study uses a data-driven approach to modelling groundwater fate and transport. These approaches have contributed to the understanding of contaminant transport in groundwater (Buckerfield et al., 2020; Knoll et al., 2019) and reduce computational requirements when compared to process-based models (Castalietti et al., 2012). Data-driven approaches have been used successfully in modelling *E. coli* behaviour in fractured rock environments (Yosri et al., 2021), predicting groundwater nitrate concentrations (Knoll et al., 2019), identifying solute transport pathways in fractured aquifers (Yosri et al., 2021), and characterizing uncertainty in coastal plain watershed systems (Samadi et al., 2018).

The goal of this study is to better understand drivers of, and relationships between, climate (seasonality), well and aquifer characteristics (geology, well depth), sampling behaviour (frequency, timing), and *E. coli* (presence, concentration). This is undertaken through a novel application of supervised machine learning techniques, namely GAMLSS, to a large dataset capturing both hydrological and microbiological variables for private wells. These variables are *collectively* assessed within explanatory models as a precursor for improving understanding of aquifer vulnerability to contamination and assessing well

water quality. This work builds on Latchmore et al. (2020), which *individually* assessed geology and testing frequency to inform testing recommendations for private well users within a health risk framework.

2.3 METHODS

2.3.1 Dataset

The analyses in this paper have been undertaken using an Ontario-specific groundwater dataset that consists of 795,023 well sample observations for 253,136 unique private wells that have been tested 1 to 446 times between 2010 and 2017, inclusive. The dataset was created through the amalgamation of Ontario’s Well Water Information System (WWIS) and Well Water Testing Database (WWTD). More information on these databases can be found in Latchmore et al. (2020).

Parameters in the dataset are described in Table 2-1, along with additional relevant dataset information and generated sub-classifications for selected variables, established according to criteria in the literature for the purpose of these analyses. The specific classification methods are presented in A.2.1. These sub-classifications are used in lieu of, or alongside, discrete values in some analyses to fit regulatory definitions or account for uncertainty.

2.3.2 Data Processing

While the original dataset was assessed for quality as described by Latchmore et al. (2020), additional cleaning, data conversion, and sub-classification (Table 2-1) were required to enable the assessment of factors driving the presence of *E. coli* in private wells

in Ontario, as described in A.2.1. Only observations associated with wells that were in use and classified as domestic or multiple use including domestic in the dataset were included in these analyses. To better understand potential relationships, selected continuous variables were classified into data bins to account for uncertainty in the data (e.g., specific capacity, *E. coli* concentration) or to align variables with well regulations, standards, and recommendations (e.g., well depth, testing frequency). In the instance of well depth, well regulations and data distribution were considered to ensure categorical bins were evenly distributed. Latitude and longitude are utilized as gradients over space rather than point locations. As such, they have been disaggregated into half-degree bins.

Table 2-1: Description of variables contained within the merged WWIS and WWTD dataset, including sub-classifications derived for the purpose of these analyses.

| Parameter | Description | Sub-classifications derived for current analyses |
|-----------------------------|--|---|
| Well Use | Intended use of well water (Domestic, Agriculture, Livestock, Commercial, Public) | Domestic and Multiple Use including Domestic |
| <i>E. coli</i> Result | Number of <i>E. coli</i> reported in sample by laboratory. Laboratory Reporting Range: 0 – 80 CFU/100 mL | non-detects (ND): 0 Category 1: 1-10 Category 2: 11-50 Category 3: 51+ |
| Total Coliforms (TC) Result | Number of TC reported in sample by laboratory. Laboratory Reporting Range: 0 – 80 CFU/100 mL | No Significant Evidence: ≤ 5 May Be Unsafe to Drink: > 5 |
| Location | Location of well geographically, in longitude and latitude | Binned into 0.5-degree ranges |
| Date of Observation | Date of water sample collection | |
| Geological Formation | Stratigraphy of geologic formation in which well is situated (originally recorded in ft) | Consolidated (further categorised as igneous, metamorphic, sedimentary) (See Table A.2-1) Unconsolidated (further categorised as high, medium, low permeability) (See Table A.2-2) |
| Pump Test | Information recorded from pump test includes static water level, water level after pumping, pump test rate, and pump | Specific Capacity (GPM/m) = Pumping Rate/Drawdown Low (0 - <3.3 GPM/m) |

| | | |
|---------------------------|---|--|
| | test duration (originally recorded in GPM/ft) | Moderate (3.3 – 16.4 GPM/m) High (>16.4 GPM/m) |
| Well Depth | Distance from ground surface to bottom of well (originally recorded in ft) and classification of well depth | Shallow/Moderate (< 12.5 m) Moderate 1 (12.5 m ≤ x < 18.3 m) Moderate 2 (18.3 m ≤ x < 24.4 m) Moderate 3 (24.4 m ≤ x < 31.1 m) Moderate 4 (31.1 m ≤ x < 41.8 m) Moderate 5 (41.8 m ≤ x < 61 m) Deep (≥ 61 m) |
| Date of Well Construction | Year well construction was completed | |
| Status | Qualitative microbiology comments based on laboratory processing of the sample (See Table A.2-3) | |

2.3.3 Statistical Analysis

Probability of *E. coli* contamination in Ontario private wells was investigated with respect to seasonality, geological formation, and well depth, using numerous data exploration and visualization techniques. The specific capacity was calculated based on pump test data in the dataset (See A2.1). To assess changes over time, trends were explored based on intra- and inter-annual patterns at different temporal resolutions. These resolutions include monthly, annual, and the entire study period. Note that 0.06% of wells account for approximately 20% of *E. coli* test results because of the large number of samples taken from these wells during the study period. Given that each *E. coli* sample represents a data point in space and time, the fact that they originate from a small number of wells helps to differentiate the impact of variable factors (e.g., seasonality) from fixed factors (e.g., geology, well characteristics). Further, the distribution of fixed variables (i.e., well depth, bottom of well stratigraphy, and specific capacity) were compared between a dataset containing all *E. coli* samples and one containing observations from individual

wells represented by the highest *E. coli* sample result. The distributions remained similar, indicating that highly sampled wells did not over-weight the models.

Before exploring more complex relationships using machine learning methods, univariate and bivariate analyses were conducted on all independent variables. Univariate analyses were conducted to explore the data distribution of each individual variable. Bivariate analyses were conducted to identify empirical relationships between individual variable pairs. Specifically, the probability of contamination given well depth and the probability of contamination given bottom stratigraphy were calculated (See A.2.1). These were followed by machine learning techniques, i.e., association rule and regression analyses. Regression analyses were chosen over other (non-regression) supervised machine learning techniques that require greater computational intensities (i.e., random forests) or that cannot be interpreted sufficiently to ensure adherence to physical processes (i.e., artificial neural networks). The generalized additive model for location, scale, and shape (GAMLSS) regression model was chosen due to the highly skewed distributions (zero-inflated) of some variables. GAMLSS is able to deal with zero-inflated variables through use of general distribution families (i.e., highly skewed with the addition of zero-inflated and zero-adjusted families) (Stasinopoulos and Rigby, 2007). The large number of observations with a zero *E. coli* count (87%) prohibits the use of linear models (LM), generalized linear models (GLM), or general additive models (GAM) (Stasinopoulos and Rigby, 2007). Association rule analysis was chosen as a supplementary technique to further explore select variables due to its ability to discover interesting relationships and strong

rules between variables in large datasets, while being considered a “fast mining algorithm” (Hahsler et al., 2005).

2.3.3.1 Regression Analyses

A series of regression analyses (R package “*gamlss*”; Rigby and Stasinopoulos, 2005) were conducted to develop explanatory models for *E. coli* concentration based on seasonality, hydrogeology, well characteristics, and human behaviour (Table A.2-4). A collinearity matrix was developed (utilizing Phi and Pearson’s coefficient) and correlated variables, as well as obvious confounders, were removed from the set of model input variables. The corresponding models use a distributional regression approach where all parameters of the conditional distribution of the response variable are modelled using explanatory variables (Rigby et al., 2019). Independent variables (Table A.2-4) were selected to develop a series of models to explain *E. coli* concentrations, each exploring different elements of the risk pathway (Figure 2-1): seasonal (Driver 1 in Figure 2-1), hydrogeological (Driver 2 in Figure 2-1), well characteristics (Driver 3 in Figure 2-1), and testing practices (Driver 4 in Figure 2-1). This method of separating models combines the power of machine learning with subject matter expertise, to understand the interactions and impacts of variables representing a specific driver of *E. coli* contamination along the risk pathway. Once developed, explanatory models for Drivers 1-3 informed development of an “informed model” based on all relevant variables in the dataset.

Based on subject matter expertise, various combinations of independent variables were included in models to assess their ability to explain the dependent variable (*E. coli* concentrations or testing frequencies). In some cases, continuous, categorical, and binary

forms of the same independent variable were assessed for performance against evaluation criteria (e.g., model option 1 uses binary bottom stratigraphy, and model option 2 uses categorical bottom stratigraphy). All models were evaluated against each other employing 10-fold cross validation, using the appropriate mixed model “fitting families”, as defined by Rigby et al. (2019). Fitting families were chosen to incorporate discrete, categorical, and continuous variables. Families chosen are as follows: zero adjusted logarithmic distribution (ZALG) and zero adjusted inverse Gaussian distribution (ZAIG) (Rigby et al., 2019).

The “best model” was identified as the one with the lowest cross validated Global Deviance (Rigby et al., 2019; Stasinopoulos and Rigby, 2005). It is important to note that this enables a comparison between models but does not reflect model accuracy. To consider model accuracy, residual analyses were conducted on the “best” models.

Once the best model was determined, models were trained (i.e., learning to fit the parameters of the independent variables) using a randomly selected dataset containing 80% of the data, and subsequently tested (i.e., assessment of trained model performance) on the remaining 20% of the data (Joshi, 2020), as a means to fit the model. This was conducted over 10 iterations with 10 unique data splits within each model, with the regression coefficients averaged across iterations to address parameter uncertainty (determine mean and variance) in the coefficients for the final explanatory model. Two-tailed hypothesis tests were used to assess the statistical significance of model variables. Note that statistical significance of variables in these models do not render the model predictive. Rather, significance refers to the importance of the variable in explaining *E. coli* presence or

concentration in a well while the magnitude of the coefficient indicates relative impact. Ultimately, the goal of these models is to explain casual relationships, not predict the probability of an event occurring (Sainani, 2014).

Finally, to assess variable importance, each independent variable was removed one by one, and cross validated Global Deviance values were calculated and compared to assess the impact. The “most important” variable to the model is defined as the variable that results in the greatest increase in cross validated Global Deviance when removed from the model.

2.3.3.2 *Analyses of Hydrogeological Settings and Well Characteristics*

Assessment of the impacts of the bottom layer stratigraphy (categorized by rock type and grain size) on *E. coli* concentration (CFU/100mL) was undertaken utilizing Association Rule Mining Analysis using the Apriori algorithm (R package “*arules*”; Hahsler et al., 2005), which identifies statistically interesting relationships in large datasets. The “interestingness” of a rule is based on four key measurements: *confidence*, which is the estimate of the conditional probability of an itemset Y given another itemset X (Hahsler et al., 2005); *support*, which is the proportion of observations in the dataset which contain the itemset X (Hahsler et al., 2005); *lift*, which is the deviation of the *support* from the expected value, given independence (Hahsler et al., 2005); and, *standardized lift*, which is the lift relative to its upper and lower bounds (McNicholas et al., 2008). Standardized lift was used as the ranking method in this study as it calls upon support, confidence, and lift, and as such presents a natural and unambiguous method of ranking association rules (McNicholas et al., 2008). All analyses were conducted with a minimum level of support

of 0.005 to increase the number of rules derived, a minimum confidence level of 0.9 to ensure a sufficient level of confidence and to narrow down derived rules, and two to six items to ensure that the relationships considered are complex, but not overly so (McNicholas et al., 2008).

2.3.3.3 Well Sampling Analyses

Frequency and timing of well testing were explored in conjunction with the index sample status for each well within the recorded dataset period. To explore whether the test message returned drove well testing frequency (subsequent test or no subsequent test), all individual wells were further reclassified into four testing status categories: first test “no significant evidence” of *E. coli*, first test “no result”, first test “may be unsafe” to drink, and first test “unsafe to drink”. While this analysis could be undertaken on any two consecutive samples, almost half of the unique wells in this dataset were only tested once over the eight-year study period. As such, the initial test also represents the only “previous” test for a large proportion of wells, with “no subsequent test” being an important behavioural decision. All values for the following calculations were standardized and plotted based on total number of tests and number of tests within each testing frequency group (see A.2.2). Decay curves were then created utilizing the nonlinear least squares (nls) method (see A.2.2) to estimate the parameters (y_o , y_f , α) of the decay equation (Watson, 2020).

Utilizing this decay function, the decay rate for each initial test status was determined and compared.

Further analyses were undertaken to determine whether user testing events coincide with typical seasonal weather, such as spring thaws and summer dry-wet patterns as well as high frequencies of adverse results. User testing was determined by summing all observations in a given month of a given year. Adverse testing results for each month were standardized with respect to year (see A.2.2).

2.4 RESULTS AND DISCUSSION

Models for each driver are described and discussed in the following sections. Each model is described in (Table A.2-5) and summarized in Figure 2-2.

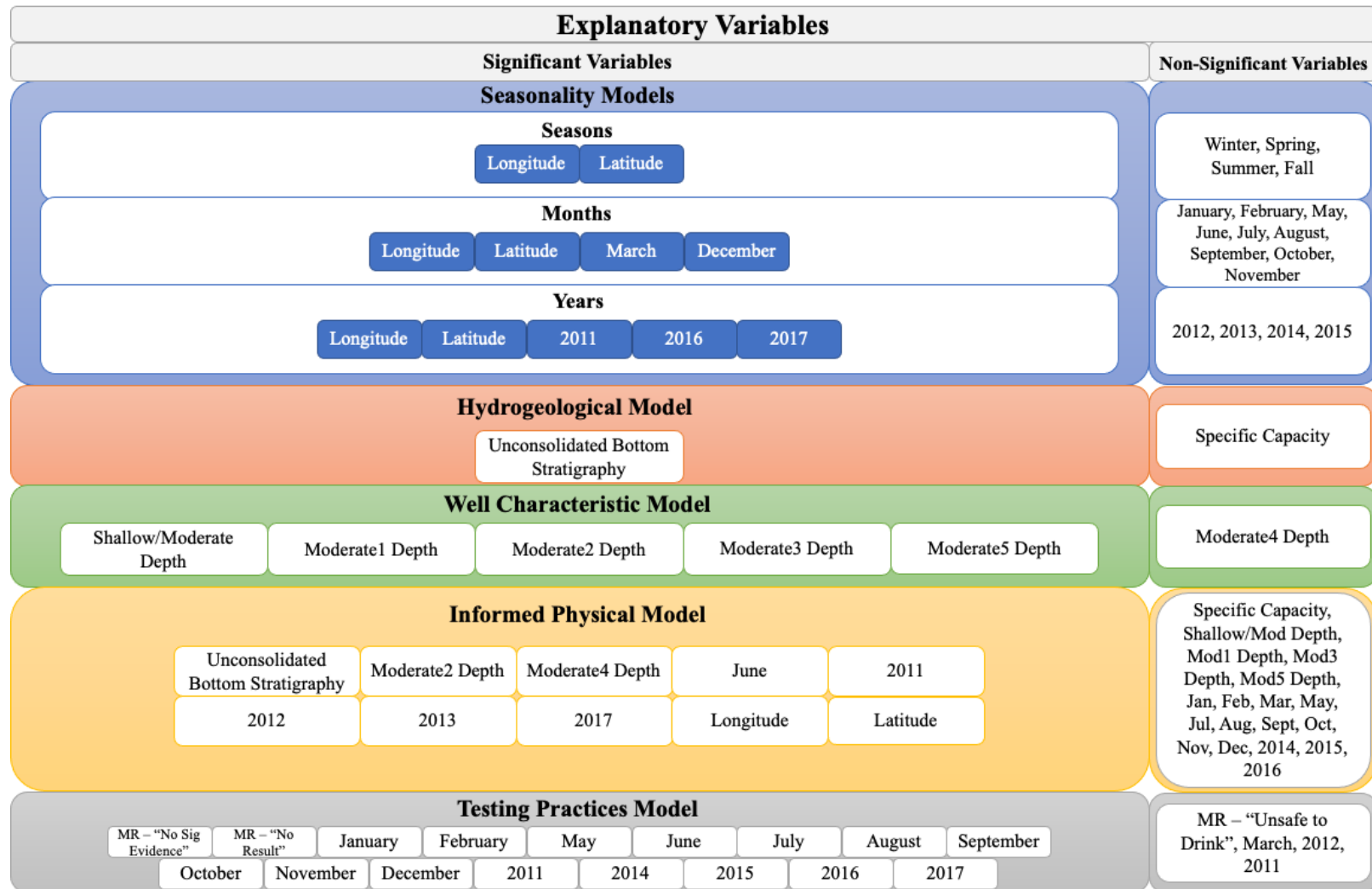


Figure 2-2: Summary of explanatory variables across “best” models for each driver (Figure 2-1).

2.4.1 Seasonal Drivers (Driver 1 in Figure 2-1)

E. coli presence and concentration in the environment is driven, in part, by seasonal changes in temperature, precipitation, and land use. Thus, an understanding of when samples are most likely to be adverse is necessary for enhanced testing awareness and recommendations. Seasonal drivers explored include season delineations and intra- and inter-annual relationships (Figure 2-2; Table A.2-5; Figures A.2.3-1-A.2.3-6). No trends emerged from the bivariate analyses, likely due to the complexity of *E. coli* fate and transport processes.

The best GAMLSS explanatory model included latitude and longitude, which were statistically significant, and Season delineation 1 (i.e., winter commencing in January), which was not statistically significant but holds explanatory value. The most significant impact on *E. coli* concentration in this model is latitude; with each increasing half-degree of latitude, *E. coli* concentrations decrease by 0.16 ± 0.01 CFU/100mL (p-value < 0.01) (Figure A.2-1; Table A.2-5). Latitude likely accounts for variations in the onset of freeze and thaw across Ontario and therefore can be considered a proxy indicator for seasonal lag. This is reflected in the 1975-2005 average first and last date for frost in different climate zones in Ontario. In a more southern zone, average first and last frosts occur on October 8th and May 3rd, respectively, compared to September 16th and June 3rd in a more northern location (OMAFRA, 2020). The more nuanced variations accounted for through latitude in particular may explain the lack of consistency in seasonal delineations within the literature (Atherholt et al., 2017; Rocha et al., 2015) as well as the lack of statistical significance for the season delineations in this model (Table A.2-5). Longitude has a

weaker relationship with *E. coli* concentration, but is a proxy for climate variations in tandem with latitude in Ontario. This is due to the presence of large bodies of water (the Great Lakes), which modify local temperature and precipitation patterns, particularly in winter. However, it is recognised that climate also drives land use and land cover, and that other variables, such as population density, vary spatially, so a compound proxy cannot be ruled out.

Seasonal delineations were subject to further refinement using individual months. The best explanatory model that emerged included all months except April, along with latitude and longitude. This model indicates that samples collected in March explain the greatest increase in *E. coli* concentrations, while samples collected in December explain the greatest decrease in *E. coli* concentrations. July emerges with some of the highest numbers of adverse *E. coli* sample results (18.2%; n = 4,699) (Figure 2-4), although these results do not represent the largest increase in *E. coli* concentrations (Figure 2-2; Table A.2-5). The agricultural season begins in April/May in Ontario, typically peaking between June and August, affecting manure spreading (Bach et al., 2002; Ontario Ministry of Environment Conservation and Parks, 2020a) and livestock grazing patterns (Invik, 2015), introducing more *E. coli* into the environment (Conboy and Goss, 2000). Further, increased ambient temperatures lead to a more sustained *E. coli* growth rate (Porter et al., 2019) coupled with increased faecal excretion rates in cattle (Invik, 2015). Increased use of summer homes increases local septic tank usage and may also increase private well water quality testing (Di Pelino et al., 2019). Further, groundwater systems are more vulnerable to microbiological contamination in the summer months when extended hot dry periods

harden the ground and lead to cracks, which act as enhanced pathogen transport pathways that are activated during sporadic heavy rainfall events (Health Canada, 2020). High user testing in July (12.2%, $n = 96,001$), coupled with increased *E. coli* loads and hydrological drivers likely contribute to the high numbers of adverse sample results in July (Health Canada, 2020). While the largest, most significant monthly increase in *E. coli* concentration occurs in March (0.25 ± 0.05 CFU/100mL, $p\text{-value} = 0.03$), March has one of the lowest numbers of user tests (6.4%, $n = 50,858$) and as a result is associated with a lower number of adverse *E. coli* observations (4.38%, $n = 1,131$) (Figure 2-4). Heavy rain and snowmelt typical of March and April (Jones et al., 2015) (not deemed explanatory by the model, so not depicted) have been associated with the flushing of *E. coli* through the system (Schuster et al., 2005). This increases risk of contamination (Health Canada, 2013), likely due to increased groundwater recharge, possibly explaining the monthly increase in *E. coli* concentrations in March and subsequent decrease in May ($p\text{-value} = 0.03$). December emerges as a month with some of the lowest numbers of adverse *E. coli* sample results (3.1%; $n = 799$) (Figure 2-4) and largest explanatory *E. coli* concentration decrease (-0.27 ± 0.05 CFU/100mL, $p\text{-value} = 0.03$). The findings for December may reflect combinations of changing processes and inputs, including frozen soils and reduced rainfall, thereby decreasing groundwater infiltration (Atherholt et al., 2017) and reducing *E. coli* availability (Bach et al., 2002). Similar to the seasons model findings, latitude is significant in the monthly model ($p\text{-value} < 0.01$), with a decrease in average *E. coli* concentration of 0.16 ± 0.01 CFU/100mL per half-degree of latitude (Figure A.2-3-Figure A.2-4; Table A.2-5).

Again, while most likely driven by climate patterns, a compound proxy cannot be ruled out.

An inter-annual assessment of *E. coli* concentration was conducted to look for trends year over year. The average *E. coli* concentration generally decreases from year to year between 2011 and 2017, with the exception of 2010 which was not identified as explanatory in the model (Figure 2-2). All years in the model except 2013, 2014, and 2015 are statistically significant and all years are statistically significantly different from one another ($p\text{-value} < 0.01$), except for 2011 to 2012 (Figure A.2-5-Figure A.2-6). The peak average *E. coli* concentrations in 2011 and 2012 are likely due to frequent flooding events causing mobilization of *E. coli* (Latchmore et al., 2020; Ontario Ministry of Environment Conservation and Parks, 2013). The years 2016 and 2017 represent a significant drop in average *E. coli* concentrations over previous years ($p\text{-value} < 0.01$), likely due to droughts in 2016 which reduced *E. coli* transport (Latchmore et al., 2020). Similar to the seasonal and monthly models, latitude and longitude are significant variables in the annual model with more northern latitudes associated with lower average *E. coli* concentrations (Figure A.2-5).

2.4.2 Hydrogeological Drivers (Driver 2 in Figure 2-1)

One of the primary drivers of pathogen transport into a well is the local hydrogeology. While the entire stratigraphic column plays a role in pathogen fate and transport, this analysis focuses on the interface between the aquifer and the well production zone. For a further exploration of the effects of overburden depth and specific bedrock types (limestone, shale, sandstone, and granite) on *E. coli* detection rates, see Latchmore et al.

(2020). The hydrogeological drivers explored here include bottom stratigraphy and specific well capacity. Among the variable groups discussed in the methods, binary bottom stratigraphy (i.e., consolidated or unconsolidated) outperformed a categorical bottom stratigraphy, averaging an improved cross validated Global Deviance.

From the dataset, initial classifications for each well were defined as consolidated (bedrock) (63.3%, $n = 499,647$) or unconsolidated (36.7%, $n = 289,426$). Of the wells completed in bedrock (i.e., consolidated), the lowest strata consisted of metamorphic (0.8%; $n = 3,814$), sedimentary (69.6%; $n = 347,958$), igneous (28.0%; $n = 139,921$), metamorphic and sedimentary (0.2%; $n = 1,086$), metamorphic and igneous (0.3%; $n = 1,712$), sedimentary and igneous (1.0%; $n = 4,774$), or all three rock types (0.1%; $n = 382$). The explanatory hydrogeological-based model summary demonstrates that an unconsolidated bottom stratigraphy increases average *E. coli* concentrations by 0.14 ± 0.02 CFU/100mL ($p\text{-value} < 0.01$), while consolidated did not provide explanatory power despite being considered a driver in the literature (Atherholt et al., 2017; Latchmore et al., 2020). To explore further, bivariate analyses were used to compare the likelihood of contamination in wells completed in consolidated (bedrock) and unconsolidated units. It was found that those completed in unconsolidated units (29.4%, $n = 7,589$) are significantly less likely to encounter contamination than those in consolidated units (70.6%, $n = 18,232$) (Table A.2-6).

An association rules analysis further examined the impact of bedrock type on *E. coli* concentrations. According to the association rules, wells completed in metamorphic bedrock had a lower probability of encountering higher *E. coli* concentrations as compared

to those completed in sedimentary bedrock. When non-detect (ND) observations were removed from the stratigraphy analyses to reduce skewing in *E. coli* concentration, wells completed in sedimentary and igneous materials had a higher probability of encountering higher *E. coli* concentrations compared to those completed in metamorphic units (Table A.2-7). These findings are supported by Conboy and Goss (2000), who found that wells completed in limestone or dolostone (76% of the sedimentary wells in the current dataset) are considered at “high risk” for pathogen contamination. The study further determined that the age of sedimentary rocks is important, as older deposits likely contain more fractures and solution channels, which act as transportation highways for pathogens (Conboy and Goss, 2000) and therefore *E. coli*. Finally, bedrock wells with minimal overburden are more likely to become contaminated due to the lack of soil available to filter pathogens before they reach fractures or channels (Conboy and Goss, 2000; Latchmore et al., 2020). Surprisingly, no association rules emerged linking *E. coli* concentrations with either bottom stratigraphy permeability or specific capacity, likely due to small numbers of observations in some subcategories.

2.4.3 Well Characteristics (Driver 3 in Figure 2-1)

Well characteristics impact the physical integrity of the well and thus influence *E. coli* ingress (Di Pelino et al., 2019). The well characteristics explored here include well depth and year of well construction. As with the hydrogeological drivers, a selection between categorical and continuous variables for well depth was undertaken, and it was determined that categorical well depth improved explanatory power. It should be noted that a smaller number of categories was originally used to align with provincial regulations (shallow,

moderate, and deep; Ontario Ministry of Environment Conservation and Parks, 2019). However, this is a skewed distribution that resulted in findings that were in contradiction to conventional understanding and could not be explained using process-based logic. This is underscoring the fact that machine learning must be used in combination with disciplinary expertise to ensure relevant models (Reichstein et al., 2019).

Well depth categories up to moderate³ were found to increase average *E. coli* concentrations by 0.14 ± 0.03 - 0.20 ± 0.04 CFU/100mL (p-value = 0.01) (Figure A.2-7-Figure A.2-8). Typically, well users assume that deep wells are protected from contamination in comparison to moderate and shallow depth wells (Kreutzwiser et al., 2010). However, the fact that deep wells were not found to be explanatory of *E. coli* concentration serves as a reminder that greater depths are not protective against contamination. A supplementary bivariate analysis underscores this finding; shallow wells were significantly different (p-value < 0.05) from deep wells at all *E. coli* concentrations, with shallow wells being more likely to return adverse samples (p-value ≤ 0.0027) (Table A.2-8). As such, increased depth is not a reason to assume sufficient protection of drinking water quality. Given a risk of complacency regarding the microbiological safety of deep wells, these findings could represent a public health threat to the 15% of well users who rely on deep wells in this dataset.

2.4.4 Informed Physical Model

The findings from the seasonal, hydrogeological, and well characteristic models were used to create an informed physical model (Figure 2-2, Figure A.2-9-Figure A.2-10). Based on model outputs (Table A.2-5) and subject matter expertise, the most explanatory

variables were combined into a single model to explore relative importance of driver variables. The final model consisted of binary bottom stratigraphy, continuous specific capacity, categorized well depth, month of test, year of test, longitude, and latitude. General trends in the informed physical model aligned with those of the individual models. The model is most sensitive to specific capacity followed by bottom of well stratigraphy, year, latitude, well depth, month, and finally longitude. This is a particularly interesting finding as the specific capacity of a well is not typically considered to be a driver of contamination risk. More work needs to be done to determine whether specific capacity is a driver of contamination, or if the model is selecting specific capacity as a proxy for other factors (e.g., high permeability related to the presence of fractures).

These results demonstrate that machine learning techniques employed in combination with disciplinary expertise are useful for developing data-driven explanatory models of the relationship between *E. coli* concentrations in private wells and the drivers of this contamination. Indeed, relative sensitivity to specific capacity makes sense but also highlights a variable that is not normally considered in this context and thus requires further process-based analysis.

2.4.5 Testing Practices (Driver 4 in Figure 2-1)

Water quality testing is critical because it is the only way to characterize well water quality, which provides important information for both well stewardship practices and human health protection. A regression analysis of testing patterns revealed that individual wells were tested on average 2.70 ± 0.004 times over the 8-year study period (Table A.2-5). Critically, this dataset does not represent all private drinking water wells in Ontario.

Many wells were excluded due to incomplete information, the inability to match a water test record to a well record, or never having had a sample submitted to a provincial laboratory for testing. As such, given estimates of the number of wells in Ontario (Ontario Ministry of Environment Conservation and Parks, 2020b), there may be approximately 345,000 additional wells not captured by this dataset that, by definition, would be classified as sampled fewer than 16 times. According to the current dataset, 98% ($n=245,708$) of individual wells were tested less than the two times per year threshold (≤ 16 tests between 2010 and 2017), with 48% ($n=119,670$) only testing once over the eight-year period. This limited testing may be attributable to complacency (e.g., history of non-adverse sample results or no concern for colour or odour), no experience of adverse health effects, or inconvenience (e.g., limited hours at sample drop off locations) (Invik, 2015).

Using regression analyses, user testing frequency was found to be impacted by the sample result message received (excluding “may be unsafe”, as it was not deemed important by the explanatory model, so not depicted), month of user test, and year of user test (Figure 2-2; Figure A.2-11-Figure A.2-12). The return of an “unsafe to drink” message, while not statistically significant, slightly increased the number of samples taken by 0.02 ± 0.008 over the study period, likely due to the health concern that this result represents to the user (Table A.2-5). A status that is returned as “no result” (i.e., processing issues, chemical testing requested, appearance or order unacceptable, interfering substances, unauthorized submitter) or “no significant evidence” was found to, on average, decrease the number of tests submitted (0.13 ± 0.006 and 0.05 ± 0.004 , respectively), likely due to the non-alarming nature of the message (Table A.2-5).

To explore user testing practices further, the message received for the first test was assessed as an indicator of subsequent testing. It was found that two fitted decay equations were required to best characterize the data; one for 15 or fewer tests and one for 16 or more tests, as the decay rates are different. For 15 or fewer tests, if the first test message was “no significant evidence” (73%, $n=183,608$ of initial samples), the well user was less likely to continue testing (decay rate = 0.96) as compared to when initial samples were “no result” (6%, $n=15,816$) (decay rate = 0.42, $p\text{-value} = < 0.0001$), or “may be unsafe” (13%, $n=31,656$) (decay rate = 0.40, $p\text{-value} = < 0.0001$), or “unsafe to drink” (8%, $n=20,341$) (decay rate = 0.40, $p\text{-value} = < 0.0001$) (Figure 2-3). Qayyum et al. (2020) found that a well that received an initial negative index test (not containing *E. coli* or total coliforms) retested 64% of the time, when compared to a 74% retesting rate when the initial index test is positive (containing *E. coli* or total coliforms). This reflects similar trends to those found in this work – decay rates for retesting were highest (i.e., less retesting) when the index test was “no significant evidence” (Qayyum et al., 2020). Additionally, all curves except “may be unsafe” and “unsafe to drink” are statistically significantly different from one another. The use of the word “unsafe” is likely a flag for concern amongst well users. Well users testing 16 or more times over the 8-year study period are likely to be routine well testers. While all of these decay curves are statistically significantly different from one another, decay rates fall within a smaller range than those who test 15 or fewer times (“no significant evidence”, decay rate = 0.11845, $p\text{-value} = < 0.0001$; “no result”, decay rate = 0.16080, $p\text{-value} = < 0.0001$; “may be unsafe”, decay rate = 0.15107, $p\text{-value} = < 0.0001$; “unsafe to

drink”, decay rate = 0.12197, p-value = < 0.0001) (Table A.2-9). Routine testing indicates greater awareness of appropriate well stewardship practices (Lavallee et al., 2020).

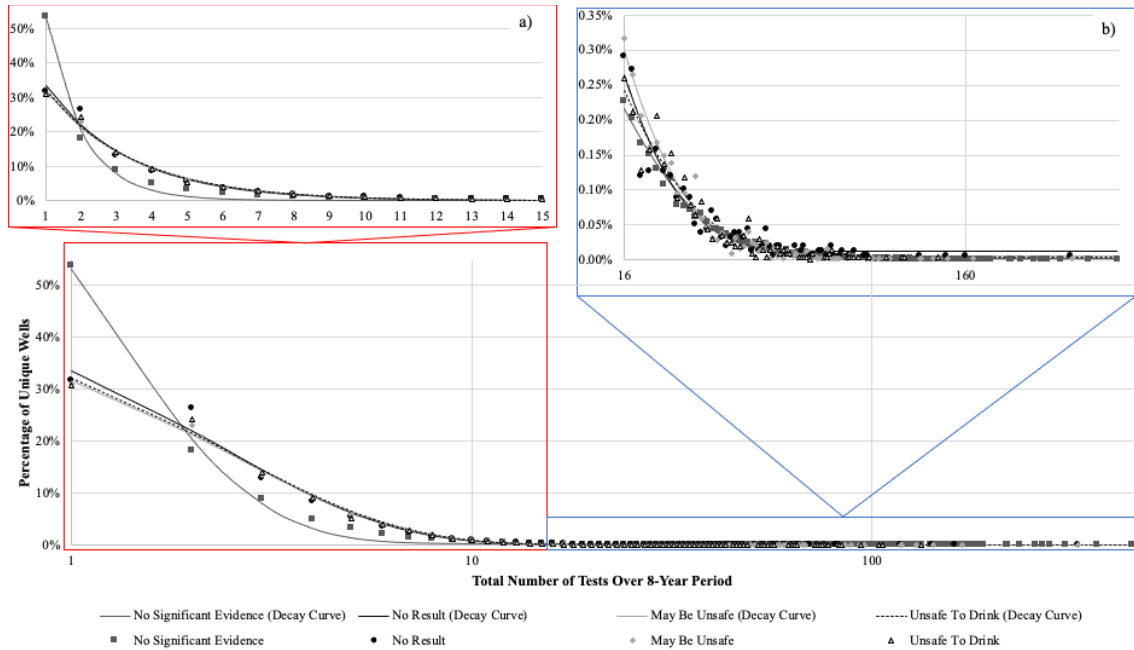


Figure 2-3: Percentage of individual wells versus number of times tested over the eight-year period given an initial sample that was “no significant evidence” (73% of wells), “no result” (6%), “may be unsafe” (13%), or “unsafe to drink” (8%). Insets show a) under two times per year threshold (i.e., 1-15 tests), and b) at or over two times per year threshold tested tail (i.e., 16-446 tests) of this curve, respectively.

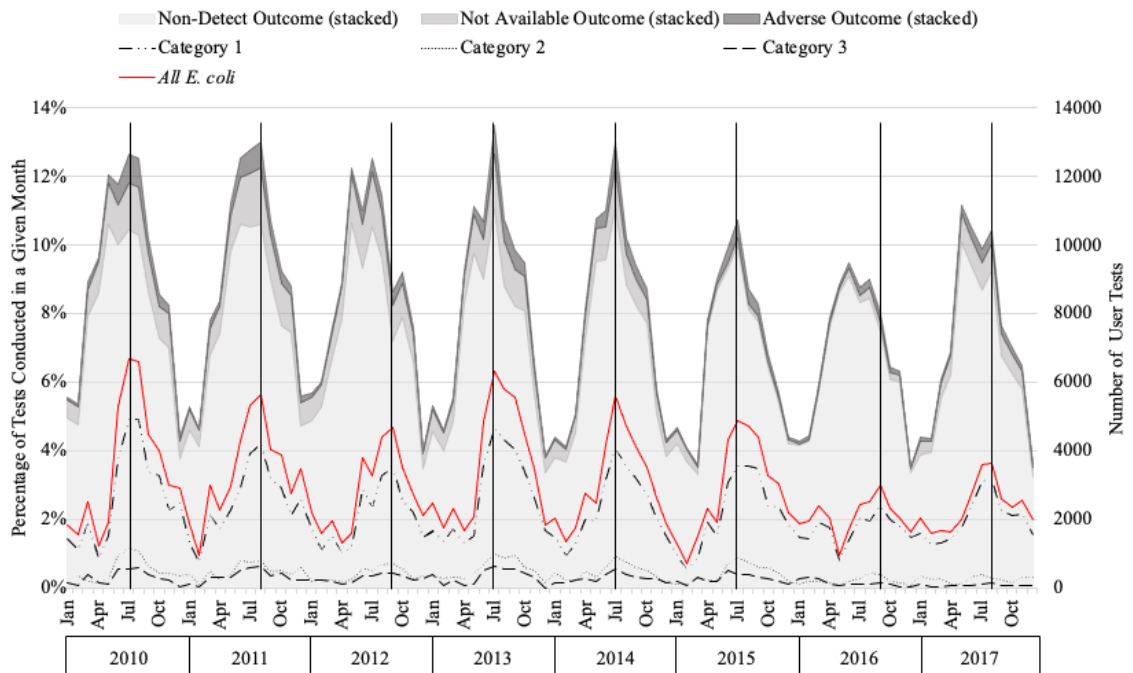


Figure 2-4: Occurrence of private well testing and adverse results over the study period, where non-detects are defined as 0 CFU/100mL, “All *E. coli*” represents the summation of the three *E. coli* concentration categories. Vertical bars are extensions of peak adverse points each year.

The majority of samples were submitted for testing in July (Figure 2-4), representing the second highest increase in user testing frequency (month over month). It is postulated that this may represent seasonal testers. Further, July is more conducive weather for driving, as well as the start of summer holiday season in Ontario when people may have more time or are using seasonal residences (and associated wells). Interestingly, the month with the largest impact on user testing frequency is January (0.26 ± 0.01 times). This may be because anyone testing in January is probably more consistent in their testing regime as January does not represent a month that is communicated as a critical testing period, and thus has the fewest user tests (Figure 2-4).

While in some years (2010, 2011, 2013) testing frequency coincided with peak adverse sample occurrences (i.e., July), this was not the case in 2012 and 2014-2017, when peak adverse sample occurrences shifted to as late as September. This offset between testing and peak *E. coli* contamination events raises the question of timing of future peaks in adverse occurrence, and whether well users have adequate contamination risk information for optimal well testing practices. Starting in 2014 there is a general trend of decreasing user testing frequency (Figure A.2-11), with no obvious explanation. Lack of user testing compliance, combined with a divergence between peak testing and *E. coli* contamination periods in latter years, and the decreasing trend in testing frequency (Figure 2-4), underscore the need to better understand well user behaviours, provide additional resources, and target educational campaigns. More specifically, enhanced methods are required to predict and communicate risk of potential contamination events to well users and there is a need for evidence-based testing regime guidelines. Increased well user outreach will improve knowledge, attitudes, and practices with respect to well sampling and testing to move well user sampling practices closer to the “temporal truth” (Latchmore et al., 2020) of *E. coli* contamination events.

2.5 STUDY LIMITATION

The WWTD is subject to methodological limitations associated with *E. coli* quantification; however, the uncertainties cannot be quantified with the available data. Further, the WWIS database is subject to data entry errors, as borehole logs are hand recorded in the field and later transcribed into an online database, some of which date back to the 1910's. This was addressed through the removal of outliers that were outside the

range of realistic values i.e., specific capacities less than zero. There is no indication that these outliers are systematic.

2.6 CONCLUSION

To the authors' knowledge, supervised machine learning approaches such as GAMLSS and Association Rule Analysis have not previously been used to assess *E. coli* contamination risk in private wells. The approaches in combination with a large private well dataset enabled the development of explanatory models of *E. coli* concentration as a function of seasonal, hydrogeologic, and well characteristic drivers. Consensus with existing literature for many findings confirms the validity of this novel approach, which also identified drivers that are not typically considered but are supported by a process-based understanding of the system. As such, these findings also demonstrate the importance of coupling machine learning approaches with disciplinary expertise. This opens up opportunities to develop better tools to understand drivers and predict contamination that can be used to evaluate and mitigate public health risk and inform better policy and stewardship practices. The results provide valuable insight into drivers of *E. coli* contamination, their relative importance, and therefore potential public health risks associated with the use of private wells in Ontario. Specifically, the following key findings were uncovered.

- The best delineation for the seasonal variable identified winter as starting in January. However, the seasonal variable was not found to be as important as latitude to explain intra-annual variations in *E. coli* concentrations due to the spatial variability of climate patterns in Ontario. Specifically, latitude was found to better

represent spatial variations in the onset of seasonal freeze and thaw events that drive *E. coli* concentrations and therefore should replace seasonal lag factors.

- The use of months as a variable demonstrates the ability to capture more granular changes in interannual peak *E. coli* concentrations. The shift in peak *E. coli* concentrations to later in the year is a finding that requires further investigation.
- Bedrock wells completed in sedimentary and igneous formations are more likely to have higher *E. coli* concentrations when compared to those completed in metamorphic or unconsolidated formations. Previously, igneous and metamorphic formations have not been differentiated in this manner.
- *E. coli* contamination is statistically significantly impacted by well depth; generally, wells up to a depth of approximately 60m are more likely to become contaminated with *E. coli*. While this is congruent with the literature, the depth threshold warrants further investigation. Further, deep wells do not emerge as reducing *E. coli* contamination. As such, increased depth does not guarantee that a contamination event will not occur - testing and stewardship are still required.
- The informed physical model was most sensitive to specific capacity followed by bottom of well stratigraphy, year, latitude, well depth, month, and longitude. The specific capacity of a well is not typically associated with contamination risk and therefore warrants further investigation.
- Testing frequency was significantly impacted by initial test message received. Frequency increased with an “unsafe to drink” result and decreased with “no significant evidence” and “no result” messages. This finding confirms the need for

well users to be educated on the temporal changes of *E. coli* contamination of groundwater wells, the impacts of the physical environment and well characteristics on *E. coli* concentrations in their well, and the need for an informed, regular testing regime to protect their health.

- While, in general, there is a correlation between when users test their wells and when the greatest frequencies and concentrations of adverse results occur, a decoupling can be observed in recent years where testing remains highest in July but peaks in adverse results occur up to three months later. This finding has potential implications for the health of well users as they may not be capturing peak *E. coli* contamination events in their wells.

This study demonstrates that a coupled-systems approach that applies machine learning techniques in combination with a large, multi-dimensional dataset can support and advance our understanding of geo-spatio-temporal relationships and interconnections that impact *E. coli* contamination in private wells. Recognition of these interconnections offers an innovative path forward for enhancing private well user awareness and stewardship. The identification of explanatory variables, their relative importance, and effects on *E. coli* concentration, in combination with other data sets (e.g., meteorological) can be used to inform and advance the development of future predictive data-driven fate and transport models.

2.7 WORKS CITED

- Atherholt, T.B., Procopio, N.A., Goodrow, S.M., 2017. Seasonality of Coliform Bacteria Detection Rates in New Jersey Domestic Wells. *Groundwater* 55, 346–361. <https://doi.org/10.1111/gwat.12482>
- Bach, S.J., McAllister, T.A., Veira, D.M., Gannon, V.P.J., Holley, R.A., 2002. Transmission and control of *Escherichia coli* O157:H7 - A review. *Can. J. Anim. Sci.* 82, 475–490. <https://doi.org/10.4141/A02-021>
- Buckerfield, S.J., Quilliam, R.S., Bussiere, L., Waldron, S., Naylor, L.A., Li, S., Oliver, D.M., 2020. Chronic urban hotspots and agricultural drainage drive microbial pollution of karst water resources in rural developing regions. *Sci. Total Environ.* 744, 1–10. <https://doi.org/10.1016/j.scitotenv.2020.140898>
- Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012. Data-driven dynamic emulation modelling for the optimal management of environmental systems. *Environ. Model. Softw.* 34, 30–43. <https://doi.org/10.1016/j.envsoft.2011.09.003>
- Conboy, M.J., Goss, M.J., 2000. Natural protection of groundwater against bacteria of fecal origin. *J. Contam. Hydrol.* 43, 1–24. [https://doi.org/10.1016/S0169-7722\(99\)00100-X](https://doi.org/10.1016/S0169-7722(99)00100-X)
- Di Pelino, S., Schuster-Wallace, C., Hynds, P.D., Dickson-Anderson, S.E., Majury, A., 2019. A coupled-systems framework for reducing health risks associated with private drinking water wells. *Can. Water Resour. J.* 44, 280–290. <https://doi.org/10.1080/07011784.2019.1581663>
- Fong, T.T., Mansfield, L.S., Wilson, D.L., Schwab, D.J., Molloy, S.L., Rose, J.B., 2007. Massive microbiological groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environ. Health Perspect.* 115, 856–864. <https://doi.org/10.1289/ehp.9430>

- Foster, S.S.D., Chilton, P.J., 2003. Groundwater: The processes and global significance of aquifer degradation. *Philos. Trans. R. Soc. B Biol. Sci.* 358, 1957–1972. <https://doi.org/10.1098/rstb.2003.1380>
- Hahsler, M., Grün, B., Hornik, K., 2005. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *J. Stat. Softw.* 14. <https://doi.org/10.18637/jss.v014.i15>
- Health Canada, 2020. Guidelines for Canadian Drinking Water Quality.
- Health Canada, 2013. Guidance for Providing Safe Drinking Water in Areas of Federal Jurisdiction.
- Invik, J., 2015. Total Coliform and Escherichia Coli Contamination in Rural Well Water in Alberta, Canada: Spatiotemporal Analysis and Risk Factor Assessment. ProQuest Diss. Theses. University of Calgary. <https://doi.org/10.11575/PRISM/28466>
- Jones, N.E., Petreman, I.C., Schmidt, B.J., 2015. High Flows and Freshet Timing in Canada : Observed Trends. Peterborough, Ontario.
- Joshi, A. V, 2020. Machine Learning and Artificial Intelligence, Machine Learning and Artificial Intelligence. <https://doi.org/10.1007/978-3-030-26622-6>
- Knoll, L., Breuer, L., Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Sci. Total Environ.* 668, 1317–1327. <https://doi.org/10.1016/j.scitotenv.2019.03.045>
- Kreutzweiser, R., de Loë, R.C., Imgrund, K., 2010. Out of Sight, Out of Mind: Private Water Well Stewardship in Ontario. Report on the Findings of the Ontario Household Water Well Owner Survey 2008, Water Policy and Governance Group, University of Waterloo, Waterloo, ON.
- Latchmore, T., Hynds, P., Brown, S., Schuster-Wallace, C., Dickson-Anderson, Sarah McDermott, K., Majury, A., 2020. Analysis of a Large Spatiotemporal

Groundwater Quality Dataset, Ontario 2010 - 2017: Informing Human Health Risk Assessment and Testing Guidance for Private Drinking Water Wells.

Lavallee, S., Hynds, P.D., Brown, R.S., Schuster-Wallace, C., Dickson- Anderson, S., Di Pelino, S., Egan, R., Majury, A., 2020. Examining influential drivers of private well users' perceptions in Ontario: A cross-sectional population study. *Sci. Total Environ.* 142952. <https://doi.org/10.1016/j.scitotenv.2020.142952>

McNicholas, P.D., Murphy, T.B., O'Regan, M., 2008. Standardising the lift of an association rule. *Comput. Stat. Data Anal.* 52, 4712–4721. <https://doi.org/10.1016/j.csda.2008.03.013>

Murphy, H.M., Prioleau, M.D., Borchardt, M.A., Hynds, P.D., 2017. Review: Epidemiological evidence of groundwater contribution to global enteric disease, 1948–2015. *Hydrogeol. J.* 25, 981–1001. <https://doi.org/10.1007/s10040-017-1543-y>

Murphy, H.M., Thomas, M.K., Schmidt, P.J., Medeiros, D.T., McFadyen, S., Pintar, K.D.M., 2016. Estimating the burden of acute gastrointestinal illness due to *Giardia*, *Cryptosporidium*, *Campylobacter*, *E. coli* O157 and norovirus associated with private wells and small water systems in Canada. *Epidemiol. Infect.* 144, 1355–1370. <https://doi.org/10.1017/S0950268815002071>

O'Dwyer, J., Hynds, P.D., Byrne, K.A., Ryan, M.P., Adley, C.C., 2018. Development of a hierarchical model for predicting microbiological contamination of private groundwater supplies in a geologically heterogeneous region. *Environ. Pollut.* 237, 329–338. <https://doi.org/10.1016/j.envpol.2018.02.052>

OMAFRA, 2020. Climate Zones and Planting Dates for Vegetables in Ontario [WWW Document]. Veg. Crop. URL
<http://www.omafra.gov.on.ca/english/crops/facts/climzoneveg.htm> (accessed 2.16.21).

- Ontario Ministry of Environment Conservation and Parks, 2020a. Nutrient Management Act.
- Ontario Ministry of Environment Conservation and Parks, 2020b. Well records - WWIS - Microsoft Access - Ontario Data Catalogue.
- Ontario Ministry of Environment Conservation and Parks, 2019. Water Supply Wells: Requirements and Best Practices.
- Ontario Ministry of Environment Conservation and Parks, 2013. Canada's Top Ten Weather Stories Archive [WWW Document]. URL <https://www.ec.gc.ca/meteo-weather/meteo-weather/default.asp?lang=En&n=3318B51C-1>
- Porter, K.D.H., Quilliam, R.S., Reaney, S.M., Oliver, D.M., 2019. High resolution characterisation of *E. coli* proliferation profiles in livestock faeces. *Waste Manag.* 87, 537–545. <https://doi.org/10.1016/j.wasman.2019.02.037>
- Qayyum, S., Hynds, P., Richardson, H., McDermott, K., Majury, A., 2020. A geostatistical study of socioeconomic status (SES), rurality, seasonality and index test results as drivers of free private groundwater testing in southern Ontario, 2012–2016. *Sci. Total Environ.* 717. <https://doi.org/10.1016/j.scitotenv.2020.137188>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195-204.
- Rigby R.A. and Stasinopoulos D.M., 2005. Generalized additive models for location, scale and shape (with discussion), *Appl. Statist.*, 54, part 3, pp 507-554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- Rigby, R., Stasinopoulos, M., Heller, G., De Bastiani, F., 2019. Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R, CRC Press.
- Rivera, A., 2017. The state of ground water in Canada. Gr. Water Canada.

- Rocha, C., Wilson, J., Scholten, J., Schubert, M., 2015. Retention and fate of groundwater-borne nitrogen in a coastal bay (Kinvara Bay, Western Ireland) during summer. *Biogeochemistry* 125, 275–299. <https://doi.org/10.1007/s10533-015-0116-1>
- Sainani, K.L., 2014. Explanatory versus predictive modeling. *PM R* 6, 841–844. <https://doi.org/10.1016/j.pmrj.2014.08.941>
- Samadi, S., Tufford, D.L., Carbone, G.J., 2018. Estimating hydrologic model uncertainty in the presence of complex residual error structures. *Stoch. Environ. Res. Risk Assess.* 32, 1259–1281. <https://doi.org/10.1007/s00477-017-1489-6>
- Schuster, C.J., Ellis, A.G., Robertson, W.J., Dominique, F., Aramini, J.J., Marshall, B.J., Medeiros, D.T., 2005. Infectious Disease Outbreaks Related to Drinking Water in Canada, 1974-2001 96, 254–258.
- Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 23, 1–46. <https://doi.org/10.18637/jss.v023.i07>
- UN Water, 2015. Water for a Sustainable World, The United Nations World Water Development Report.
- Watson, D., 2020. Fitting exponential decays in R [WWW Document]. URL https://douglas-watson.github.io/post/2018-09_exponential_curve_fitting/ (accessed 11.2.20).
- Yosri, A., Dickson-Anderson, S., Siam, A., El-Dakhakhni, W., 2021. Transport pathway identification in fractured aquifers: A stochastic event synchrony-based framework. *Adv. Water Resour.* 147, 103800. <https://doi.org/10.1016/j.advwatres.2020.103800>

**CHAPTER 3 CONVERTING LAND USE – LAND COVER
TO *E. COLI* CONTAMINATION POTENTIAL
CLASSES FOR GROUNDWATER WELLS:
UTILIZING A LARGE ONTARIO-BASED DATASET**

Summary of Paper 2: Converting Land Use – Land Cover to *E. coli* Contamination Potential Classes for Groundwater Wells: Utilizing a Large Ontario-based Dataset, by K.White, C. Schuster-Wallace, and S. Dickson-Anderson, Journal of Applied Spatial Analysis and Policy, Under Review, ASAP-D-23-00035.

Maps of Major Land Use - Land Cover Categories and Respective *E. coli* Contamination Potential Classes for Groundwater Wells in Ontario, by K. White, S. Dickson-Anderson, and C. Schuster-Wallace, Federated Research Data Repository, doi: 10.20383/102.0530. 2023.

Summary:

This research expands on the first chapter of this thesis, exploring a new Land Use-Land Cover (LULC) dataset to advance the understanding of the relationship between LULC and *E. coli* presence in wells. This work utilizes a novel approach by combining an innovative IPCC approach to consistent treatment of uncertainties of findings with a supervised regression-based machine learning approach, GAMLSS, to identify and support relationships between *E. coli* presence in wells and LULC categories. The results of this demonstrated:

- The innovative new method for assessing the impact of LULC on *E. coli* contamination potential for wells was successful in creating a geospatial raster dataset based on regression-based machine learning approaches and literature consensus methods.

- The highest *E. coli* contamination potential was in areas used for pastoral/agricultural purposes, followed by urban uses, aggregate mines, and open/barelands, followed by forests, bedrock, scrublands, and disturbed lands, and finally the lowest *E. coli* contamination potential classes included wetlands, water, and grasslands.
- Mapping from this research can be used to inform policy makers of geographic based well characteristic best practices, locations for wellhead protection areas, and targeted outreach campaigns.

3.1 ABSTRACT

Land Use-Land Cover (LULC) types have been used as a proxy for *E. coli* sources and transport mechanisms. This study aims to advance the understanding of the relationship between LULC and *E. coli* presence in wells for 11 major LULC categories. This represents a novel approach for assessing the potential for well contamination and consequent groundwater management strategies.

The approach combines the IPCC approach to consistent treatment of uncertainties of findings in the existing literature with new understanding from a combination of large datasets. Generalized Additive Models for Location, Shape, and Scale (GAMLSS) regression analyses were used to identify and support relationships between a large dataset of *E. coli* presence in wells and LULC data.

Risk classes were determined from a combination of literature classification and regression analyses. A raster dataset for Ontario that identifies areas of low to very high potential for *E. coli* presence in wells was created from these findings. However, gaps remain in understanding the relationship between some LULC categories and the presence of *E. coli* in wells. This approach can be applied to other large datasets and broader geographic regions to address these gaps and generate similar raster datasets. These outcomes are instructive in the development of land management and source water protection plans, as well as well stewardship strategies.

Keywords: Groundwater Policy, Machine Learning, Large Dataset, Private Well, *E. coli* Contamination

3.2 INTRODUCTION

E. coli contamination of groundwater requires the presence of an *E. coli* source. Some typical sources of *E. coli* include leaking septic tanks, manure-based fertilizers, and domestic and wild animal waste. Because these sources occur in specific types of landscapes, Land Use - Land Cover (LULC) types have often been explored as a predictor of *E. coli* prevalence in the natural environment (Dusek et al., 2018; Gregory et al., 2019; Jabbar et al., 2019). Pastoral and agriculture and urban LULC types have received the most attention to date in the literature. However, consideration and assessment of other LULCs, such as scrubland, disturbance, and bedrock, typically associated with lower *E. coli* prevalence, are not present in the literature.

The transport of *E. coli* into groundwater from surface deposition requires a mechanism (e.g., rainfall) and a flow path such as root or insect holes, fractures, or improperly sealed or abandoned wells. While a source and pathway are required for *E. coli* to be transported from the land into groundwater, additional conditions must also be met for *E. coli* to be present in a well. Specifically, there must either be *E. coli* present in the groundwater supplying the well screen or it must be hydraulically connected to the surface, for example through a cracked well head or casing. Conversely, some environments effectively remove *E. coli*, for example some types of vegetation (antimicrobial properties) and soils (filtration). LULCs are useful proxies because they can be used to summarize the human and animal activity in an area, which are indicative of *E. coli* sources, as well as represent potential transport characteristics based on the typical physical environment.

The goal of this work is to uncover new knowledge and improve upon existing knowledge of relationships between *E. coli* in the natural environment and presence in wells using LULC categories as a surrogate measure of *E. coli* loading and transport through the subsurface. To achieve this, the first objective is to aggregate minor LULC categories ($n = 32$) into more descriptive and simplified categories. The second objective is to explore the relationships between LULC categories from the first objective ($n = 11$) and *E. coli* presence in wells utilizing a large private well dataset (White et al., 2021). This novel approach for assessing the potential for well contamination was developed using Ontario, Canada as a case study. The approach results in two raster datasets: the first depicting the simplified LULC categorization structure (amalgamating similar LULC categories into more comprehensive categories), the second depicting the *E. coli* contamination potential classes for private wells based on LULC. Together, these datasets will provide data to inform policy that impacts LULC driven *E. coli* contamination, particularly with respect to private well characteristics, public health-based outreach strategies, land use and management plans, and delineation of source water protection zones.

3.3 METHODS

3.3.1 Data Sources

LULC raster datasets were acquired from the Ontario GeoHub, which is a part of the Ontario Open Government Initiative. Specifically, the Southern Ontario Land Resource Information System (SOLRIS) 3.0 (Ontario Ministry of Natural Resources and Forestry, 2019) and the Ontario Land Cover Compilation (OLCC) 2.0 (Ontario Ministry of Natural

Resources and Forestry, 2016) were used as input databases. While the OLCC 2.0 covers all of Ontario (96.63° W to 71.67° W and 40.65° N to 57.44° N), it utilizes the outdated SOLRIS 1.2 to classify Southern Ontario. In more recent years, the SOLRIS 3.0 was developed utilizing improved methods and data captured from 2000 to 2015. For this reason, the combined dataset developed in this work updates the Southern Ontario region in the OLCC 2.0 dataset with the SOLRIS 3.0 data (83.10° W to 74.33° W and 41.67° N to 45.88° N). This was possible as both datasets utilise the same resolution (15m), projection (MNR Lambert Conformal Conic), and datum (North American 1983 CSRS). *E. coli* occurrences in wells, with the respective geographic locations, were obtained from an Ontario-specific groundwater dataset that combines the Ontario Water Well Information System (WWIS) and the Water Well Testing Database (WWTD) (White et al., 2021). The groundwater dataset originally consists of 795,023 well sample observations for 253,136 unique private wells that have been tested 1 to 446 times between 2010 and 2017, inclusive. Once the LULC and groundwater dataset were combined for analysis, data categorised as “not available” were removed, resulting in a final dataset of 709,485 water sample observations. Within this combined dataset, 96% of water sample observations are non-detect for *E. coli* (0 CFU/100 mL), and 4% of water sample observations report an *E. coli* concentration in CFU/100 mL.

3.3.2 LULC Categorization

The OLCC 2.0 and SOLRIS 3.0 datasets consist of 29 and 32 LULC categories, respectively. The different LULC categories in these datasets needed to be synthesized and standardised before being combined into a single dataset. Based on category descriptions,

all minor categories (i.e., original 29 categories for OLCC 2.0 and 32 categories for SOLRIS 3.0) were reclassified into 11 major categories using the minor category descriptions to aggregate similar LULC categories. A No Data category was used to categorize pixels where data are unavailable due to cloud cover or inaccuracies; $n = 1,027$ obs (Table 3-1). The OLCC 2.0 and the SOLRIS 3.0 were combined using the new summarized major LULC categories to create a new updated raster dataset (White et al., 2022, Figure 3-1).

Table 3-1: Summary of major LULC categories.

| Unit Name | Code Number | Unit Description | Source Unit Name |
|--|-------------|---|---|
| Pastoral/ Agricultural (n = 276, 308 obs) | 1 | Continuous row crops and mixed crops are rotated with perennial crops. Also includes hay/pasture, orchards, vineyards, nurseries, rural properties not in production, urban brown fields, and clearings within forests. | SOLRIS 3.0: Undifferentiated |
| | | | SOLRIS 3.0: Tilled |
| Urban (n = 221, 890 obs) | 2 | Highways, roads, linear frequencies of structures more than 10 per 500m or 4 per 1-hectare box. Green space or permeable surfaces can be present. | OLCC 2.0: Agriculture and Undifferentiated |
| | | | |
| Aggregate Mines (n = 308 obs) | 3 | Open-pit aggregate extraction site. Includes associated infrastructure such as roads, buildings, weigh scales, and ponds. | SOLRIS 3.0: Transportation |
| | | | SOLRIS 3.0: Built-Up Area Pervious |
| Open/Bareland (n = 153 obs) | 4 | Unconsolidated materials subject to active processes (i.e., wave energy, erosion, etc.). Containing <25% tree or shrub coverage. | SOLRIS 3.0: Built-Up Area Impervious |
| | | | OLCC 2.0: Community/Infrastructure |
| Aggregate Mines (n = 308 obs) | 3 | Open-pit aggregate extraction site. Includes associated infrastructure such as roads, buildings, weigh scales, and ponds. | SOLRIS 3.0: Extraction – Aggregate |
| | | | OLCC 2.0: Sand/Gravel/Mine Tailing/Extraction |
| Open/Bareland (n = 153 obs) | 4 | Unconsolidated materials subject to active processes (i.e., wave energy, erosion, etc.). Containing <25% tree or shrub coverage. | SOLRIS 3.0: Open Beach/Bar |
| | | | SOLRIS 3.0: Open Alvar |
| Open/Bareland (n = 153 obs) | 4 | Unconsolidated materials subject to active processes (i.e., wave energy, erosion, etc.). Containing <25% tree or shrub coverage. | SOLRIS 3.0: Open Sand Barren and Dune |
| | | | SOLRIS 3.0: Treed Sand Barren and Dune |
| Open/Bareland (n = 153 obs) | 4 | Unconsolidated materials subject to active processes (i.e., wave energy, erosion, etc.). Containing <25% tree or shrub coverage. | OLCC 2.0: Open Cliff and Talus |
| | | | |

| | | | |
|--|---|---|---------------------------------------|
| Scrubland (n = 21, 448 obs) | 5 | Often tree cover <25%, shallow substrates, low shrubs (<2m). Can have rapidly draining soils, may contain exposed bedrock. | OLCC 2.0: Alvar |
| | | | OLCC 2.0: Sand Barren and Dune |
| | | | SOLRIS 3.0: Fen |
| | | | SOLRIS 3.0: Sparse Treed |
| | | | SOLRIS 3.0: Shrub Alvar |
| | | | SOLRIS 3.0: Treed Alvar |
| | | | OLCC 2.0: Fen |
| | | | OLCC 2.0: Heath |
| | | | OLCC 2.0: Sparse Treed |
| | | | SOLRIS 3.0: Extraction – Peat/Topsoil |
| Disturbance (n = 1, 658 obs) | 6 | Includes peat removed for consumptive (e.g., trees and peat) and non-consumptive (e.g., conservation, erosion protection) uses. Includes forest clear cut or burns <10 years old (low trees, dead trees, mosses, herbaceous). | OLCC 2.0: Disturbance |
| Bedrock (n = 3, 204 obs) | 7 | Exposed bedrock, often times <25% vegetation. | SOLRIS 3.0: Open Cliff and Talus |
| | | | SOLRIS 3.0: Treed Cliff and Talus |
| | | | SOLRIS 3.0: Open Bareland |
| | | | OLCC 2.0: Bedrock |
| | | | SOLRIS 3.0: Forest |
| Forest (n = 136, 151 obs) | 8 | >60% tree cover (coniferous, deciduous, mixed). | SOLRIS 3.0: Deciduous Forest |
| | | | SOLRIS 3.0: Mixed Forest |
| | | | SOLRIS 3.0: Coniferous Forest |
| | | | SOLRIS 3.0: Plantations – Treed |
| | | | SOLRIS 3.0: Cultivated |
| | | | SOLRIS 3.0: Hedge Rows |
| | | | OLCC 2.0: Treed Upland |
| | | | OLCC 2.0: Deciduous Treed |
| | | | OLCC 2.0: Mixed Treed |
| | | | OLCC 2.0: Coniferous Treed |
| Wetlands (n = 28, 581 obs) | 9 | Water table seasonally or permanently at, near, or above substrate surface (can be gently flowing). Vegetation can include woody plants, trees (often <25%), shrubs | OLCC 2.0: Plantations – Treed |
| | | | OLCC 2.0: Cultivated |
| | | | OLCC 2.0: Hedge Rows |
| | | | SOLRIS 3.0: Marsh |
| | | | SOLRIS 3.0: Treed Swamp |
| | | | SOLRIS 3.0: Thicket Swamp |
| | | | SOLRIS 3.0: Bog |
| | | | OLCC 2.0: Mudflats |
| | | | OLCC 2.0: Marsh |

| | | | |
|-----------------------------------|----|---|------------------------------------|
| | | (often hydrophytic, often >25%). | OLCC 2.0: Swamp |
| | | | OLCC 2.0: Bog |
| | | | SOLRIS 3.0: Open Tallgrass Prairie |
| | | | SOLRIS 3.0: Tallgrass Savannah |
| | | | SOLRIS 3.0: Tallgrass Woodland |
| | | | OLCC 2.0: Open Tallgrass Prairie |
| | | | OLCC 2.0: Tallgrass Savannah |
| | | | OLCC 2.0: Tallgrass Woodland |
| Grasslands (n = 70 obs) | 10 | Ground layer dominated by prairie graminoids, tree cover < 60%, shrub cover <25%. | |
| Water (n = 18, 687 obs) | 11 | Permanent water >2m deep, vegetation coverage <25%. | |
| | | | SOLRIS 3.0: Open Water |
| | | | OLCC 2.0: Clear Open Water |
| | | | OLCC 2.0: Turbid Water |

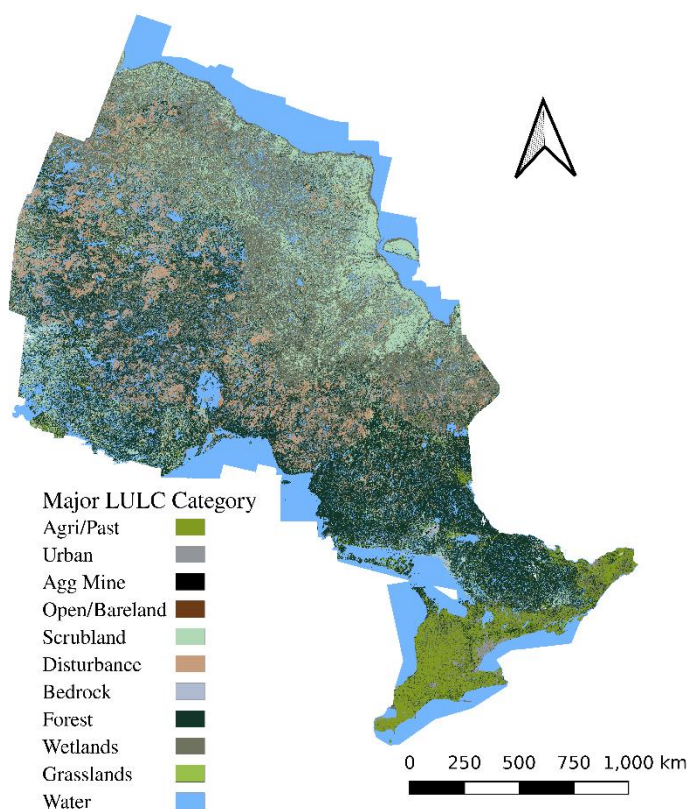


Figure 3-1: Map of Ontario based on newly defined major LULC categories (White et al., 2022).

3.3.3 *E. coli* Contamination Potential Estimation

A series of regression analyses (R package “gamlss”; (Rigby et al., 2005) were conducted to identify the relationships between the LULC categories and the presence or absence of *E. coli* in tested wells. Similar to White et al. (2021), regression analyses were selected as the best method for uncovering these relationships as they require minimal computational requirements and are highly interpretable, allowing for adherence to known physical processes. Further, the specific method of generalized additive model for location, scale, and shape (GAMLSS) was selected over other methods (i.e., linear models, generalized linear models, and general additive models) due to its ability to deal with highly skewed data distributions. This specification is important due to both the large variation in observations in different LULC categories (Table 3-1) and the large number of zero *E. coli* counts in the water sample observation data (96% of observations) (Stasinopoulos and Rigby, 2007; White et al., 2021).

The corresponding models use the independent variable, LULC category, to explain *E. coli* presence in well samples as a proxy for well contamination potential. Data subsets were created to account for the skewed data in both datasets. With respect to water sample observations, a subset was created of observations where *E. coli* presence was detected. LULC data were separated into two subsets; one with less than 30,000 observations per category and the other with at least 135,000 observations per category. A total of six regression analyses were run using combinations of the two full datasets and the three subsets. This was undertaken to further investigate how different LULCs relate to the presence of *E. coli* in wells. Models using datasets that included non-detect *E. coli*

observations use the zero adjusted logarithmic distribution (ZALG) fitting family models, while those that excluded non-detect *E. coli* observations use the logarithmic (LG) fitting family. A description of fitting families can be found in Rigby et al. (2005).

Each model was developed based on 10 iterations with 10 unique data splits (80% training, 20% testing). The regression coefficients were averaged across iterations to address parameter uncertainty (mean and variance) (White et al., 2021). The potential for *E. coli* occurrence in wells associated with each of the LULC categories within each of the six regression models was then summarized and ordered from highest *E. coli* contamination potential (high positive coefficient) to lowest *E. coli* contamination potential (high negative coefficient). Scores for each LULC category were summed across all six models, and standardized based on the number of models that a LULC category was present in. This was undertaken to identify a robust order of LULC from highest to lowest *E. coli* contamination potential.

To ensure that the large range in the number of water sample observations between LULC categories did not influence the LULC impact potential findings from the model, the findings were compared against those from a literature review. The literature review also highlighted which LULC categories lack evidence. This comparison was conducted using the method described by Mastrandrea et al. (2010). A level of confidence was assigned to each summarized model finding, based on an assessment of both the degree of evidence within the identified literature and the level of agreement among the literature. The evaluation criteria are included in Table A.3.1-1, and summarized literature used to assign a level of confidence can be found in Table A.3.1-2.

3.4 RESULTS AND DISCUSSION

Summarized regression analyses (Table A.3.1-3) were used to determine the order of *E. coli* contamination potential for wells, from highest contamination potential to lowest contamination potential. Table 3-2 shows the scores assigned based on the summary of each regression analysis.

Table 3-2: Summary of regression models exploring varying subsets of data including and excluding *E. coli* non-detect water sample results. Ordering of LULC categories (i.e., 1 -10) is based on the averaged coefficient of 10 iterations of cross validated regression analyses, where one results in the highest adverse *E. coli* contamination potential and 10 the least adverse potential.

| LULC Category | All LULC categories (with all <i>E. coli</i> data) | All LULC categories (with <i>E. coli</i> presence subset) | LULC subset with <30,000 Obs.* in category (with all <i>E. coli</i> data) | LULC subset with <30,000 Obs.* in category (with <i>E. coli</i> presence subset) | LULC subset with >135,000 Obs.* in category (with all <i>E. coli</i> data) | LULC subset with >135,000 Obs.* in category (with <i>E. coli</i> presence subset) |
|-----------------------|--|---|---|--|--|---|
| Open/Bareland | 1 | 1 | 1 | 1 | X | X |
| Aggregate Mine | 2 | 2 | 2 | 2 | X | X |
| Pastoral/Agricultural | 3 | 3 | X | X | 1 | 1 |
| Urban | 4 | 4 | X | X | 2 | 2 |
| Disturbance | 5 | 5 | 3 | 3 | X | X |
| Bedrock | 6 | 6 | 4 | 4 | X | X |
| Forest | 7 | 7 | X | X | 3 | 3 |
| Scrubland | 8 | 8 | 5 | 5 | X | X |
| Wetland | 9 | 9 | 6 | 6 | X | X |
| Water | 10 | 10 | 7 | 7 | X | X |
| Grassland | 11 | NA | 8 | NA | X | X |

• NA represents a LULC category that was to be included in the regression analysis but did not contain any water sample

observations

- X represents a LULC category that was not considered in an analysis
-

Table 3-3 presents the derived *E. coli* contamination potential classification for each LULC category along with the respective level of confidence from the literature (Mastrandrea et al., 2010). Notably, LULCs that result in the greatest prevalence of *E. coli* in groundwater also have the greatest consensus in literature. The pastoral/agricultural category (n = 276,308) was classed as very high as a result of robust consensus in the literature, strong statistical association with *E. coli* presence in wells, and a relatively small standard deviation within the regression analyses. This is likely due to the high *E. coli* loadings from animal presence and spreading of manure as fertilizer (Gregory et al., 2019; Jabbar et al., 2019), and the potential of disturbed soil structures to create preferential flow pathways due to soil tilling (Hua, 2017). Urban LULCs are associated with a high prevalence of *E. coli* in groundwater due to the higher occurrence of domestic animals, leaking septic tanks, and presence of wastewater treatment plants (Hua, 2017; Jabbar et al., 2019; Paule-Mercado et al., 2016). With respect to low *E. coli* contamination potential for wells, wetlands (n = 28,581) were identified in this class through both the regression analyses and the literature. This is likely due to the natural purification abilities of wetlands, as they are made up of soils with higher sorption capacities and support vegetation with roots that reduce bacteria (e.g., via antimicrobial excretions) (Dordio et al., 2008). Aggregate mines were found to have a consistent statistical association with *E. coli* presence in wells, though they had relatively high standard deviations within regression analyses. Despite high standard deviations, aggregate mines are concluded to result in high

adverse impact potential due to a correlation between heavy metals and *E. coli* presence as well as the prevalence of preferential flow paths (Armah, 2014; Somaratne and Hallas, 2015). Conversely, while regression analyses identified open/bareland as having the highest adverse *E. coli* contamination potential in groundwater, the low number of water sample observations available for the current analysis ($n = 153$) combined with the low level of confidence derived from the literature, resulted in this LULC category being classified as high, rather than very high (Table 3-3). More generally, an examination of the level of confidence shows that LULC categories with fewer than 25,000 water sample observations have lower level of confidence due to a lack of literature, likely due to similar data limitations as this study. An interesting exception is grasslands; although there were few water samples in this LULC category ($n = 70$) in the current dataset, there was widespread consensus in the literature that grasslands have a low *E. coli* contamination potential for groundwater, which corresponded well to the regression analyses.

Table 3-3: Summary of LULC *E. coli* contamination potential and respective level of confidence with respect to literature. Where ★★★★★ represents high confidence, and ★ represents low confidence.

| Classification of <i>E. coli</i> contamination potential | Land Use - Land Cover | Level of confidence based on Literature | Ranking based on regression ¹ | Revised level of confidence based on regression analysis plus literature |
|--|---|---|--|--|
| Very High | Pastoral/Agricultural ($n = 276, 308$ obs) | ★★★★★ | 2 (****) | ★★★★★ |
| High | Open/Bareland ($n = 153$ obs) | ★ | 1 (****) | ★★★ |
| | Aggregate Mines ($n = 308$ obs) | ★ | 2 () | ★★★★ |

| | | | | |
|---|---------------------------------|------|----------|-------|
| Moderate | Urban (n = 221, 890 obs) | ★★★★ | 3 (****) | ★★★★★ |
| | Disturbance (n = 1, 658 obs) | ★ | 4 () | ★ |
| | Bedrock (n = 3,204 obs) | | 5 () | ★ |
| | Forest (n = 136, 151 obs) | ★★★★ | 6 (**) | ★★★★★ |
| | Scrubland (n = 21, 448 obs) | ★ | 7 (****) | ★ |
| Low | Wetlands (n = 28, 581 obs) | ★★ | 8 (****) | ★★★ |
| | Water (n = 18, 687 obs) | ★ | 9 (****) | ★★★★ |
| | Grasslands (n = 70 obs) | ★★★ | 10 () | ★★★★★ |
| ¹ Statistical significance of variable from regression analyses based on LULC subsets with <30, 000 and >135,000 observations. **** 99.9%, *** 99%, **95%, *90%. | | | | |

LULC categories with zero to two stars require more exploration to improve the literature-based level of confidence of the *E. coli* contamination potential for wells. It is important to note that five out of seven (71%) LULC categories are assessed as low confidence in the literature due to limited evidence rather than low agreement across articles. This work successfully added statistically significant findings to literature on the LULC categories of agriculture/pastoral, open/barelands, urban, forest, scrubland, wetlands, and water. While not statistically significant, this work also added insights to the other categories of aggregate mines, disturbance, bedrock, and grasslands. Further, this work emphasises specific LULC categories where more research is needed to increase the overall level of confidence (a goal of ★★★★★) in the *E. coli* contamination potential

of all LULC categories for wells. Based on the findings summarized in Table 3.2, a spatial dataset for *E. coli* well contamination potential across Ontario (resolution 15m x 15m) was created (White et al., 2022, Figure 3-2). Future research should seek to build on this study in an effort to continue to fill highlighted LULC contamination potential knowledge gaps.

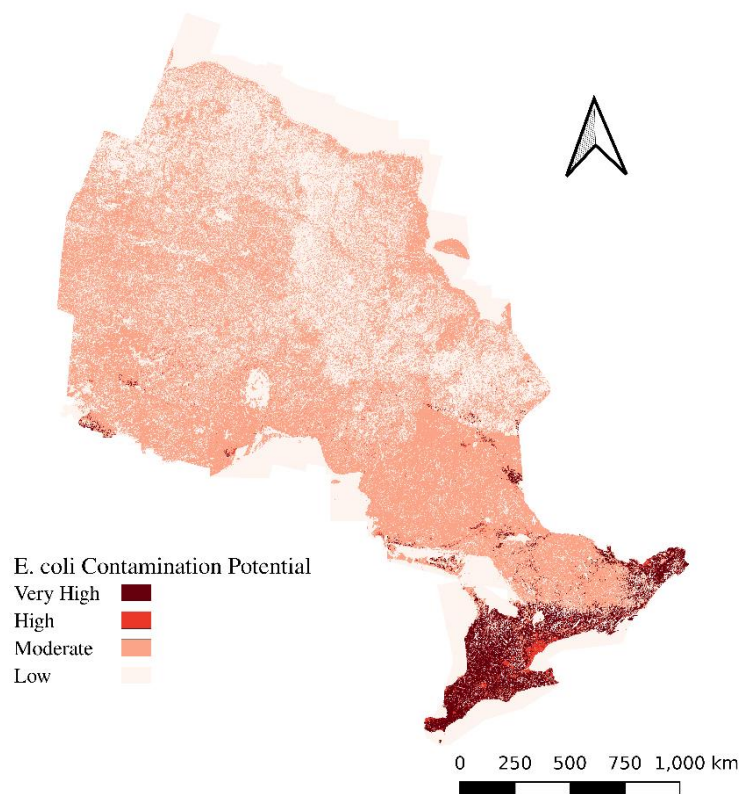


Figure 3-2: Map of *E. coli* contamination potential classes for Ontario based on derived raster data (White et al., 2022).

Based on Figure 3-2, Southern Ontario demonstrates a higher likelihood of *E. coli* contamination potential for wells. This is explained by the greater presence of at least two high risk LULC categories, namely pastoral/agriculture and urban. This coupling of “very high” (pastoral/agricultural) and “high” (urban, open/barelands, aggregate mines)

contamination potential classes can be seen in other heavily urbanized regions outside of Southern Ontario such as Sudbury, Englehart, Thunder Bay, and Atikokan. This consistent coupling results in contamination potential “hotspots” across Ontario near large urban centers. The identification of high and very high *E. coli* contamination potential areas could inform public health-based campaigns educating well users of the importance of well water testing, treatment, and maintenance. These spatially targeted campaigns could improve stewardship of some of the most vulnerable private wells.

3.5 POLICY IMPLICATIONS

The LULC-driven *E. coli* contamination potential map presented in this work offers unique policy and regulatory implications for well characteristics and locations, public health protection, land use and management strategies, and well source water protection. While deeper wells are often preferred as there is reduced opportunity for the well water to be under the direct influence of surface water, in the high and very high *E. coli* contamination potential classes it should be a requirement that new construction wells are deeper than the current regulations require (20ft, with the exception of where only a shallow aquifer exists reducing requirement to 10ft; O.Reg. 903: Wells). As the *E. coli* loadings are higher in these areas, this will help to protect well users from *E. coli* contamination events.

For wells located in high and very high classes it is advisable for them to be treated as stringently as municipal wells. Currently, municipal wells in Ontario are treated as vulnerable areas under the *Clean Water Act, 2006* and are supported by wellhead protection areas (WHPA). The purpose of WHPAs is to require planning policies to limit certain land

use activities (a list of 21 have been identified in Section 107 of the *Clean Water Act*) within specific zones surrounding a well where activities may impact groundwater quality in the well (i.e., Zone A – 100 m from well, Zone B – 2-year travel time from well, Zone C – 5-year travel time, Zone D - 25-year travel time). Current well drilling regulations for private wells (O.Reg 903: Wells) require a minimum distance of only 30m from a source of contamination, and sources of contamination are not as extensive as the *Clean Water Act*. For new wells located in any *E. coli* contamination potential class, increasing the required 30m from potential contaminant sources to the minimum WHPA Zone-A's 100m would decrease the vulnerability of private wells. Further, for new private wells located in high or very high classes, WHPA's Zone-B should also be considered, ensuring that the potential risks of the 21 land use activities are clearly outlined and provided to the well user. As private wells have been identified as a potential source of contamination in WHPAs (Ausable Bayfield Source Protection Authority, 2015), this policy-driven improved protection of private wells and knowledge transfer to private well users will result in decreased vulnerability for private well users, municipal wells, and highly vulnerable aquifers.

3.6 CONCLUSIONS

A new method for assessing the impact of LULC on *E. coli* contamination potential for wells was applied to a case-study of Ontario, Canada. Utilising an extensive dataset of microbiological sampling in private wells, this derived raster data represents one of the first to provide geospatial LULC *E. coli* contamination potential of groundwater. By undertaking a robust regression analysis that included multiple iterations, data splits, and

subset analyses to account for skewed data distributions, this dataset represents a new way to assess *E. coli* contamination potential for wells in Ontario. Comparing regression analyses to existing literature has further contributed to and strengthened knowledge of potential contamination levels of different LULC. Specifically, consensus was found between regression analyses and literature for the following:

- pastoral/agriculture LULC type (categorised as very high in terms of *E. coli* contamination potential for wells in Ontario);
- urban LULC type (categorised as high in terms of *E. coli* contamination potential for wells in Ontario);
- aggregate mines LULC type (categorised as high in terms of *E. coli* contamination potential for wells in Ontario);
- forest LULC type (categorised as moderate in terms of *E. coli* contamination potential for wells in Ontario); and
- water and grasslands LULC types (categorised as low in terms of *E. coli* contamination potential for wells in Ontario).

Literature-based level of confidence was increased based on the findings in this work for the following:

- open/barelands increased confidence from 1 to 2 stars;
- aggregate mines increased in confidence from 1 to 3 stars;
- bedrock increased in confidence from 0 to 1 star;
- water increased in confidence from 1 to 3 stars; and
- grasslands increased in confidence from 3 to 4 stars.

While progress was made, more research is still required for specific LULC categories such as scrublands, disturbance, bedrock, water and open/barelands to continue to increase

level of confidence. Application of this approach to other large datasets and regions can further improve our understanding of the role of LULC in *E. coli* contamination potential for wells. Further, the raster datasets presented in this work can inform policies and regulations to reduce and mitigate the impacts of microbiological contamination of wells (private and municipal) to protect the health of people relying on drinking water wells.

3.7 WORKS CITED

- Armah, F.A., 2014. Relationship Between Coliform Bacteria and Water Chemistry in Groundwater Within Gold Mining Environments in Ghana. *Water Quality, Exposure and Health* 2014 5:4 5, 183–195. <https://doi.org/10.1007/S12403-014-0110-1>
- Arnaud, E., Best, A., Parker, B.L., Aravena, R., Dunfield, K., 2015. Transport of *Escherichia coli* through a Thick Vadose Zone. *J Environ Qual* 44, 1424–1434. <https://doi.org/10.2134/jeq2015.02.0067>
- Ausable Bayfield Source Protection Authority, 2015. Protecting Our Drinking Water Together: An Introduction to Drinking Water Source Protection.
- Caldeira, K., Chopin, T., Gaines, S., Haugan, P., Hemer, M., Howard, J., Konar, M., Krause-Jensen, D., Lindstad, E., Lovelock, C.E., Michelin, M., Nielsen, F.G., Northrop, E., Parker, R., Roy, J., Smith, T., Some, S., Tyedmers, P., 2019. The Ocean as a Solution to Climate Change: Five Opportunities for Action. Washington, D.C.
- Dordio, A., Carvalho, A.J.P., Pinto, A.P., 2008. Wetlands: Water “Living Filters”?, in: *Wetlands: Ecology, Conservation and Restoration*. pp. 15–71.

Dusek, N., Hewitt, A.J., Schmidt, K.N., Bergholz, P.W., 2018. Landscape-Scale Factors Affecting the Prevalence of *Escherichia coli* in Surface Soil Include Land Cover Type, Edge Interactions, and Soil pH.

Elliott, A.H., Semadeni-Davies, A.F., Shankar, U., Zeldis, J.R., Wheeler, D.M., Plew, D.R., Rys, G.J., Harris, S.R., 2016. A national-scale GIS-based system for modelling impacts of land use on water quality. *Environmental Modelling and Software* 86, 131–144. <https://doi.org/10.1016/j.envsoft.2016.09.011>

Gregory, L.F., Harmel, R.D., Karthikeyan, R., Wagner, K.L., Gentry, T.J., Aitkenhead-Peterson, J.A., 2019. Elucidating the Effects of Land Cover and Usage on Background *Escherichia coli* Sources in Edge-of-Field Runoff. <https://doi.org/10.2134/jeq2019.02.0051>

Hua, A.K., 2017. Land Use Land Cover Changes in Detection of Water Quality: A Study Based on Remote Sensing and Multivariate Statistics. *J Environ Public Health* 2017. <https://doi.org/10.1155/2017/7515130>

Jabbar, F.K., Grote, K., Tucker, R.E., 2019. A novel approach for assessing watershed susceptibility using weighted overlay and analytical hierarchy process (AHP) methodology: a case study in Eagle Creek Watershed, USA. *Environmental Science and Pollution Research* 26, 31981–31997. <https://doi.org/10.1007/s11356-019-06355-9>

Larned, S.T., Scarsbrook, M.R., Snelder, T.H., Norton, N.J., Biggs, B.J.F., 2004. Water quality in low-elevation streams and rivers of New Zealand: Recent state and trends

in contrasting land-cover classes. *N Z J Mar Freshwater Res* 38, 347–366.

<https://doi.org/10.1080/00288330.2004.9517243>

Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R., Plattner, G.-K., Yohe, G.W., Zwiers, F.W., 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Intergovernmental Panel on Climate Change (IPCC).

Namugize, J.N., Jewitt, G., Graham, M., 2018. Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. *Physics and Chemistry of the Earth* 105, 247–264. <https://doi.org/10.1016/j.pce.2018.03.013>

Ontario Ministry of Natural Resources and Forestry, 2019. Southern Ontario Land Resources Information System (SOLRIS) 3.0 [WWW Document]. URL <https://geohub.lio.gov.on.ca/documents/lio::southern-ontario-land-resource-information-system-solris-3-0/about>

Ontario Ministry of Natural Resources and Forestry, 2016. Ontario Land Cover Compilation v.2.0 [WWW Document]. URL <https://geohub.lio.gov.on.ca/documents/7aa998fdf100434da27a41f1c637382c/about>

Paule-Mercado, M.A., Ventura, J.S., Memon, S.A., Jahng, D., Kang, J.H., Lee, C.H., 2016. Monitoring and predicting the fecal indicator bacteria concentrations from agricultural, mixed land use and urban stormwater runoff. *Science of the Total Environment* 550, 1171–1181. <https://doi.org/10.1016/j.scitotenv.2016.01.026>

- Petersen, F., Hubbart, J.A., 2020. Quantifying escherichia coli and suspended particulate matter concentrations in a mixed-land use appalachian watershed. *Water (Switzerland)* 12, 532. <https://doi.org/10.3390/w12020532>
- Rigby, R.A., Stasinopoulos, D.M., Lane, P.W., 2005. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C Appl Stat* 54, 507–554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- Rivera, A., 2014. Canada's Groundwater Resources. Fitzhenry & Whiteside, Markham.
- Somaratne, N., Hallas, G., 2015. Review of risk status of groundwater supply wells by tracing the source of coliform contamination. *Water (Switzerland)* 7, 3878–3905. <https://doi.org/10.3390/w7073878>
- Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw* 23, 1–46. <https://doi.org/10.18637/jss.v023.i07>
- Tate, K.W., Atwill, E.R., Bartolome, J.W., Nader, G., 2006. Significant *Escherichia coli* Attenuation by Vegetative Buffers on Annual Grasslands. *J Environ Qual* 35, 795–805. <https://doi.org/10.2134/jeq2005.0141>
- White, K., Dickson-Anderson, S., Majury, A., McDermott, K., Hynds, P., Brown, R.S., Schuster-Wallace, C., 2021. Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable,

geo-spatio-temporal dataset. Water Res 197, 117089.

<https://doi.org/10.1016/j.watres.2021.117089>

White, K., Dickson-Anderson, S., Schuster-Wallace, C., 2022. Land Use Land Cover in Ontario and Respective Impact to *E. coli* Contamination in Private Wells.

<https://doi.org/10.20383/102.0530>

CHAPTER 4 EXPLORING THE ROLE OF RAINFALL

INTERMITTENCY ON *E. COLI* CONTAMINATION

EVENTS IN PRIVATE WELLS: AN ONTARIO CASE

STUDY

Summary of Paper 3: Exploring the role of rainfall intermittency on *E. coli* contamination events in private wells: an Ontario case study, by K. White, A. Yosri, S. Dickson-Anderson, and C. Schuster-Wallace, Will be submitted to Journal of Weather and Climate Extremes by August 2023.

Summary:

This research expands on the first chapter of this thesis, exploring how rainfall intermittency patterning impacts *E. coli* presence in private wells. This work utilizes a mixture of supervised (spectral clustering) and unsupervised machine learning (random forest-ensembled decision tree-based classification, GAMLSS, MARS, LASSO) approaches to move beyond using standard rainfall lag time towards rainfall patterning. The results of this demonstrated:

- While rainfall intermittency could not predict *E. coli* presence on its own, the optimal antecedent time period to consider in predicting *E. coli* presence in wells was 36 days, though a more computationally conservative lag time of 16 days is also appropriate.
- If there was no rainfall in the 36 days prior to the date of observations, it is highly unlikely that *E. coli* will be found in a private well, due to the lack of transport mechanism.
- Regression techniques identified the increasing the number of consecutive wet days or number of wet day/dry day cycles increases the likelihood of *E. coli* presence

and decreases the severity of contamination events. This is tied to the concept of dilution theory in groundwater.

4.1 ABSTRACT

Rainfall events alter likelihood and severity of groundwater contamination. While likelihood is typically captured through standardized lag times, these are not ideal. Thus, the goal of this work is to utilize a large Ontario-specific dataset to assess whether it is possible to move beyond standard lag times. Machine learning techniques were applied to various rainfall characteristics and a large *E. coli* presence/absence dataset to explore the existence of unique patterns that improve upon the information provided by lag times. The presence and severity of *E. coli* in private wells cannot be predicted using rainfall intermittency patterns alone. However, it was determined that optimal lag times are less than 36-days; thus, if no rainfall occurs 36 days prior, wells are significantly less likely to become contaminated. Increasing the number of consecutive wet days or the number of wet/dry cycles increases the likelihood of lower-level *E. coli* concentrations occurring in wells.

Keywords:

Machine learning, private wells, *E. coli*, rainfall intermittency, rainfall lag times, large dataset

4.2 INTRODUCTION

It has been well documented that rainfall events increase the likelihood of drinking water source contamination, which is associated with the occurrence of diarrhea in consumers (Carlton et al., 2014; De Roos et al., 2020; Gleason and Fagliano, 2017; Jagai et al., 2015; Kraay et al., 2020; Levy et al., 2016). More specifically, rainfall after a dry

period has been linked to an increase in diarrhea occurrence, while rainfall after a wet period was found to be protective against diarrhea (Levy et al., 2016). The higher likelihood post dry period has been associated with the “concentration effect”, which is a flushing of pathogens that have accumulated on the land (Kraay et al., 2020), remobilizing them through the surface and groundwater systems (Levy et al., 2016). Conversely, the lower likelihood caused by rainfall after a wet period is attributed to the “dilution effect”, in which pathogens on the surface and in the groundwater have been mobilised and are diluted by recent, prior rainfalls (Levy et al., 2016). These relationships demonstrate that intermittency, a core characteristic of rainfall patterns (Trenberth et al., 2017), plays an important role in introducing waterborne pathogens to drinking water sources, including wells.

Lag times represent important variables for environmental disease modeling, particularly waterborne diseases, as they are a proxy for pathogen transport through terrestrial and aquatic environments (including groundwater), latency period (time from exposure to symptoms), and in the case of hospitalizations, time to seek care (though not relevant in this work). Current approaches used to examine relationships between weather conditions and adverse water quality or waterborne disease incidence typically identify and apply lag times. These lags can be identified through data fitting or a priori methods (Table 4-1). Studies adopting such approaches show a great variability with respect to the relationships being explored, water sources, and variables utilized, and therefore in the lag times identified. The variability in approaches introduces additional uncertainty into

predictive models, particularly when attempting to assess potential impacts of climate change on drinking water contamination and, by extension, waterborne disease burdens.

Table 4-1: Summary of literature exploring the relationship between rainfall and its respective outcome (i.e., water quality or human health).

| Lag Time | Relationship Explored | Water Source | Measurement of Study | How Lag Time Determined | Author |
|-----------------|--|----------------------------------|-----------------------------|--------------------------------|------------------------------|
| 1 day | Microbiological contamination in wells after rainfall event | Groundwater Well | Water Quality | Data Fitting | (Godfrey et al., 2005) |
| 2 days | Linkage between ER visits and extreme precipitation event | Surface Water | Human Health | Data Fitting | (Gleason and Fagliano, 2017) |
| | Linkage between heavy rainfall and first time pediatric hospitalizations for enteric bacterial or viral infections | Locally Supplied Drinking Water | Human Health | Data Fitting | (Lai et al., 2020) |
| 3 days | Human risk of exposure to faecally contaminated water from river source | River | Water Quality | Data Fitting | (Buckerfield et al., 2019) |
| | Linkage between rainfall events and well water quality | Groundwater Wells | Water Quality | A Priori | (Wu et al., 2016) |
| 4 days | Linkage between rainfall events and pediatric hospital visits for AGI | Municipal Drinking Water Systems | Human Health | A Priori | (Drayna et al., 2010) |

| | | | | | |
|------------------|---|--|--|--------------|----------------------------|
| 8 days | Linkage between ER visits and extreme precipitation event | Undefined Drinking Water Sources (exposed to CSOs) | Human Health (ER Visit with GI Illness) | A Priori | (Jagai et al., 2015) |
| 8-16 days | Linkage between heavy precipitation and hospital visit | Surface Water (River) | Human Health (AGI Cases in Hospital) | Data Fitting | (De Roos et al., 2020) |
| 2 weeks | Linkage between heavy rainfall and diarrhea incidence in humans | Mainly surface water (river), rarely rainwater and groundwater | Human Health (Diarrhea Incidence) | A Priori | (Carlton et al., 2014) |
| | <i>E. coli</i> removal efficiency after dry and wet periods in bench-scale column | Laboratory | Water Quality | A Priori | (Chandrasena et al., 2019) |
| 4 weeks | Linkage between heavy rainfall events and waterborne disease outbreaks due to contaminated surface water | Surface Waters | Human Health (Disease Outbreaks (2 or more individuals falls ill)) | Data Fitting | (Curriero et al., 2001) |
| | Linkage between heath outbreaks and drinking water (Note: this does not include <i>E. coli</i> specific outbreak) | Groundwater and Surface Water | Human Health (Disease Outbreaks) | Data Fitting | (Nichols et al., 2009) |
| 6 weeks | Linkage between waterborne disease outbreaks and high-impact weather events | Drinking Water Source (specific source unknown) | Human Health (Disease Outbreaks) | A Priori | (Thomas et al., 2006) |

| | | | | | |
|---------------------|--|-------------|---|--------------|-------------------------|
| 2 months | Linkage between heavy rainfall events and waterborne disease outbreaks due to contaminated groundwater | Groundwater | Human Health (Disease Outbreaks (2 or more individuals falls ill) | Data Fitting | (Curriero et al., 2001) |
|---------------------|--|-------------|---|--------------|-------------------------|

While rainfall intermittency is not the only factor contributing to the likelihood of wells becoming contaminated with *E. coli* (e.g., White et al., 2021, White et al., 2022), rainfall intermittency patterns may provide an improved approach over the current use of fixed lag times, as use of fixed lag times are deemed impossible by Guzman Herrador et al. (2015). As such, the goal of this work is to assess if it is possible to move beyond the various current rainfall lag times. This is achieved by: i) applying machine learning techniques (i.e., cluster, classification, regression, and association rule analyses) to identify and describe unique, meaningful clusters of patterns based on relevant characteristics (i.e., maximum consecutive wet and dry days, number of wet and dry days, total rainfall, mean rainfall, standard deviation of rainfall, maximum single day rainfall, number of wet/dry cycles); ii) identifying all unique intermittency patterns associated with *E. coli* presence in private wells; and, iii) exploring whether land use-land cover (LULC) *E. coli* Contamination Potential classes and bottom of well stratigraphy can improve predictive accuracy of analyses. This is undertaken to explore potential improvements in analyses that can be used as support tools for well owners and decision-makers for improved well stewardship.

4.3 METHODS

4.3.1 Datasets

The analyses in this paper have been undertaken utilizing Ontario-specific private well data, Daymet rainfall data, and LULC *E. coli* Contamination Potential class data (Table 4-2).

Table 4-2: Summary of datasets and variables used in this study.

| Dataset | Variables | Specifications | Reference |
|---|--------------------------------------|---|--|
| Private Well Dataset | <i>E. coli</i> observations | Data available from 2010 to 2017 | (Latchmore et al., 2020; White et al., 2021) |
| | Geological Formation | | |
| | Date of Observation | Point observations | |
| | Date of Observations | | |
| | Latitude and Longitude | | |
| Rainfall Data | Precipitation | Data available from 1980 to present day | (Thornton et al., 2021) |
| | Air Temperature | | |
| | Vapour Pressure | Grid size 1km x 1km | |
| LULC <i>E. coli</i> Contamination Potential Class Data | Land Use - Land Cover <i>E. coli</i> | Data available from 2000 to 2015 | (White et al., 2022) |
| | Contamination Potential Classes | | |
| | | Grid size 15m x 15m | |

4.3.2 Data Processing

4.3.2.1 Private Well Dataset

The private well dataset is the amalgamation of the Well Water Testing Database (WWTD), containing information on microbiological testing completed on private wells in Ontario, and the Well Water Information System (WWIS), containing information on

well characteristics. For more information on these datasets, see Latchmore et al. (2020) and White et al. (2021). The categorical and binary sub-classes of *E. coli* observations created for analyses in White et al. (2021) were utilized for these analyses. Categorical sub-classes are defined as non-detects (ND) (0 CFU/100mL), Category 1 (1-10 CFU/100mL), Category 2 (11-50 CFU/100mL), and Category 3 (51+ CFU/100mL). Binary sub-classes are defined as non-adverse (0 CFU/100mL) and adverse (1+ CFU/100mL). Binary sub-classes of geological formation were also used, i.e., consolidated (metamorphic, sedimentary, igneous rock) and unconsolidated materials (sand, gravel, till) (White et al., 2021).

4.3.2.2 Rainfall Data

Data available through Daymet (Thornton et al., 2021) includes precipitation, which is defined as daily total precipitation in all forms in millimeters water-equivalent. As this study is only interested in rainfall, the fraction falling as liquid was separated from the total precipitation using the rain-snow identifier equation (Jennings et al., 2018) based on average daily air temperature (T_{avg}), relative humidity (RH), and the likelihood of snow (PS) as follows:

$$T_{avg_{i,n}} = \frac{T_{min_{i,n}} + T_{max_{i,n}}}{2} \quad [\text{Eq. 4.1}]$$

where the subscripts *i* and *n* represent the date and Daymet Tile ID, respectively, T_{min} is the minimum temperature, T_{max} is the maximum temperature, and T_{avg} is the average temperature.

Relative humidity is then calculated according to (Murray, 1966):

$$RH_{i,n} = \frac{VP_{i,n}}{6.11 * 10^{7.5T_{avg_{i,n}}/237.7 + T_{avg_{i,n}}}} * 100\% \quad [Eq. 4.2]$$

where VP is the unsaturated vapour pressure, and RH is the relative humidity.

Finally, the likelihood of snow is calculated utilizing the bivariate equation from Jennings et al. (2018) and used to calculate rainfall.

$$PS_{i,n} = \frac{1}{1 + e^{-10.04 + 1.41T_{avg_{i,n}} + 0.09RH_{i,n}}} \quad [Eq. 4.3]$$

$$Rainfall_{i,n} = \begin{cases} PS_{i,n}, & PS_{i,n} < 0.5 \\ 0, & PS_{i,n} \geq 0.5 \end{cases} \quad [Eq. 4.4]$$

where PS is the likelihood of snow, P is the precipitation, and Rainfall is the rainfall.

Extracted rainfall data employed in the analyses includes raw continuous rainfall data, binary rainfall categories (dry: ≤ 0.1 mm/day; wet: > 0.1 mm/day), and rainfall pattern-based variables (see Section 2.3.1).

4.3.2.3 Linking Unique Datasets

The Spatial Analyst toolbox in ArcGIS was used to link rainfall and LULC data to the private well dataset, using the private well dataset as the limiting dataset (i.e., only dataset with a discrete number of data points). For rainfall data, unique private well locations ($n = 253,136$) were assigned to 1km-by-1km grid cells corresponding to the original rainfall data (i.e., Daymet grid). Once each unique well was assigned a grid cell, for each water sample observation ($n = 795,023$), rainfall values and corresponding categories (i.e., dry versus wet) for the 60 days prior to the well sample date were recorded. Sixty days is used as a starting point of analysis as it corresponds to the longest lag time

found in literature (Curriero et al., 2001). For LULC data, unique private well locations were assigned to 15m-by-15m grid cells corresponding to the LULC *E. coli* Contamination Potential class data (White et al., 2022).

4.3.3 Data Analytics Techniques

4.3.3.1 Cluster and Classification Techniques

Cluster and classification analyses are machine learning techniques that are typically applied to identify groups in observations (clustering) and the relationship between inputs and categorical outputs (classification). Cluster analysis is applied in an unsupervised manner with the objective of categorizing similar observations based on a distance measure (Xu and Wunsch, 2008). K-means, model-based, and spectral clustering are the typical approaches used for such purposes, where *i*) in K-means clustering, a number of clusters K is assumed and the resulting within-cluster sum of squared distance is subsequently evaluated (Haggag et al., 2021); *ii*) in model-based clustering, a finite mixture model is used to discretize observations into unimodal components representing the clusters (Fraley and Raftery, 2002); and *iii*) in spectral clustering, observations are conceptualized as nodes of a directed graph and commonalities are subsequently identified based on eigenanalysis (Lin et al., 2021). It should be highlighted that while K-means and model-based clustering are the most common clustering approaches, their application is considered an iterative process since the number of clusters is changed continually until an optimal value is achieved. In contrast, spectral clustering represents a more computationally efficient alternative since the optimal number of clusters is directly identified based on the number of nearly zero eigenvalues. As such, in this study spectral clustering analysis is applied to

group similar rainfall patterns at different well locations and subsequently identify the characteristics of those associated with *E. coli* contamination.

Classification analysis, in contrast to clustering, aims to explore the relationship between continuous/categorical inputs and a set of categorical outputs (Goforth et al., 2022; Zounemat-Kermani et al., 2021). Decision tree, artificial neural network, and support vector machine are among the most efficient classification approaches that can be applied separately, individually within a resampling ensemble method (i.e., random forest, boosting, bagging), or collectively within an average model (Zounemat-Kermani et al., 2021). Model evaluation metrics (i.e., recall, precision, f-score, and accuracy) are typically utilized to assess the model capability to reflect the input-output relationship.

In this study, a random forest-ensembled decision tree-based classification (referred to as the classification model hereafter) is used to determine the antecedent time period that is most important for predicting *E. coli* presence in private wells, as well as to investigate the coupling of rainfall, LULC *E. coli* Contamination Potential classes, and stratigraphic data that better represent mobilization and transport processes than rainfall alone. The application of decision tree-based classification relies on hierarchical treelike decisions according to the input values (Zumel and Mount, 2019). When applied within a random forest ensemble method, bootstrapped decision trees are developed independently and are subsequently integrated within an average classifier. It should be emphasized that other ensemble techniques can be employed; however, the random forest approach is selected in this study as its efficacy has been confirmed in related applications (Bindal and Singh, 2019; Chen et al., 2020; He et al., 2022; Naghibi et al., 2020).

Classification model findings of the defined antecedent time period were utilized to derive new rainfall pattern-based variables to be used in regression analyses (Table 4-3).

Table 4-3: Summary of all regression analysis variables, based on classification model finding of n-36 days representing the defined antecedent time period for *E. coli* presence prediction.

| Variable Name | Variable Description | Data Range | Data Categories* |
|---|--|-------------------|--|
| <i>E. coli</i> | Number of <i>E. coli</i> reported in sample by laboratory | 0 – 81 CFU/100mL | Non-detect (non-adverse): 0 Adverse: 1+ Category 1: 1 - 10 Category 2: 11 – 50 Category 3: 51+ |
| LULC | LULC <i>E. coli</i> Contamination Potential classes | 1 – 4 | Low: 1 Moderate: 2 High: 3 Very High: 4 |
| Stratigraphy | Stratigraphy of geologic formation in which well is situated | 1 – 2 | Consolidated: 1 Unconsolidated: 2 |
| Maximum Number of Consecutive Wet Days | Within the defined antecedent time period maximum number of consecutive wet days | 0 – 18 days | All Data Category 1: 0 - 1 Category 2: 2 Category 3: 3 - 5 Category 4: 6 - 18 Adverse Data Category 1: 1 - 2 Category 2: 3 Category 3: 4 Category 4: 5 – 12 |
| Maximum Number of Consecutive Dry Days | Within the defined antecedent time period, | 2 – 37 days | All Data Category 1: 2 - 5 Category 2: 6 - 8 |

| | | | |
|---------------------------|---|--------------|---|
| | maximum number of consecutive dry days | | Category 3: 9 - 12 Category 4: 13 - 37 Adverse Data Category 1: 2 - 5 Category 2: 6 - 7 Category 3: 8 - 10 Category 4: 11 - 36 |
| Number of Wet Days | Count of wet days over the defined antecedent time period | 0 – 31 days | All Data Category 1: 0 - 7 Category 2: 8 - 11 Category 3: 12 - 14 Category 4: 15 - 31 Adverse Data Category 1: 1 - 9 Category 2: 10 - 12 Category 3: 13 - 15 Category 4: 16 – 27 |
| Number of Dry Days | Count of dry days over the defined antecedent time period | 6 – 37 days | All Data Category 1: 6 - 21 Category 2: 22 - 24 Category 3: 25 - 28 Category 4: 29 - 37 Adverse Data Category 1: 10 - 20 Category 2: 21 - 23 Category 3: 24 - 26 Category 4: 27 - 36 |
| Total Rainfall | Sum of daily rainfall over defined antecedent time period | 0 – 320.6 mm | All Data Category 1: 0 – 50.3 Category 2: 50.4 – 88.3 Category 3: 88.4 – 123.9 |

| | | | |
|---------------------------------------|--|---------------|--|
| | | | Category 4: 124 – 320.6 Adverse Data Category 1: 0.1 – 74.4 Category 2: 74.5 – 105.9 Category 3: 106 – 138.9 Category 4: 139 – 297 |
| Mean of Rainfall | Mean of daily rainfall amounts over defined antecedent time period (i.e., total rainfall/number of wet days) | 0 – 51.32 mm | All Data Category 1: 0 – 5.64 Category 2: 5.65 – 7.37 Category 3: 7.38 – 9.31 Category 4: 9.32 – 51.32 Adverse Data Category 1: 0.1 – 6.55 Category 2: 6.56 – 8.21 Category 3: 8.22 – 10.09 Category 4: 10.1 – 45.6 |
| Standard Deviation of Rainfall | Standard deviation of daily rainfall amounts over defined antecedent time period | 0 – 30.31 mm | All Data Category 1: 0 – 3.93 Category 2: 3.94 – 5.75 Category 3: 5.76 – 7.79 Category 4: 7.80 – 30.31 Adverse Data Category 1: 0 – 4.90 Category 2: 4.91 – 6.55 Category 3: 6.56 – 8.52 Category 4: 8.53 – 28.1 |
| Maximum Single Day Rainfall | Maximum value of a single day rainfall amount over defined antecedent time period | 0 – 112.76 mm | All Data Category 1: 0 – 13.89 Category 2: 13.9 – 20.59 Category 3: 20.6 – 28.09 Category 4: 28.1 – 112.76 |

| | | | |
|---|--|--|-------------------------|
| | | | Adverse Data |
| | | | Category 1: 0.1 – 17.4 |
| | | | Category 2: 17.5 – 23.7 |
| | | | Category 3: 23.8 – 31.0 |
| | | | Category 4: 31.1 – 104 |
| | | | All Data |
| | | | Category 1: 0 - 7 |
| | | | Category 2: 8 - 10 |
| | | | Category 3: 11 – 12 |
| | | | Category 4: 13 - 24 |
| | | | Adverse Data |
| | | | Category 1: 1 - 9 |
| | | | Category 2: 10 - 11 |
| | | | Category 3: 12 - 13 |
| | | | Category 4: 14 - 24 |
| <p>* With the exception of <i>E. coli</i>, LULC, and stratigraphy, data categories were created to ensure an equal frequency in each category, there are two ranges based on the consideration of All <i>E. coli</i> data and only adverse <i>E. coli</i> data</p> | | | |

4.3.3.2 Regression-based Techniques

Regression-based techniques are supervised machine learning approaches used for the estimation of relationships between a dependent variable and one or more independent variables. The techniques used in this study include Multivariate Adaptive Regression Splines (MARS), Generalized Additive Model for Location, Scale, and Shape (GAMLSS), LASSO (Least Absolute Shrinkage and Selection Operator) Regression, and Partial Least Squares (PLS; adverse *E. coli* data only). MARS, first introduced by Friedman (1991), is a procedure for fitting adaptive non-linear regression that uses piecewise basis functions to define relationships between a dependent variable and a set of independent variables, allowing for higher degree interactions between independent variables. This technique utilizes hinge functions whereby at specific threshold(s) (knot), the linear relationship between the dependant and independent variable changes, represented by an ensemble of linear functions joined by one or more hinges. MARS has been successfully used in similar studies such as groundwater potential mapping (Park et al., 2017) and creating groundwater spring potential maps (Yousefi et al., 2020). GAMLSS is a technique for fitting regression type models where the distribution of the response variable does not have to be exponential and has the ability to deal with highly skewed and kurtotic distributions (Stasinopoulos and Rigby, 2007); this technique has been successfully used for the *E. coli* contamination dependent variable in White et al. (2021). LASSO regression is a technique that utilizes linear regression as a basis, and improves the results by introducing an L1 penalty term (penalizing the sum of absolute values of the weights versus the L2 penalty which penalizes the sum of squares of the weights), with the goal of minimizing the prediction error, to

combat overfitting of the model and allow for variable selection (Ranstam and Cook, 2018). This technique has been successfully used to determine variable importance for groundwater recharge potential (Pourghasemi et al., 2020). PLS is a multivariate statistical technique enabling comparison between multiple response and explanatory variables, designed in part to deal with multicollinearity (Pirouz, 2006). This technique has been successfully used in literature to predict groundwater quality and susceptibility to contamination (Ncibi et al., 2020; Sakizadeh and Ahmadpour, 2016). Regression techniques were employed to uncover trends between dependent variables, *E. coli* presence/absence and concentration, and independent variables (i.e., maximum number of consecutive wet days, maximum number of consecutive dry days, number of wet days, number of dry days, total rainfall, mean of rainfall, standard deviation of rainfall, maximum single day rainfall, number of wet-dry cycles, LULC, and stratigraphy) after classification and cluster analyses did not uncover unique rainfall-based patterns. MARS, GAMLSS, LASSO regression, and PLS are all used first for variable selection (from the 11 initial independent variables), then to create explanatory models targeting *E. coli* contamination in wells.

Models are created by first fine-tuning the model parameters, employing variable selection techniques, fine-tuning based on the new subset of significant variables, then training and testing on 10 unique data splits, encouraging a robust model output. Fine-tuning of MARS employs cross validation to identify the optimal degree of interactions and iterations, LASSO also uses cross validation to identify the optimal shrinkage penalty (lambda) value, and GAMLSS fitting families were chosen based on knowledge of the

dependent variable in the models (i.e., zero-inflated binomial for all *E. coli* data and Poisson Inverse Gaussian for adverse *E. coli* data). As each of the four techniques has the ability to perform a version of variable selection, models were built using 10-fold cross validation and variable importance was subsequently compared across three techniques for all *E. coli* data (MARS, LASSO, GAMLSS) and four techniques for adverse *E. coli* data (MARS, LASSO, GAMLSS, PLS). If a variable was considered insignificant by two or more techniques, it was removed from future analyses.

Based on the variable importance findings, model parameters were re-tuned utilizing cross validation. Ten unique data splits of 80% training and 20% testing were used to create a robust model and assess model performance (White et al., 2021). Model performance was compared using the coefficient of determination (R^2) and root mean squared error.

4.3.3.3 Association Rule Analysis

Association rule analysis (ARA) was employed for further data exploration purposes with the goal of uncovering additional trends in the all *E. coli* and adverse *E. coli* datasets. For all *E. coli* data, the goal was to understand variables associated with a present/absent observation. For adverse *E. coli* data, the goal was to understand variables associated with low, moderate, or high concentration observations. To conduct ARA the *a priori* algorithm “arules” (R package) was used (Hahsler et al., 2005), which identifies statistically interesting relationships in large datasets. The “interestingness” of a rule is based on four key measurements: confidence, which is the estimate of the conditional probability of an itemset Y given another itemset X (Hahsler et al., 2005); support, which is the proportion of observations in the dataset that contain the itemset X (Hahsler et al.,

2005); lift, which is the deviation of the support from the expected value assuming independence (Hahsler et al., 2005); and, standardized lift, which is the lift relative to its upper and lower bounds (McNicholas et al., 2008). Following White et al. (2021), standardized lift was used as the ranking method in this study. All analyses were conducted with a minimum level of support of 0.005 to increase the number of rules derived, and two to six items to ensure that the relationships considered are complex, but not overly so (McNicholas et al., 2008).

4.4 RESULTS AND DISCUSSION

4.4.1 Improving on Traditional Lag Times

An initial random forest analysis was undertaken to identify the most important days corresponding to *E. coli* presence in wells when considering wet and dry days. It can be seen that the most important time periods (identified as the lowest ranked) in predicting *E. coli* presence in private wells are from day n to $n-5$ where n represents the date of the sample collection. However, later days were also notable for their lower rankings, so, for the purpose of this exploratory study, the definition of important was expanded to include days n to $n-6$, $n-9$, $n-22$, and $n-45$ (i.e., days where the median fell below the dotted line in Figure 4-1A). To assess the impact of the various antecedent time period cut-offs to the predictive model, model accuracy was explored by adding one antecedent day at a time (Figure 4-1B).

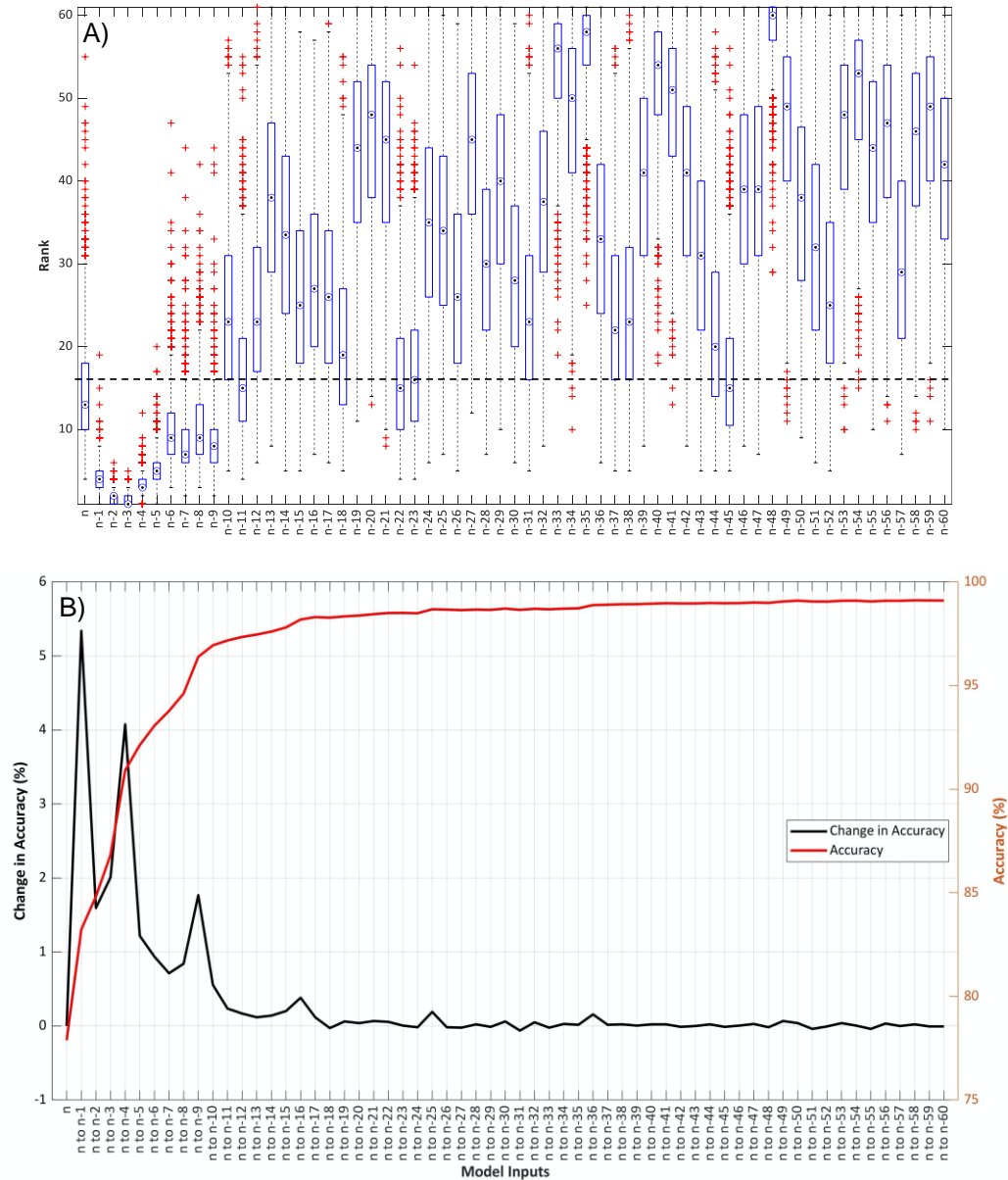


Figure 4-1: Summary output of random forest analysis exploring adverse *E. coli* observations for conditions from day n to day n-60. Figure A depicts the ranking order of antecedent days, where lower ranks represent more important variables. The dotted line highlights the highest mean rank of the antecedent days considered (i.e., n to n-9, n-22, and n-45). Figure B depicts the change in accuracy after the addition of one further antecedent day (i.e., n to n-45 represents the model accuracy when days n to n-45 are included in the model).

The random forest analysis shows that a majority of the increase in accuracy occurs between the n to $n-5$ days (Figure 4-1B). The increase in accuracy tails at the $n-16$ -day mark and the last slight increase in model accuracy is at the $n-36$ -day mark. This exploratory study primarily focuses on the n to $n-36$ -day period to capitalize on the highest model accuracy. However, cluster analyses also included the n to $n-5$ and n to $n-16$ days to see if trends would become more evident using less data.

4.4.2 Identification of Rainfall Intermittency Patterns

It has been well documented in the literature that rain after dry periods can result in pulses of *E. coli* in water bodies (“concentration effect”) (e.g., Baral et al., 2018; Carlton et al., 2014; Kraay et al., 2020; Levy et al., 2016; Mohanty et al., 2015) and rain after wet periods may not result in contamination (“dilution effect”) (Carlton et al., 2014), although if infrastructure (e.g., combined sewer systems, septic tanks) is overwhelmed, subsequent contamination events may occur (Kraay et al., 2020). Given the importance of dry-wet patterns on *E. coli* contamination in groundwater and using the identified 36-day lag period, the most frequent 10% dry-wet patterns associated with *E. coli* contamination in wells were identified utilizing a spectral clustering approach as described earlier in the Methods section (Figure 4-2).

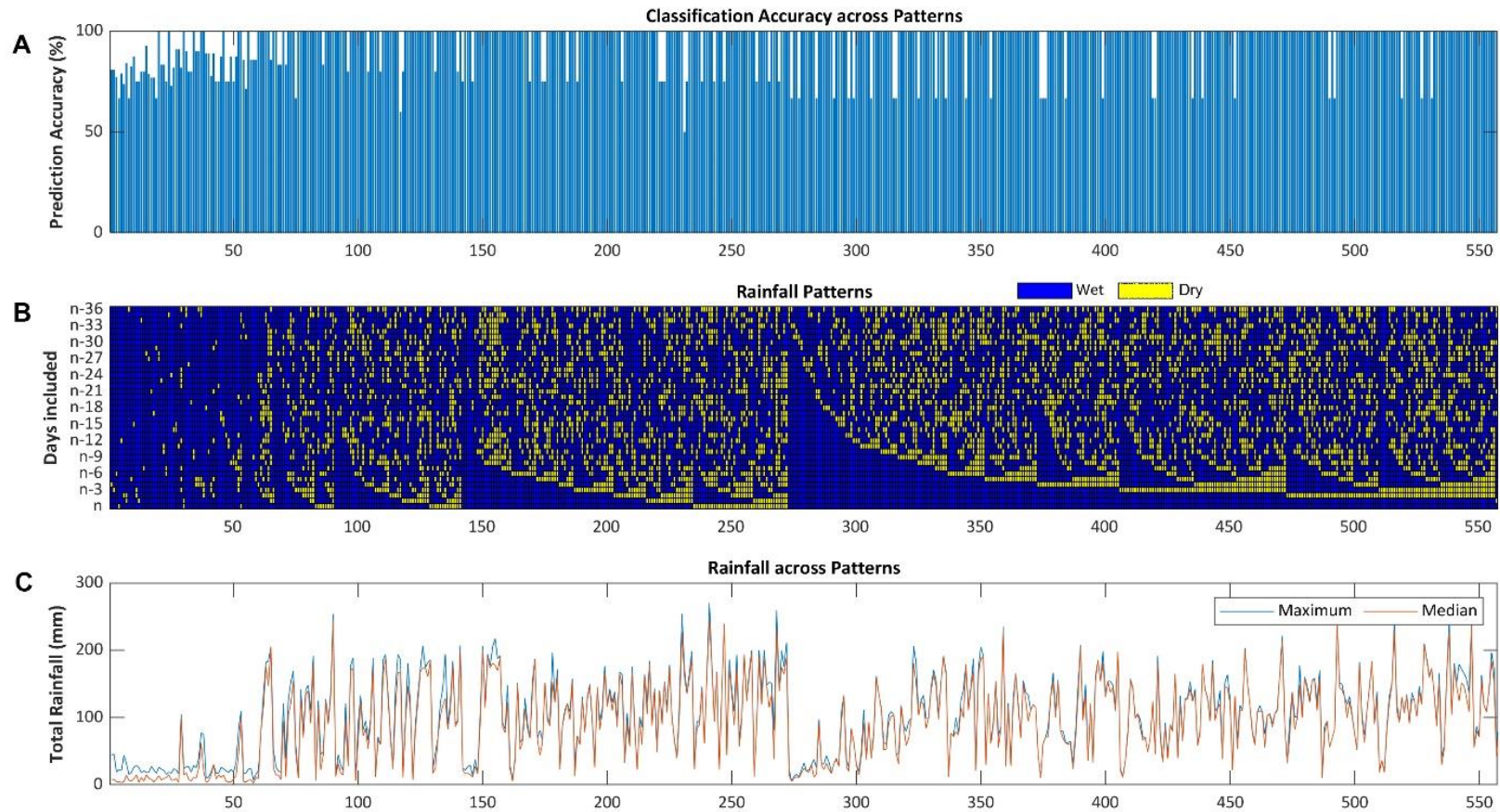


Figure 4-2: Summary of the top 10% (based on frequency) of dry-wet rainfall patterns leading to *E. coli* contamination in wells. A - represents the prediction accuracy of each pattern in predicting *E. coli* contamination. B - represents the dry-wet pattern over the 36-days preceding a contamination event, where blue represents wet days and yellow represents dry days. C – represents the maximum and median total rainfall found within a specific pattern, across all pattern occurrences. Note: the x-axis represents the order of patterns from most to least frequent.

The most frequent patterns (n~30) contained many “wet days”, but upon further investigation, the total accumulated rainfall was very low ($15.7 \text{ mm} \pm 15.4$). This suggests that there does not need to be a lot of rainfall to mobilize *E. coli*. What is seemingly more important is the occurrence of dry days leading up to a contamination event, allowing for *E. coli* accumulation to occur on the surface (“concentration effect”). The most ideal dry-wet sequence leading to *E. coli* contamination in wells was of many wet days with low total rainfall leading up to a contamination event, which improves contaminant transport conditions as the soil remains “primed” (i.e., damp enough to improve soil porewater connectivity), followed by a few dry days to encourage *E. coli* accumulation, and finally wet days right before the contamination event. This latter observation held through all of the most frequently occurring patterns, albeit with the total number of consecutive wet days varying from 3 to 30.

4.4.3 Identification of Meaningful Rainfall Intermittency Clusters

Following on from the wet-dry pattern analyses, cluster analysis was utilized as a means to determine identifiable clusters of rainfall patterns that result in *E. coli* contamination in a well. The purpose of this analysis was to determine specific characteristics of clusters to provide more meaningful insights for well users and policy makers. A critical element of this analysis was to determine whether or not the clusters associated with contamination events are significantly different from clusters that do not result in contamination events. As a starting point, spectral analysis was used to determine the optimal number of clusters based on the fraction of variance explained in the model, up

to a maximum of 30 clusters. Based solely on the machine learning method, the optimal number of clusters was established to be 14.

When the clustering technique was set to equal 14, the outputted results did not provide meaningful information. Instead, there were several clusters with similar characteristics that could be easily confused with one another (e.g., multiple clusters had similar rainfall amounts and wet to dry ratios) (Figure A.4-1). Upon further user-based clustering, it was determined that the optimal number of clusters for meaningful output is 6 clusters, also supported by the original kick point of the spectral analysis (Figure A.4-2; Figure 4-3). Clusters were determined using wet and dry day patterns (Figure 4-3A). Once identified, clusters were further described using *E. coli* concentration classes, *E. coli* LULC contamination potential classes, and, stratigraphy (Figure 4-3B). For clarification, *E. coli* concentrations, LULC, and stratigraphy were not included in cluster creation, rather as a method to describe the unique clusters. Inclusion of these variables in analyses is explored in Section 4.4.4.

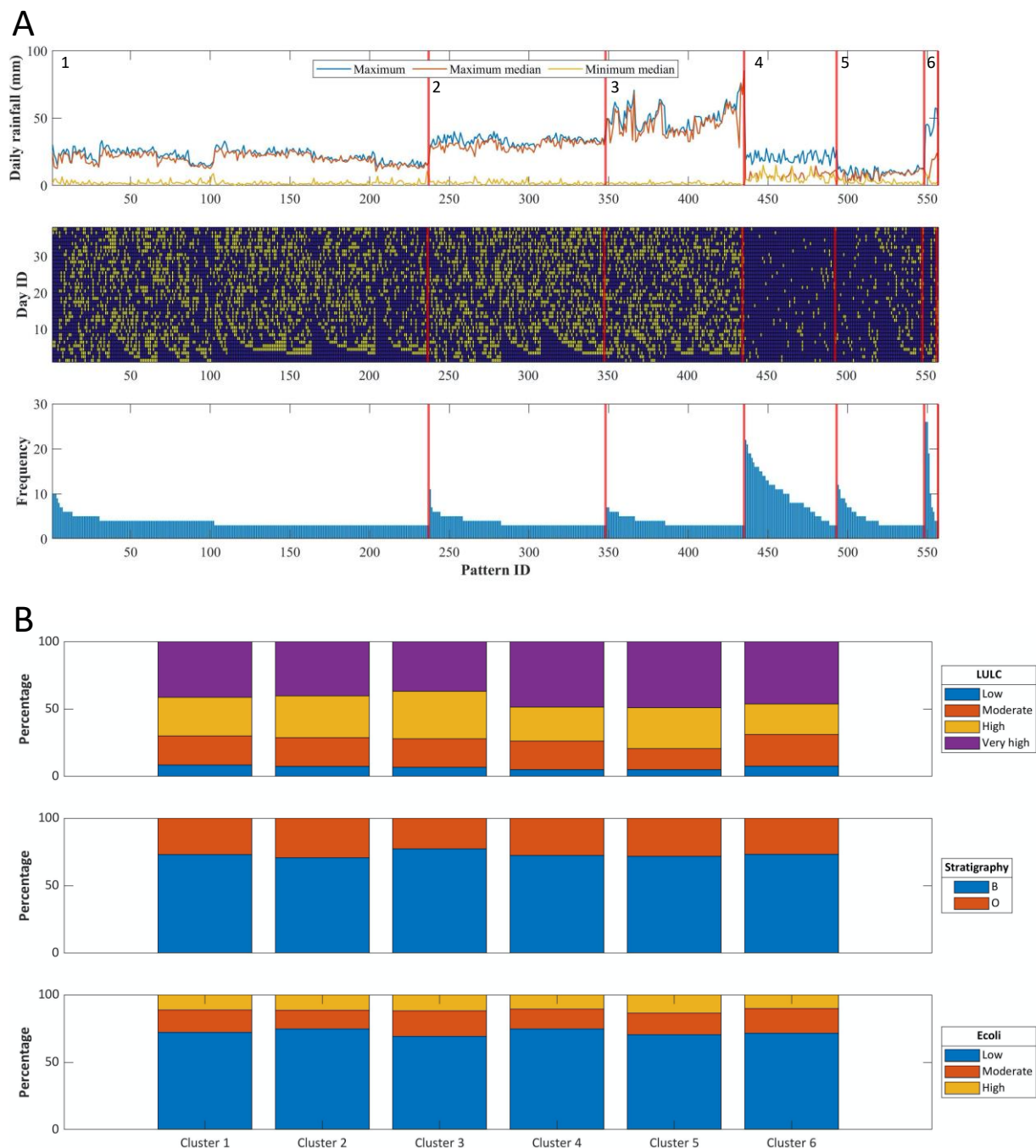


Figure 4-3: Box A depicts results from clustering showing rainfall pattern clusters resulting in *E. coli* contamination in wells. Box B depicts a summary of these 6 clusters based on the breakdown of LULC, Stratigraphy, and *E. coli* contamination severity, where the y-axis represents the percentage of the cluster.

The cluster that is associated with the highest concentrations of *E. coli* contamination (combination of high and moderate *E. coli* counts) is cluster 3, followed by cluster 5 and then 6. Further, cluster 3 contains the highest distribution of wells finished in bedrock, the highest accumulated rainfall amount, and the highest distribution of LULC of “high *E. coli* contamination potential”. So, it can be concluded that the combination of LULCs that increase the prevalence of *E. coli* and result in high surface runoff, stratigraphy that is likely to have preferential flow paths to wells, and higher average rainfall amounts result in the highest concentrations of *E. coli* contamination. A notable characteristic of cluster 5 is its highest distribution of LULC of “very high” and “high *E. coli* contamination potential” and its low average rainfall amount. Thus, it seems that if high *E. coli* loading is possible based on LULC, the required rainfall amount to transport *E. coli* through the system into a well is lower; this is a similar trend to that seen in cluster 4. Notably, cluster 4 has a higher proportion of low adverse *E. coli* contamination when compared to cluster 5. This may be attributed to cluster 4’s high dry day ratio - while there are many opportunities for the concentration effect to take place, the saturation level of the vadose zone may result in fewer connected pathways and rainfall amounts may be insufficient to push any or all contaminants through the system. In support of this explanation and further demonstration of the complexities involved in pathogen mobilization, 3 of 6 meaningful clusters in the non-adverse *E. coli* results had very similar high dry day ratios, that in these instances resulted in no presence of *E. coli* in wells. A full summary of clusters based on adverse *E. coli* observations can be found in Table 4-4.

Table 4-4: Summary of clusters based on adverse and non-adverse *E. coli* observations from n to n-36 days.

| Adverse <i>E. coli</i> Observations | | | | | | | Non-Adverse <i>E. coli</i> Observations | | | | | | |
|---|--------------|--------------|-----------|--------------|--------------|--------------|---|--------------|--------------|--------------|--------------|--------------|--------------|
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | All Dry Days | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
| Number of unique patterns | 237 | 111 | 87 | 58 | 55 | 9 | 1 | 84 | 72 | 6 | 5 | 4 | 4 |
| Number of observations | 875 | 412 | 326 | 602 | 257 | 119 | 24473 | 30570 | 7401 | 2877 | 588 | 24780 | 354 |
| Max number of wet/dry days | 21/35 | 26/34 | 21/35 | 8/36 | 13/36 | 23/36 | | 6/36 | 21/35 | 2/36 | 19/35 | 2/37 | 21/28 |
| Mean number of wet days | 11.3±4.68 | 12.93±4.53 | 12.26±3.3 | 1.71±1.26 | 4.44±3.33 | 6.67±7.3 | 0/37 | 1.80±.95 | 7.58±4.79 | 1.33±0.52 | 11.6±7.3 | 1.5±1 | 17.25±5.56 |
| Mean number of dry days | 25.7±4.68 | 24.07±4.53 | 24.74±3.3 | 35.29±1.26 | 32.56±3.33 | 30.33±7.3 | | 35.2±0.95 | 29.42±4.79 | 35.67±0.52 | 25.4±7.3 | 35.5±1 | 19.75±5.56 |
| Max rainfall range | 31.41 | 38.8 | 84.05 | 25.055 | 13.09 | 56.32 | -- | 40.61 | 39.31 | 57.64 | 17.96 | 18.14 | 40.01 |

| | | | | | | | | | | | | |
|--|------------|------------|-----------|-------------|-----------|-----------|-------------|------------|-------------|-----------|-------------|------------|
| (i.e., the maximum of each pattern's range within cluster) | | | | | | | | | | | | |
| Minimum | 5 | 25.13 | 35.85 | 2.49 | 1.34 | 33.84 | 20.98 | 18.46 | 44.41 | 12.9 | 0 | 37.82 |
| Mean | 20.37±4.98 | 31.55±2.95 | 49.93±9.6 | 15.08±5.05 | 7.64±2.85 | 42.35±7.9 | 29.73±4.52 | 27.67±4.86 | 49.84±4.44 | 16.27±2.1 | 12.74±8.6 | 38.99±1.03 |
| Max/Min | | | | | | | | | | | | |
| number of consecutive wet days | 13/1 | 13/1 | 8/1 | 3/1 | 6/1 | 6/1 | 3/1 | 8/1 | 2/1 | 8/1 | 1/1 | 6/1 |
| Mean | 1.93±1.32 | 1.98±1.45 | 1.82±1.25 | 1.39±0.52 | 1.42±0.79 | 1.94±1.46 | 1.47±0.57 | 1.92±1.22 | 1.14±0.38 | 2.15±1.68 | 1±0 | 2.38±1.37 |
| Max/Min | | | | | | | | | | | | |
| number of consecutive dry days | 35/1 | 34/1 | 18/1 | 36/1 | 33/1 | 34/1 | 36/1 | 33/1 | 34/1 | 16/1 | 37/1 | 8/1 |
| Mean | 4.04±4.41 | 3.45±3.29 | 3.37±2.69 | 16.64±10.26 | 8.37±8.34 | 7.8±10.56 | 17.09±10.19 | 6.56±7.24 | 17.83±14.84 | 4.7±3.78 | 15.78±12.81 | 2.63±1.94 |

| | | | | | | | | | | | | | |
|------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| LULC ratio: | | | | | | | | | | | | | |
| Very High | 0.412 | 0.4 | 0.368 | 0.485 | 0.488 | 0.462 | 0.397 | 0.418 | 0.373 | 0.398 | 0.318 | 0.282 | 0.456 |
| High | 0.287 | 0.311 | 0.353 | 0.252 | 0.305 | 0.227 | 0.285 | 0.318 | 0.366 | 0.33 | 0.52 | 0.393 | 0.319 |
| Moderate | 0.216 | 0.216 | 0.212 | 0.213 | 0.156 | 0.235 | 0.26 | 0.203 | 0.193 | 0.213 | 0.104 | 0.246 | 0.173 |
| Low | 0.085 | 0.073 | 0.067 | 0.05 | 0.051 | 0.076 | 0.058 | 0.062 | 0.069 | 0.059 | 0.058 | 0.079 | 0.052 |
| Stratigraphy ratio: | | | | | | | | | | | | | |
| Bedrock | 0.915 | 0.927 | 0.933 | 0.95 | 0.949 | 0.924 | 0.633 | 0.644 | 0.599 | 0.581 | 0.93 | 0.799 | 0.554 |
| Overburden | 0.085 | 0.073 | 0.067 | 0.05 | 0.051 | 0.076 | 0.367 | 0.356 | 0.401 | 0.419 | 0.07 | 0.201 | 0.446 |
| <i>E. coli</i> ratio: | | | | | | | | | | | | | |
| High | 0.112 | 0.114 | 0.12 | 0.106 | 0.136 | 0.101 | --- | --- | --- | --- | --- | --- | --- |
| Moderate | 0.167 | 0.138 | 0.19 | 0.148 | 0.16 | 0.185 | | | | | | | |
| Low | 0.721 | 0.748 | 0.69 | 0.746 | 0.704 | 0.714 | | | | | | | |

To ensure that the clusters associated with adverse *E. coli* contamination in wells are unique, pattern clusters were also explored for non-adverse *E. coli* results. Comparing the patterns between adverse and non-adverse observations found that there were 105 wet-dry patterns (around 0.5% of the patterns resulting in adverse observation) found in both analyses. The most common pattern found in non-adverse observations ($n = 24,473$) was 37 dry days (n to $n-36$); a lack of rainfall means that there is no transport mechanism to introduce contaminants into the system. While this pattern also resulted in adverse observations ($n = 322$), a well is statistically significantly ($p\text{-value} < 0.01$) less likely to experience an *E. coli* contamination event. However, none of the other clusters demonstrate strong differentiating trends sufficient to identify patterns of drivers associated with *E. coli* contamination in wells.

Given that exploring clusters for patterning between n to $n-36$ days did not provide any significant trends, additional cluster analyses were conducted for patterning using n to $n-5$ days (63 patterns; 692,578 observations; 84% cluster accuracy) and n to $n-16$ days (5,853 patterns; 436,283 observations; 92% cluster accuracy) to determine if less data would offer clearer trends. However, similar to the n to $n-36$ day trends, the n to $n-5$ (Table A.4-1) and n to $n-16$ day (Table A.4-2) periods did not uncover any significant trends. Generally, for both the n to $n-5$ and n to $n-16$ day cluster analyses, clusters capturing adverse observations have slightly higher proportions of very high LULC *E. coli* contamination potential (0.414 ± 0.026) and bedrock bottom of well stratigraphy (0.718 ± 0.019) when compared to clusters capturing non-adverse observations (0.389 ± 0.01 ; $p < 0.001$ and 0.638 ± 0.01 ; $p < 0.001$, respectively). This supports previous

findings that identified higher *E. coli* contamination potential associated with specific LULCs (i.e., agricultural/pastoral) (White et al., 2023a) and wells finished in consolidated formations (i.e., bedrock) (White et al., 2021). Clusters from the n to n-16 analysis show a trend in the number of wet and dry days within a pattern. Specifically, the adverse clusters have slightly fewer mean wet days (6.92 ± 6.14) and slightly more mean dry days (10.1 ± 6.14) than the non-adverse clusters (8 ± 6.81 ; $p < 0.001$ and 9 ± 6.81 ; $p < 0.001$, respectively). This trend supports literature discussing the impact of the concentration/dilution effect on contamination events, where these events are slightly more likely to occur with higher dry:wet day ratios. Since the cluster analyses using less data (i.e., n to n-5 and n to n-16 days) were not substantially more insightful than the n to n-36 day analysis, the remainder of these analyses only consider the n to n-36 day antecedent conditions to capitalize on the highest model accuracy.

4.4.4 Addition of Variables to Improve Accuracy

As previously noted, LULC *E. coli* Contamination Potential classes and stratigraphic data, that better represent loading, mobilization, and transport processes than rainfall alone, may improve predictive accuracy. Overall, compared to rainfall alone, the classification accuracy slightly increases from 98.33% at n-36 days to 98.52% at n-36 days when rainfall and LULC are used to predict *E. coli* contamination, followed closely by rainfall and stratigraphy (~98.42% at n-36 days). When all three predictors are used, the classification accuracy drops slightly (~98.27% at n-36 days) which may be due to too many redundant or irrelevant variables being considered, introducing unnecessary complexity and possible overfitting of the model, or confounding variables, all of which result in a decrease in

predictive accuracy (Čeperić et al., 2017). This makes it very clear that rainfall is the main driver associated with *E. coli* contamination in wells. While the addition of other variables slightly increases the predictive accuracy, the additional computational and data requirements would generally not be worth the resulting improvements.

4.4.5 Regression Techniques

4.4.5.1 Variable Selection

Exploration utilizing classification and clustering techniques provided an identification of the defined antecedent time period (i.e., $n - 36$ days) for contamination potential within wells, but was unable to identify unique rainfall-based patterns driving these contamination events. To better explain these occurrences, regression-based techniques were utilized (i.e., MARS, GAMLSS, LASSO Regression, PLS). Model building began with 11 independent variables (maximum number of consecutive wet days, maximum number of consecutive dry days, number of wet days, number of dry days, total rainfall, mean of rainfall, standard deviation of rainfall, maximum single day rainfall, number of wet-dry cycles, LULC, and stratigraphy). In an attempt to simplify models, variable selection was conducted using the four regression techniques and compared to identify common non-significant variables. When all *E. coli* data were included, with a dependent variable of presence/absence of *E. coli*, MARS, GAMLSS, and LASSO regression all identified number of dry days as non-significant in explaining *E. coli* presence/absence in a well; two of three techniques (MARS and GAMLSS) identified maximum single day rainfall and standard deviation of rainfall as non-significant, and one of three techniques (MARS) identified LULC, number of wet days, consecutive number of

wet days, stratigraphy, and total rainfall as non-significant. Therefore, maximum single day rainfall, standard deviation of rainfall, and number of dry days were removed from subsequent analyses focused on all *E. coli* data.

Considering only adverse *E. coli* observations (1 – 81 CFU/100mL), MARS only considered LULC as significant. The other three techniques (GAMLSS, LASSO, and PLS) provided more insights into variable selection, identifying standard deviation of rainfall as non-significant in explaining *E. coli* concentration in a well. Two of the techniques identified maximum single day rainfall (GAMLSS and PLS), consecutive dry days (GAMLSS and LASSO), and number of dry days (GAMLSS and LASSO) as non-significant, and one of three identified number of wet days (GAMLSS), mean rainfall (PLS), and total rainfall (PLS) as non-significant. This resulted in the removal of the same variables as for all *E. coli* data (i.e., standard deviation of rainfall, maximum single day rainfall, and number of dry days) and the additional removal of consecutive dry days, in subsequent analyses exploring adverse *E. coli* observations.

4.4.5.2 Regression Outputs

While regression techniques do not produce confident predictive models due to low R^2 values, models provided explanatory trends of all *E. coli* and adverse *E. coli* observations (Table 4-5). The variables mean and total rainfall were found to be significant when exploring all *E. coli* observations. As mean and total rainfall increases, there is a greater likelihood of *E. coli* presence in wells. Further, MARS models identified a threshold (knot) at 12.25mm mean rainfall, above which (≥ 12.25 mm) the likelihood of *E. coli*

presence increases 2x faster than values less than this threshold (< 12.25 mm). However, these variables were not deemed significant.

Based on summary outputs, LULC *E. coli* Contamination Potential class was identified as a notable variable. Generally, the higher the LULC *E. coli* Contamination Potential class, the more likely *E. coli* will be found in a well and at a higher concentration. The output of the adverse *E. coli*-based MARS model identifies a threshold (knot) at the high class (3); for values less than 3 (representing the transition from low to moderate and from moderate to high) the likelihood of *E. coli* presence increased at a rate 2x greater than the transition from high to very high *E. coli* contamination potential.

Models show that when considering consecutive wet days, the more consecutive wet days, the more likely it is that *E. coli* is present, but the greater the number of consecutive wet days, the lower concentration levels. This supports the theory that rainfall is required to introduce pathogens into a well, as well as the concentration and dilution effects - higher concentrations occur following fewer consecutive wet days and lower concentrations occur following more consecutive wet days. The number of dry-wet cycles depicts a similar finding with an increased number of cycles increasing the likelihood of experiencing *E. coli* presence while also decreasing the severity of the contamination event. Further, the MARS technique identifies a threshold (knot) at 11 cycles, whereby there is a gentler increase in likelihood of *E. coli* presence less than 11 cycles, and an average 3x greater rate increase past the 11 cycles.

Table 4-5: Summary of regression technique outputs for all *E. coli* and adverse *E. coli* observations.

| | All <i>E. coli</i> Observations | | | Adverse <i>E. coli</i> Observations | | | |
|--|---------------------------------|--------------------|------------------|-------------------------------------|-------------------|------------------|-------|
| | MARS* | GAMLSS | LASSO | MARS* | GAMLSS | LASSO | PLS** |
| Intercept | 100% | -5.96±3.97E-2 | -6.04E-2±1.23E-3 | 100% | 2.17±3.13E-2 | 7.35±5.46E-1 | |
| Stratigraphy | NS | -3.37E-1±6.43E-3 | -1.08E-2±2.02E-4 | NS | 6.23E-2 ± 9.06E-3 | 1.20 ± 1.82E-1 | 2 |
| LULC | NS | 6.58E-2 ± 2.99E-3 | 2.18E-3±1.00E-4 | 100% | 8.65E-2 ± 5.09E-3 | 1.60 ± 6.75E-2 | 1 |
| Consecutive Wet Days | NS | 1.73E-1 ± 3.41E-3 | 6.06E-3±1.22E-4 | NS | -1.76E-2±4.29E-3 | -2.43E-1±1.16E-1 | 3 |
| Consecutive Dry Days | 50% | 5.04E-2 ± 1.20E-3 | 2.01E-3±3.18E-5 | NS-VR | NS-VR | NS-VR | NS-VR |
| Number of Wet Days | NS | -6.86E-2 ± 2.57E-3 | -3.39E-3±7.96E-5 | NS | NS | NS | 5 |
| Total Rainfall | NS | 1.11E-3 ± 1.44E-4 | 1.52E-4±5.14E-6 | NS | NS | 7.21E-3±1.65E-3 | 6 |
| Mean Rainfall | 100% | 6.88E-2 ± 1.10E-3 | 1.27E-3±4.21E-5 | NS | NS | NS | NS |
| Number of Cycles | 100% | 2.04E-1 ± 2.32E-3 | 7.23E-3±7.37E-5 | NS | -7.73E-3±1.64E-3 | -1.21E-1±5.05E-2 | 3 |
| NS-VR – Not significant in model based on variable reduction in section 4.4.5.1 | | | | | | | |
| NS – Not deemed significant in model | | | | | | | |
| * Cell value represents the percentage of the 10 training-testing data splits a given variable was deemed significant in | | | | | | | |
| ** Cell value represents the summarized variable importance ranking across the 10 training-testing data splits | | | | | | | |

4.4.5.3 Association Rule Analyses

Association rule analyses (ARA) conducted on all *E. coli* data found that, when only interested in rules containing an adverse event, all rules require stratigraphy to be consolidated. This could be due to the existence of preferential flow paths due to fracturing, resulting in rapid transport of *E. coli* from the surface to the well. Additionally, of the eight rules identified, five required 13-24 wet-dry cycles. Similar to the regression-based techniques, this may reflect the concentration and dilution effects. Four of eight rules identified a total rainfall range of 124 – 321mm required for *E. coli* to be present in a well. Exploration of ARA results based on adverse *E. coli* data, found that the majority of rules resulting in an adverse *E. coli* category of moderate or high required the LULC contamination potential class to be very high. Rainfall patterns specifically did not appear to result in any notable trends, despite previously having been identified as the primary driver of contamination (accuracy assessment with additional variables).

4.5 STUDY LIMITATIONS

The WWTD is subject to methodological limitations associated with *E. coli* quantification uncertainties which cannot be quantified with the available data. Further, the WWIS database is subject to data entry errors, as borehole logs are hand recorded in the field and later transcribed into an online database, some of which date back to the 1910's. Rainfall values were not measured directly, rather they are based on an estimated conversion from precipitation to rainfall, but the uncertainties from these calculations cannot be quantified as actual values are not known.

4.6 CONCLUSION

Studies exploring the impact of rainfall lag times on *E. coli* contamination of water and diarrhea occurrence in the human population have been well documented, though there is no strong consensus. While data availability can be a limitation in other studies, this study utilizes a large dataset and an array of machine learning techniques, including clustering, classification, and regression techniques, in an attempt to improve understanding of rainfall lag times and strengthen understanding of key rainfall pattern characteristics. A combination of classification and clustering techniques defined an antecedent time period to consider in predicting *E. coli* presence in wells was n to $n-36$ days, although a computationally conservative lag time of n to $n-16$ days is acceptable. Additionally, if there is no rainfall present during the n to $n-36$ days, there is a statistically significantly lower chance that a well will be contaminated, likely due to the lack of a transport mechanism. Although this work was conducted on a large dataset, clustering and classification techniques were not able to identify any additional confident trends when considering antecedent rainfall conditions leading to well contamination.

Although only one rainfall intermittency pattern surfaced as important (zero rain in the preceding period), regression techniques did identify some insights to explain *E. coli* presence and severity in wells:

- Increasing LULC contamination potential classes resulted in an increased likelihood of *E. coli* presence and severity in wells, with a threshold at the high class (3) resulting in severity increasing at a 2x faster rate from classes low to moderate and moderate to high, versus from high to very high;

- Increasing the number of consecutive wet days and number of wet-dry cycles increases the likelihood of *E. coli* presence and decreases the likely severity of the contamination events. Further, a knot is present at 11 cycles where likelihood of *E. coli* presence increases at a rate 3x faster past the 11-cycle point, and;
- Association rule analysis suggests that wells located in consolidated material are at a higher likelihood of *E. coli* presence given a source of contamination and rainfall to mobilise it.

This study demonstrates that, even with a large multi-dimensional dataset, the presence and severity of *E. coli* in private wells cannot be predicted utilizing only rainfall intermittency patterns. However, the identification of explanatory variables, their relative importance, and effects on *E. coli* presence, in combination with other explanatory considerations (e.g., well characteristics, topographic characteristics) can be used to inform and advance the development of future predictive data-driven fate and transport models.

4.7 WORKS CITED

- Baral, D., Speicher, A., Dvorak, B., Admiraal, D., Li, X., 2018. Quantifying the relative contributions of environmental sources to the microbial community in an urban stream under dry and wet weather conditions. *Appl. Environ. Microbiol.* 84. <https://doi.org/10.1128/AEM.00896-18>
- Bindal, S., Singh, C.K., 2019. Predicting groundwater arsenic contamination: Regions at risk in highest populated state of India. *Water Res.* 159, 65–76. <https://doi.org/10.1016/j.watres.2019.04.054>
- Buckerfield, S.J., Quilliam, R.S., Waldron, S., Naylor, L.A., Li, S., Oliver, D.M., 2019. Rainfall-driven *E. coli* transfer to the stream-conduit network observed through

- increasing spatial scales in mixed land-use paddy farming karst terrain. *Water Res.* X 5. <https://doi.org/10.1016/j.wroa.2019.100038>
- Carlton, E.J., Eisenberg, J.N.S., Goldstick, J., Cevallos, W., Trostle, J., Levy, K., 2014. Heavy rainfall events and diarrhea incidence: The role of social and environmental factors. *Am. J. Epidemiol.* 179, 344–352. <https://doi.org/10.1093/aje/kwt279>
- Čeperić, E., Žiković, S., Čeperić, V., 2017. Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. *Energy* 140, 893–900. <https://doi.org/10.1016/j.energy.2017.09.026>
- Chandrasena, G.I., Deletic, A., Hathaway, J.M., Lintern, A., Henry, R., McCarthy, D.T., 2019. Enhancing *Escherichia coli* removal in stormwater biofilters with a submerged zone: balancing the impact of vegetation, filter media and extended dry weather periods. *Urban Water J.* 16, 460–468. <https://doi.org/10.1080/1573062X.2019.1611883>
- Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., Ren, H., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* 171, 115454. <https://doi.org/10.1016/j.watres.2019.115454>
- Curriero, F.C., Patz, J.A., Rose, J.B., Lele, S., 2001. The Association Between Extreme Precipitation and Waterborne Disease Outbreaks in the United States, *Research Articles | Peer Reviewed | Curriero et al. Public Health.*
- De Roos, A.J., Kondo, M.C., Robinson, L.F., Rai, A., Ryan, M., Haas, C.N., Lojo, J., Fagliano, J.A., 2020. Heavy precipitation, drinking water source, and acute gastrointestinal illness in Philadelphia, 2015-2017. *PLoS One* 15, e0229258. <https://doi.org/10.1371/journal.pone.0229258>
- Drayna, P., McLellan, S.L., Simpson, P., Li, S.H., Gorelick, M.H., 2010. Association between rainfall and pediatric emergency department visits for acute gastrointestinal

- illness. Environ. Health Perspect. 118, 1439–1443.
<https://doi.org/10.1289/ehp.0901671>
- Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97, 611–631.
<https://doi.org/10.1198/016214502760047131>
- Friedman, J., 1991. Multivariate Adaptive Regression Splines. *Annals of Statistics* 19.
- Gleason, J.A., Fagliano, J.A., 2017. Effect of drinking water source on associations between gastrointestinal illness and heavy rainfall in New Jersey. *PLoS One* 12, e0173794. <https://doi.org/10.1371/journal.pone.0173794>
- Godfrey, S., Timo, F., Smith, M., 2005. Relationship between rainfall and microbiological contamination of shallow groundwater in Northern Mozambique. *Water SA* 31, 609–614. <https://doi.org/10.4314/wsa.v31i4.5152>
- Goforth, E., Yosri, A., El-Dakhakhni, W., Wiebe, L., 2022. Rapidly Prediction of Power Infrastructure Forced Outages: Data-Driven Approach for Resilience Planning. *J. Energy Eng.* 148, 04022016. [https://doi.org/10.1061/\(asce\)ey.1943-7897.0000836](https://doi.org/10.1061/(asce)ey.1943-7897.0000836)
- Guzman Herrador, B.R., De Blasio, B.F., MacDonald, E., Nichols, G., Sudre, B., Vold, L., Semenza, J.C., Nygård, K., 2015. Analytical studies assessing the association between extreme precipitation or temperature and drinking water-related waterborne infections: A review. *Environ. Heal. A Glob. Access Sci. Source* 14. <https://doi.org/10.1186/s12940-015-0014-y>
- Haggag, M., Yorsi, A., El-Dakhakhni, W., Hassini, E., 2021. Infrastructure performance prediction under Climate-Induced Disasters using data analytics. *Int. J. Disaster Risk Reduct.* 56, 102121. <https://doi.org/10.1016/j.ijdr.2021.102121>
- Hahsler, M., Grün, B., Hornik, K., 2005. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *J. Stat. Softw.* 14. <https://doi.org/10.18637/jss.v014.i15>

- He, S., Wu, J., Wang, D., He, X., 2022. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere* 290, 133388. <https://doi.org/10.1016/j.chemosphere.2021.133388>
- Jagai, J.S., Li, Q., Wang, S., Messier, K.P., Wade, T.J., Hilborn, E.D., 2015. Extreme precipitation and emergency room visits for gastrointestinal illness in areas with and without combined sewer systems: An analysis of Massachusetts data, 2003–2007. *Environ. Health Perspect.* 123, 873–879. <https://doi.org/10.1289/ehp.1408971>
- Jennings, K.S., Winchell, T.S., Livneh, B., Molotch, N.P., 2018. Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. *Nature Communications* 9, 1–9. <https://doi.org/10.1038/s41467-018-03629-7>
- Kraay, A.N.M., Man, O., Levy, M.C., Levy, K., Ionides, E., Eisenberg, J.N.S., 2020. Understanding the impact of rainfall on diarrhea: Testing the concentration-dilution hypothesis using a systematic review and meta-analysis. *Environ. Health Perspect.* <https://doi.org/10.1289/EHP6181>
- Lai, H., Hales, S., Woodward, A., Walker, C., Marks, E., Pillai, A., Chen, R.X., Morton, S.M., 2020. Effects of heavy rainfall on waterborne disease hospitalizations among young children in wet and dry areas of New Zealand. *Environment International* 145. <https://doi.org/10.1016/j.envint.2020.106136>
- Latchmore, T., Hynds, P., Brown, R.S., Schuster-Wallace, C., Dickson-Anderson, S., McDermott, K., Majury, A., 2020. Analysis of a large spatiotemporal groundwater quality dataset, Ontario 2010–2017: Informing human health risk assessment and testing guidance for private drinking water wells. *Sci. Total Environ.* 738. <https://doi.org/10.1016/j.scitotenv.2020.140382>
- Levy, K., Woster, A.P., Goldstein, R.S., Carlton, E.J., 2016. Untangling the Impacts of Climate Change on Waterborne Diseases: A Systematic Review of Relationships between Diarrheal Diseases and Temperature, Rainfall, Flooding, and Drought. *Environ. Sci. Technol.* <https://doi.org/10.1021/acs.est.5b06186>

- Lin, L., Tang, C., Dong, G., Chen, Z., Pan, Z., Liu, J., Yang, Y., Shi, J., Ji, R., Hong, W., 2021. Spectral clustering to analyze the hidden events in single-molecule break junctions. *J. Phys. Chem. C* 125, 3623–3630. <https://doi.org/10.1021/acs.jpcc.0c11473>
- McNicholas, P.D., Murphy, T.B., O'Regan, M., 2008. Standardising the lift of an association rule. *Comput. Stat. Data Anal.* 52, 4712–4721. <https://doi.org/10.1016/j.csda.2008.03.013>
- Mohanty, S.K., Saiers, J.E., Ryan, J.N., 2015. Colloid Mobilization in a Fractured Soil during Dry-Wet Cycles: Role of Drying Duration and Flow Path Permeability. *Environ. Sci. Technol.* 49, 9100–9106. <https://doi.org/10.1021/acs.est.5b00889>
- Murray, F.W., 1966. On the Computation of Saturation Vapor Pressure. *J. Appl. Meteorol. Climatol.* 6, 203–204. [https://doi.org/https://doi.org/10.1175/1520-0450\(1967\)006%3C0203:OTCOSV%3E2.0.CO;2](https://doi.org/10.1175/1520-0450(1967)006%3C0203:OTCOSV%3E2.0.CO;2)
- Naghibi, S.A., Hashemi, H., Berndtsson, R., Lee, S., 2020. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *J. Hydrol.* 589, 125197. <https://doi.org/10.1016/j.jhydrol.2020.125197>
- Ncibi, K., Chaar, H., Hadji, R., Baccari, N., Sebei, A., Khelifi, F., Abbes, M., Hamed, Y., 2020. A GIS-based statistical model for assessing groundwater susceptibility index in shallow aquifer in Central Tunisia (Sidi Bouzid basin). *Arab. J. Geosci.* 13. <https://doi.org/10.1007/s12517-020-5112-7>
- Nichols, G., Lane, C., Asgari, N., Verlander, N.Q., Charlett, A., 2009. Rainfall and outbreaks of drinking water related disease and in England and Wales. *J. Water Health* 7, 1–8. <https://doi.org/10.2166/wh.2009.143>
- Park, S., Hamm, S.Y., Jeon, H.T., Kim, J., 2017. Evaluation of logistic regression and multivariate adaptive regression spline models for groundwater potential mapping using R and GIS. *Sustain.* 9. <https://doi.org/10.3390/su9071157>

- Pirouz, D.M., 2006. An Overview of Partial Least Squares. SSRN Electron. J. 4, 1–55.
<https://doi.org/10.2139/ssrn.1631359>
- Pourghasemi, H.R., Sadhasivam, N., Yousefi, S., Tavangar, S., Ghaffari Nazarlou, H., Santosh, M., 2020. Using machine learning algorithms to map the groundwater recharge potential zones. J. Environ. Manage. 265, 110525.
<https://doi.org/10.1016/j.jenvman.2020.110525>
- Ranstam, J., Cook, J.A., 2018. Statistical nugget LASSO regression.
<https://doi.org/10.1002/bjs.10895>
- Sakizadeh, M., Ahmadpour, E., 2016. Geological impacts on groundwater pollution: a case study in Khuzestan Province. Environ. Earth Sci. 75, 1–12.
<https://doi.org/10.1007/s12665-015-4944-z>
- Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. J. Stat. Softw. 23, 1–46.
<https://doi.org/10.18637/jss.v023.i07>
- Thomas, M.K., Charron, D.F., Waltner-Toews, D., Schuster, C., Maarouf, A.R., Holt, J.D., 2006. A role of high impact weather events in waterborne disease outbreaks in Canada, 1975-2001. Int. J. Environ. Health Res. 16, 167–180.
<https://doi.org/10.1080/09603120600641326>
- Thornton, P.E., Shrestha, R., Thornton, M., Kao, S.C., Wei, Y., Wilson, B.E., 2021. Gridded daily weather data for North America with comprehensive uncertainty quantification. Sci. Data 8. <https://doi.org/10.1038/s41597-021-00973-0>
- Trenberth, K.E., Zhang, Y., Gehne, M., 2017. Intermittency in precipitation: Duration, frequency, intensity, and amounts using hourly data. J. Hydrometeorol. 18, 1393–1412. <https://doi.org/10.1175/JHM-D-16-0263.1>
- White, K., Dickson-Anderson, S., Majury, A., McDermott, K., Hynds, P., Brown, R.S., Schuster-Wallace, C., 2021. Exploration of *E. coli* contamination drivers in private

- drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset. *Water Res.* 197, 117089. <https://doi.org/10.1016/j.watres.2021.117089>
- White, K., Schuster-Wallace, C., Dickson-Anderson, S., 2022. Converting Land Use – Land Cover to *E. coli* Contamination Potential Classes for Groundwater Wells: Utilizing a Large Ontario-based Dataset. Manuscript submitted for publication
- Wu, J., Yunus, M., Islam, M.S., Emch, M., 2016. Influence of Climate Extremes and Land Use on Fecal Contamination of Shallow Tubewells in Bangladesh. *Environ. Sci. Technol.* 50, 2669–2676. <https://doi.org/10.1021/acs.est.5b05193>
- Xu, R., Wunsch, D., 2008. Clustering. John Wiley & Sons, Ltd.
- Yousefi, S., Sadhasivam, N., Pourghasemi, H.R., Ghaffari Nazarlou, H., Golkar, F., Tavangar, S., Santosh, M., 2020. Groundwater spring potential assessment using new ensemble data mining techniques. *Meas. J. Int. Meas. Confed.* 157, 107652. <https://doi.org/10.1016/j.measurement.2020.107652>
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: A review. *J. Hydrol.* <https://doi.org/10.1016/j.jhydrol.2021.126266>
- Zumel, N., Mount, J., 2019. Practical Data Science with R, Frontiers in Bioscience - Landmark. Simon and Schuster.

CHAPTER 5 TOWARDS A COUPLED-SYSTEMS

APPROACH FOR THE EXPLORATION OF *E. COLI*

CONTAMINATION IN PRIVATE DRINKING

WATER WELLS

Summary of Paper 4: Towards a Coupled-System Approach for the Exploration of *E. coli* Contamination in Private Drinking Water Wells, by K. White, C. Schuster-Wallace, and S. Dickson-Anderson, Will be submitted to Water Research by August 2023.

Summary:

This research pulls together chapters 2 through 4 by combining identified trends in well characteristics, human behaviour, hydrogeology, LULC based *E. coli* contamination potential, and rainfall intermittency patterning and exploring new variables including geographically driven seasonal delineations, snow water equivalent, and a user's closest drop-off location accessibility, and geographic testing patterns, into a coupled-systems model. This coupled-systems model explores how the physical and social environments come together to impact *E. coli* in private wells in Ontario. The results of this demonstrated:

- A novel geographically-driven seasonal threshold introduced for Ontario improves understanding of impacts of physical and social variables on *E. coli* presence in private wells;
- Human behaviour-based variables significant in most models, validating the need to use a coupled-systems approach;
- A threshold of convenience testing (<35 minutes one-way driving to sample drop-off) versus habitual testing (>50 minutes one-way driving to sample drop-off) emerged; and,
- LULC *E. coli* contamination potential has the greatest and most robust association with *E. coli* presence and concentration in wells for all models and seasons.

5.1 ABSTRACT

In Canada, there is a heavy reliance on private drinking water sources which presents unique challenges due to a lack of governmental regulations and limited resources for maintenance, management, and protection, which collectively may lead to negative impacts on human health. The purpose of this study is to develop an initial model capturing meteorological, hydrogeological, microbiological, well characteristic, and human behaviour variables, as a first step towards a novel coupled-systems approach to explaining the presence and concentration of *E. coli* in private wells.

Innovative supervised regression-based machine learning techniques were applied to a compilation dataset of 5 unique datasets with 795,023 observations and 33 variables. Three different regression techniques, univariate and bivariate analyses, and variable discretization as well as a new approach to identify seasonal delineations for better articulation of discrete hydro-meteorological regimes (i.e., summer, autumn and spring, winter) were utilized to improve the understanding of relative impacts and importance of physical and human behaviour variables on well contamination.

Key findings include: human behaviour-based variables are important in most models; a coupled-systems approach is necessary to increase explanatory power; Land Use-Land Cover has the greatest impact on *E. coli* presence and concentration; weather-based variables are only significant in the “shoulder” (spring/autumn) and “winter” models, demonstrating the importance of the concentration-dilution effect; and, critical well characteristic variables include well depth and bottom of well stratigraphy. These findings

demonstrate that applying regression-based machine learning techniques to a large, coupled-systems dataset, will have a wide array of applications both geographically and conceptually. Further, it offers an innovative way forward to improve knowledge and understanding of private well contamination and stewardship for wells owners and policy makers.

5.2 INTRODUCTION

There is a heavy reliance on groundwater as a drinking water source both globally (35%; UN Water, 2015) and within the Canadian context (33%; Boisvert and Cotteret, 2021). Private groundwater sources, relied on for drinking water by 12% (~4.5 million) of Canadians (Rivera, 2017), often present challenges including lack of regulation and limited resources for maintenance, management, and protection (Murphy et al., 2016). The dependence on these private drinking water systems results in an increased health risk due to potential exposure to microbial contamination, such as *Escherichia coli* (*E. coli*), causing acute gastrointestinal illness (AGI). To mitigate this vulnerability to illness caused by the reliance on private wells, it is important to better understand how and when these pathogens are entering and travelling through the groundwater system and the potential exacerbation or mitigation of risk through well stewardship. This can be improved through an understanding of the impact of weather patterns, hydrogeological conditions, contamination sources and loadings, and well conditions (location, construction, and maintenance) on the fate and transport of contaminants in aquifers (O'Dwyer et al., 2018)

combined with the knowledge, attitudes, and practices of private wells users (Lavallee et al., 2021).

In Ontario, winters are cold, with precipitation often falling as snow and high likelihood of frozen ground and water bodies. Spring seasons are often more unpredictable with temperatures beginning to increase above zero, resulting in mixed precipitation (snow, rain, and sleet) in conjunction with thaw of soils, snowpacks, and river and lake ice. Summers are often hot and humid, particularly in southern Ontario, with the potential for intense summer storms generating significant runoff events and localised flooding. Autumn seasons are often characterized by cooling temperatures resulting in mixed precipitation. These geographically-driven unique seasons impact the loading, fate, and transport of *E. coli* in the subsurface (Michel et al., 1999; Reynolds et al., 2020; White et al., 2021). White et al. (2021) explored the impacts of seasonality (latitude and longitude, quarterly seasons) on *E. coli* contamination in private wells and found that latitude was an important variable and seasonality was introduced as a possible driver of contamination. However, the findings were not conclusive, likely due to the use of static four-month blocks across all latitudes. This is not surprising, as seasons vary significantly across Ontario in terms of average conditions, onset, and duration due to the large geographic extent (OASDI, 2023), hence the greater investigation of seasons in this article.

As seasonally driven heavy precipitation events, snow melt, and wet-dry patterning are tied to acute gastrointestinal illness (AGI) in humans (De Roos et al., 2020; Kraay et al., 2020; Namugize et al., 2018; Whitman et al., 2008), the uncertainty caused by spatially

dependent seasonal onset results in differential risks to private well users. This has been explored in depth by White et al. (2023b), which found a critical antecedent period of 36-days during which rainfall patterns are associated with *E. coli* presence and concentrations in wells. Specifically, no rainfall in those 36 antecedent days significantly reduced the likelihood of *E. coli* in wells. Additionally, increasing the number of consecutive wet days, as well as the number of wet/dry cycles, was associated with an increased likelihood of *E. coli* presence in wells, but at reduced concentrations.

LULC types are associated with unique loading and transport mechanisms that may drive *E. coli* contamination in wells or protect against it. For example, agricultural areas increase likelihood of contamination due to animal presence increasing loading, and areas such as wetlands are protective against contamination due to their natural filtration properties (White et al., 2023a). While several studies have aimed to explain or predict *E. coli* presence or concentration in surface water or groundwater based on LULC (Jabbar et al., 2019; Namugize et al., 2018; Paule-Mercado et al., 2016; White et al., 2023a), none have included a full range of LULC types as an independent variable in a data-driven multivariate regression model for private wells, as presented here.

A well user's knowledge, attitudes, and practices (KAP), as well as perceptions regarding their well, have been linked to well stewardship behaviour and therefore human exposure to groundwater-borne pathogens (Lavallee et al., 2023, 2021). This opens the opportunity to further explore how human behaviours affect and are affected by the presence and concentration of *E. coli* in private wells. For example, a link has been

established in the literature between increased frequency of *E. coli* occurrence and increased sampling frequency (Latchmore et al., 2020), though there is no evidence to suggest this is a causal relationship, rather the association likely exists because contamination events can only be detected if they are actively tested for. Further, it has been found that well users will be more likely to repeat testing to seek reassurance that their water is safe, either out of habit or after receiving an adverse sample result (Qayyum et al., 2020). Building on these different bodies of research, greater coupling through the integration of variables that represent both natural and human systems (Di Pelino et al., 2019) in a single data-driven model is anticipated to capture more comprehensive and realistic connections to better inform risk mitigation strategies for private well users.

This work aims to develop an advanced, data-driven model based on environmental and human behaviour processes, towards the development of a coupled-systems model exploring *E. coli* presence in private wells (Figure 5-1). The model is based on a dataset that captures meteorological, hydrogeological, microbiological, well characteristic, and human behaviour variables. This is undertaken through a novel application of supervised regression machine learning techniques, basing results in process knowledge, to improve the understanding of relative impacts and importance of physical and human behaviour variables on well contamination, a feat too complex for typical process-based models. In developing the model, an additional objective is to develop a geographically-driven approach to capture variations in seasonality associated with latitude.

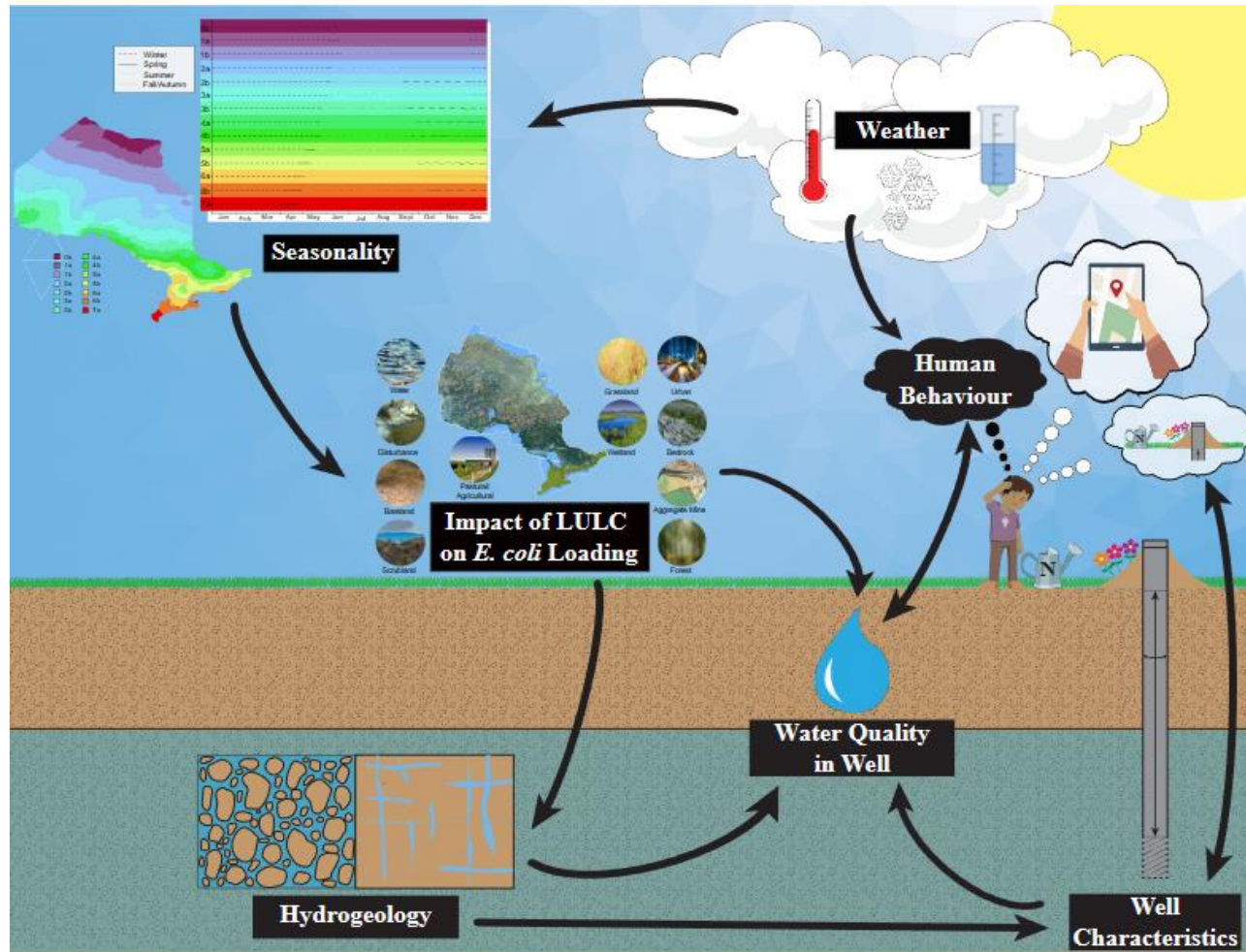


Figure 5-1: Summary of key elements of coupled-systems model explored in this work to explain *E. coli* contamination events in private wells.

5.3 METHODS

5.3.1 Datasets

The analyses in this paper have been undertaken in part using an Ontario-specific groundwater dataset that contains 795,023 well sample observations for 253,136 unique private wells detailing microbial testing, well characteristics, and hydrogeological characteristics, combining the Well Water Testing Dataset (WWTD) and the Well Water Information System (WWIS). These datasets have been previously explored in Latchmore et al. (2020) and White et al. (2021). Additional datasets were combined with the private well dataset to explore the impacts of topographic, meteorological, land use land cover, and transportation network variables on *E. coli* contamination and the impact of sample drop-off accessibility (Table 5-1).

Table 5-1: Summary of datasets and variable used in this study.

| Dataset | Variables | Specifications | Reference |
|---|---|--|--|
| Private Well Data (Well Water Testing Database (WWTD) augmented by Well Water Information System (WWIS)) | <i>E. coli</i> observations Well location Well depth Bottom stratigraphy Specific capacity Testing date Testing frequency | Data available from 2010 to 2017 Point observations | (Latchmore et al., 2020; White et al., 2021) |
| Domestic Well Data (WWIS only) | Casing Material Drilling Method Well Location | Data available from 1899 to 2021 Point observations | (Ontario Ministry of Environment Conservation and Parks, 2020) |

| | | | |
|---|---|---|---|
| Meteorological Data (Daymet V4.0) | Snow Water Equivalent (SWE) | Data available from 1980 to present day Grid size 1km x 1km | (Thornton et al., 2020) |
| | Rainfall | Data available from 2010 to 2017 Grid size 1km x 1km | (White et al., 2023b) |
| | Plant Hardiness Zones | Data available from 1981 – 2010 Grid size 2km x 2km | (McKenney et al., 2014) |
| Topographic Data (Provincial Digital Elevation Model (PDEM)) | Elevation | Data available from 2020 Grid size 30m x 30m | (Ontario Ministry of Natural Resources and Forestry, 2019) |
| LULC Data | LULC <i>E. coli</i> contamination potential classes | Data available from 1999 to 2015 Grid size 15m x 15m | (White et al., 2023a) |
| Roads Network | Roads network Posted Speed | Data available from 2015 to 2020 Network | (DMTI Spatial Inc., 2021) |

5.3.2 Data Processing

External datasets were linked to the private well dataset using the Spatial Analyst toolbox in ArcGIS and new variables created from the combined dataset (Table 5-2). Given the range of spatial resolutions represented in these different datasets, all values were linked to the unique well location (point data) in the new combined dataset. For snow water

equivalent (SWE) and rainfall data, the day of the water quality test was used in combination with well location to create spatio-temporal datasets.

Table 5-2: Summary of all variables.

| Parameter | Description | Continuous Data Range | Sub-Classifications Derived for Current Analysis |
|---|--|---|---|
| <i>E. coli</i> Result (from White et al., 2021) | Number of <i>E. coli</i> reported in sample by laboratory. Laboratory Reporting Range: 0 – 80 CFU/100 mL | 0-81 CFU/100mL | non-detects (ND): 0 Category 1: 1-10 Category 2: 11-50 Category 3: |
| Location (from White et al., 2021) | Location of well geographically, in longitude and latitude | Latitude: -85.5 - -75.0 Longitude: 42.5 – 49.0 | Binned into 0.5 degree ranges |
| Date of Observation (from White et al., 2021) | Date of water sample collection | Jan 1, 2010 – Dec 31, 2017 | |
| Geological Formation (from White et al., 2021) | Stratigraphy of geologic formation in which well is situated (originally recorded in ft) | N/A | Consolidated (further categorised as igneous, metamorphic, sedimentary) (1) Unconsolidated (0) |
| Specific Capacity (from White et al., 2021) | Information recorded from pump test includes static water level, water level after pumping, pump test rate, and pump test duration (originally recorded in GPM/ft) | 0.001 – 1919 GPM/ft | Low (0 - <3.3 GPM/m) Moderate (3.3 – 16.4 GPM/m) High (>16.4 GPM/m) |
| Well Depth (from White et al., 2021) | Distance from ground surface to bottom of well | 1 - 760m | Shallow/Moderate (< 12.5 m) |

| | | | |
|---|--|---|---|
| | (originally recorded in ft) and classification of well depth | | Moderate 1 ($12.5 \text{ m} \leq x < 18.3 \text{ m}$) Moderate 2 ($18.3 \text{ m} \leq x < 24.4 \text{ m}$) Moderate 3 ($24.4 \text{ m} \leq x < 31.1 \text{ m}$) Moderate 4 ($31.1 \text{ m} \leq x < 41.8 \text{ m}$) Moderate 5 ($41.8 \text{ m} \leq x < 61 \text{ m}$) Deep ($\geq 61 \text{ m}$) |
| Rainfall – Consecutive Number of Wet Days (from White et al., 2023b) | Maximum number of consecutive wet days | 0 – 18 days | |
| Rainfall – Number of Dry-Wet Cycles (from White et al., 2023b) | Count of number of wet to dry day transitions | 0 – 24 | |
| LULC <i>E. coli</i> Contamination Potential Classes (from White et al., 2023a) | Refers to how human activities and natural elements on the land may impact <i>E. coli</i> contamination in wells | 1 – 4 | Very High (1) High (2) Moderate (3) Low (4) |
| SWE – Maximum Number of Consecutive Melt Days (derived for current analyses) | Maximum number of consecutive melt days | 5-days: 0 – 6 days 36-days: 0 – 37 days | |
| SWE – Cumulative Melt During Maximum Consecutive Melt Day Period (derived for current analyses) | Sum of daily melt over period of maximum consecutive melt days over defined antecedent time period | 5-days: 0 – 130.8mm 36-days: 0 – 238.7mm | |

| | | | |
|--|--|---|--|
| SWE – Single Day Maximum Melt (derived for current analyses) | Maximum value of a single day melt event over the defined antecedent time period | 5-days: 0 – 129mm 36-days: 0 – 158.2mm | |
| SWE –Number of Melt Days (derived for current analyses) | Count of melt days over the defined antecedent time period | 5-days : 0 – 6 days 36-days: 0 – 37 days | |
| SWE –Total Melt (derived for current analyses) | Sum of daily melt over defined antecedent time period | 5-days: 0 – 130.8mm 36-days: 0 – 238.7mm | |
| Total Water (derived for current analyses) | Sum of rainfall and snow melt | 5-days: 0 – 158.8mm 36-days: 0 – 336.9mm | |
| Well Density (derived for current analyses) | Assigns density value to tile based on number of wells contained within, and neighbouring | 0 –366 wells/km ² | Category 1 ($1 \leq x < 6$) Category 2 ($6 \leq x < 12$) Category 3 ($12 \leq x < 24$) Category 4 ($x \geq 24$) |
| Flow Accumulation¹ (derived for current analyses) | Accumulated flow into a tile based on surrounding elevations and resulting flow directions | 0 - 68160808 | Category 1 ($0 \leq x < 1$) Category 2 ($1 \leq x < 10$) Category 3 ($10 \leq x < 100$) Category 4 ($x \geq 100$) |
| Drilling Method (derived for current analyses) | Method used to drill well during original construction | 1-4 | Category 1 (Boring, Digging, Cable Tool) Category 2 (Other Methods, Rotary Conventional, Rotary Reverse) |

| | | | |
|---|--|----------|--|
| | | | Category 3 (Jetting, Diamond, Rotary Air) Category 4 (Air Percussion, Driving) |
| Casing Diameter (from White et al., 2021) | Diameter of well casing at surface | 0 – 6.4m | |
| Bottom Casing Material (derived for current analyses) | Material used for well's casing at the bottom of well | 1-4 | Category 1 - Potential Cracking (Plastic, Concrete, Fiberglass) Category 2 - Open Hole Category 3 - Potential Corrosion (Galvanized Steel, Steel) Category 4 - Stainless Steel |
| Testing Frequency (from White et al., 2021) | Number of times user tested well over 8-year study period | 1 – 446 | |
| Testing Ratio (derived for current analyses) | Within a 1km by 1km grid, ratio of tested to non-tested domestic wells | 0-1 | |
| Testing Hotspot (derived for current analyses) | Statistically significant hot and cold spots on a map, weighted using testing ratios | -3 - 3 | 99 th Percentile Hotspot (3) 95 th Percentile Hotspot (2) 90 th Percentile Hotspot (1) Not Significant (0) 90 th Percentile Coldspot (-1) 95 th Percentile Coldspot (-2) |

| 99 th Percentile Coldspot (-3) | | |
|---|---|--------------------|
| Closest Drop-off Location – Hours/Month (derived for current analyses) | Number of hours per month the closest drop-off location is open for drop-offs | 0 – 208 hrs |
| Closest Drop-off Location – Days/Month (derived for current analyses) | Number of days per month the closest drop-off location is open for drop-offs | 0 – 20 days |
| Closest Drop-off Location – Open Evenings (derived for current analyses) | Whether the closest drop-off location is open for drop-offs in the evenings (after 6pm) | 0,1 |
| Closest Drop-off Location – Open Lunches (derived for current analyses) | Whether the closest drop-off location is open for drop-offs over lunch | 0,1 |
| Closest Drop-off Location – Seasonal Hours (derived for current analyses) | Whether the closest drop-off location has seasonal—based hours | 0,1 |
| Closest Drop-Off Location Driving Distance (derived for current analyses) | Closest location based on driving distance (km) | 0 – 449.0 km |
| Closest Drop-Off Location Driving Time (derived for current analyses) | Closest location based on driving time (min) | 0 - 757.7 min |
| Month of Test (from White et al., 2021) | Month water sample was submitted | 1 (Jan) - 12 (Dec) |
| Year of Test (from White et al., 2021) | Year water sample was submitted | 2010 - 2017 |

| | | | |
|--|--|-------|---|
| First Test Result Status (from White et al., 2021) | Returned testing message of first water sample submission | 1 - 4 | MR - “no significant evidence” (1) MR - “no result” (2) MR - “may be unsafe” (3) MR - “unsafe to drink” (4) |
|--|--|-------|---|

¹Continuous data range value represents the number of 30mx30m tiles that flow into a given tile

5.3.2.1 Seasonal Transition Identification

As precipitation falling as snow rather than rain contributes differently to *E. coli* contamination in groundwater (Invik et al., 2019), and latitude is a driver of seasonal timings, which in turn is a driver of ground permeability (i.e., frozen, dry, saturated) (White et al., 2021), this work moves beyond a fixed definition of seasons to one that incorporates latitudinal effects. The Canadian Plant Hardiness Zones (Ontario contains 0b to 7a) were used as a starting point for a more flexible delineation of seasons. These zones were originally established to guide plant choices developed using a combination of temperature, precipitation, wind speed, and snow depth (McKenney et al., 2014). Within each plant hardiness zone, the onset and cessation of winter were delineated based on SWE.

To use SWE for the determination of seasonal transitions, a dataset was developed that associates SWE (pulled from Daymet) with spatial extents of each Plant Hardiness Zone. This was achieved through a multi-step process. A 2km x 2km grid was superimposed over the Daymet data (1x1km grid cells) for Ontario, such that each 2 x 2 km cell includes four Daymet grid cells. A systematic sampling technique was then employed to develop a location dataset by extracting the location of the centre of the NW Daymet grid cell within each 2 x 2 km grid cell. Daily SWE data were then extracted from the Daymet database for each location in the location dataset across the years of study (January 1, 2010 to December 31, 2017). Absolute values of snow water equivalent (SWE) were then converted into changes in SWE (ΔSWE) calculated as:

$$\Delta SWE = SWE_{t+1} - SWE_t$$

where t is the day of the SWE measurement. A positive ΔSWE represents snow accumulation and a negative ΔSWE represents snowmelt.

To identify the onset of winter and spring within a particular zone, all occurrences of snow accumulation (positive ΔSWE) or snowmelt (negative ΔSWE) (y-axis), respectively, were plotted against their respective month and day (x-axis) in separate boxplots. The onset of winter was defined as the date representing the 75th percentile of SWE accumulation. Similarly, the onset of spring was defined as the date representing the 25th percentile of SWE melt. The onset of summer and autumn were determined using the Ontario climate zones, with the average date of last spring frost as the threshold for summer onset, and average date of first fall frost as the threshold for autumn onset (OMAFRA, 2020; PlantMaps, 2022) (Figure 5-2 and Table A.5-1).

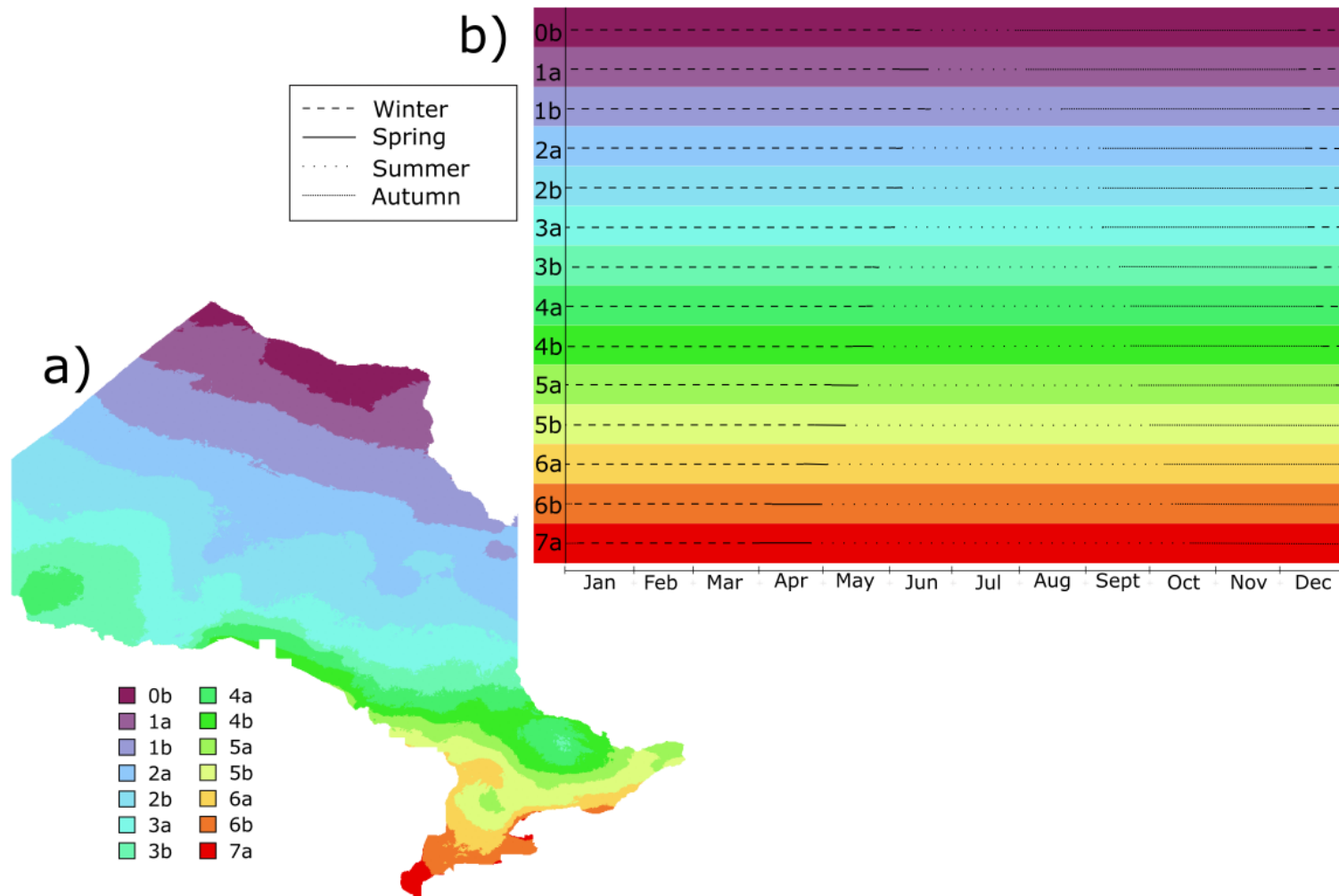


Figure 5-2: Summary of Ontario seasons (b) based on spatial distributions of Hardiness Zones (a; derived by McKenney et al. (2014)).

5.3.2.2 *Physical Model Variables*

While some variables have already been identified as important with respect to well contamination in the literature through previous analyses of this dataset (i.e., *E. coli* observation, location, date of observation, geological formation, specific capacity, well depth, rainfall, LULC, critical antecedent periods) (White et al., 2023a, 2023b, 2021), others are being explored for the first time in the context of this private well dataset (i.e., Δ SWE, well density, flow accumulation, drilling method, casing material). Additionally, this is the first time that all variables have been explored simultaneously and in conjunction with human system variables (section 5.3.3) with the goal of explaining *E. coli* contamination events.

Seasonality represents the different hydroclimatic processes that are prevalent in the different seasons. For the purpose of this work, it is assumed that precipitation in winter will only fall in frozen form (i.e., Δ SWE), precipitation in summer will only fall in liquid form (i.e., rainfall), and precipitation in spring and autumn (“shoulder” model) will be mixed (i.e., Δ SWE and rainfall). Precipitation in general is an important driver of private well contamination. Precipitation drives infiltration, which transports *E. coli* from the surface, through the subsurface, into a well. For infiltration to occur, there must be a connected path of liquid water through the soil. However, the infiltration rate is dependent on antecedent precipitation and temperature conditions, as these impact the degree of saturated pore connectivity through saturation and groundwater phases (i.e., solid, liquid), respectively. As such, rainfall and snowmelt intermittency are explored in this work. White

et al. (2023b) identified two critical antecedent periods for contamination using rainfall data. The first represents the best explanatory power (36-days) and the second represents the best trade-off between computational requirements and explanatory power (5-days). Given these findings, the same antecedent periods were applied to Δ SWE in the current analyses. Similar to rainfall intermittency (White et al. 2023b), SWE variables considered include maximum number of consecutive melt days, cumulative melt during max consecutive melt day period, single day maximum melt, total number of melt days, and total melt. For the spring and autumn seasons, total water represents the summation of rainfall and snow melt as a measure of water that is available for infiltration.

Flow accumulation is introduced to assess whether there could be water pooling around a well or if water is diverted away by the natural topography. Flow accumulation is calculated using the flow direction and flow accumulation functions and the multiple flow direction (MFD) algorithm in the ArcGIS Pro hydrology toolset. The MFD algorithm partitions flow, proportional to slope, from one cell to all downslope neighbors (Qin et al., 2007). Elevation data from the Provincial Digital Elevation Model (PDEM) (Ontario Ministry of Natural Resources and Forestry, 2019), are used as an initial input to determine the flow direction of surface runoff. The resulting raster file (30m x 30m) is then processed with the flow accumulation function, calculating the accumulated weight of all cells flowing into each downslope cell. Each calculated cell value represents the number of cells that flow into it. For this work, flow accumulation was also divided into categories where

0-1 represents no/low accumulation, 1-10 represents low/moderate accumulation, 10-100 moderate/high accumulation, and 100+ high/very high accumulation (Table 5-2).

Well density is introduced to explore the possibility of preferential flow paths that can be introduced either between the surface and subsurface along the well bore, or between wells due to disturbances in the subsurface (e.g., fracking). Well density was determined using the kernel density function in the ArcGIS Pro density toolset, which calculates the density of point features around each output raster cell (applied at the center of each cell). Kernel density conceptually works by fitting a smoothly curved surface over each point (i.e., well location), with a two-kilometre search radius extending from each input point, known as the kernel surface (ESRI, 2022). As all well locations are weighted equally, each kernel surface has a value of one, where the highest surface value is at the point (well location) and reaches zero at the search radius (2 km). Kernel densities are then calculated based on a 1km x1km raster grid by summing the kernel surfaces that overlay the raster cell centre (ESRI, 2022). Calculated kernel densities are outputted in the form of a 1km-by-1km raster dataset. The resulting kernel densities spanned a large range (0-366), so categorical ranges were created ensuring an equal distribution of data in each range (Table 5-2).

Well drilling methods and well casing material data were obtained from the WWIS database and assigned to unique wells. Regression techniques were used to rank the drilling method and bottom of well casing material based on their *E. coli* contamination potential, following the same method as White et al. (2023a), where *E. coli* concentration was used

as the dependent variable and drilling method or bottom of well casing material used as independent variables in two different models. Casing materials and drilling methods were then ranked based on largest to smallest negative impact to *E. coli* concentration. Using this ranking system, combined with the number of observations in the private well dataset (i.e., with the goal of creating similar bin sizes), methods and materials were binned into classes from one to four, where one represents greater *E. coli* contamination potential (Table 5-2).

5.3.2.3 Human Behavior Model Variables

While testing frequency within the private well dataset was explored in prior studies (Latchmore et al., 2020; White et al., 2021), testing ratios, testing hotspots, drop-off location proximity and accessibility have not been explored to date. Ontario was divided into 1km x 1km grid cells, and well locations from the private well dataset (i.e., wells that have been tested at least once) and all domestic wells from the WWIS (i.e., locations of drilled wells regardless of testing history) were plotted and summarized within each cell. The testing ratio was calculated by taking the number of wells from the private well dataset (tested at least once over the 8-year period) divided by the total number of domestic wells located in the cell. To reduce skewing of the data, cells were only considered to have a robust ratio value if there were three or more domestic wells present. With gridded testing ratios determined, unique domestic wells were assigned a ratio value from their corresponding cell. Further, the optimized hotspot analysis from the Mapping Clusters Toolset in ArcGIS Pro was utilized to identify the hotspots (higher testing ratios) and

coldspots (lower testing ratios) of testing in Ontario. This technique not only uses the testing ratio values to identify clustering, but also assigns a confidence rating to each.

Within Ontario, there are many locations that accept water sample drop-offs for Public Health Ontario's free water sampling initiative. To identify all possible drop-off locations, each public health unit's website was searched (34 units) for mention of private well testing. Across the 34 public health unit websites, 178 drop-off locations were clearly identified by these websites. The position of all 178 locations were identified, while hours were found for 168 drop-off locations. Closest drop-off locations for each domestic well were identified using the Origin-Destination (OD) Cost Matrix tool from the Network Analyst Toolset in ArcGIS Pro. For this analysis a Roads Network (DMTI Spatial Inc., 2021) was acquired from DMTI Spatial Inc., containing data on not only the road network itself but also the posted speeds. For this analysis, drop-off locations were labelled as destinations and domestic well locations were labelled as origins. The tool then calculated the driving distance (km) and time (min) it would take for a well user to reach the closest drop-off location. Finally, the hours/month, days/month, evening availability, lunch availability, and seasonal operation details of the nearest drop-off locations were assigned to each well.

Bivariate analyses were conducted comparing driving time to nearest drop-off location to testing frequency, hotspot occurrence, coldspot occurrence, day of week, year and month of test, and first test result status (i.e., no significant evidence of *E. coli*, no result, may be unsafe to drink, or unsafe to drink - *E. coli* detected in the sample). Driving

times were aggregated into groups of specific minimum driving times in 5 or 10 minute increments (e.g., a group labelled 10 minutes represents wells with a single direction driving time of greater than 10 minutes but less than the threshold of the next bin). Driving times less than 30 minutes and longer than 70 minutes were binned into 10 minute time steps, while driving times between 30 and 70 minutes were binned into 5 minute time steps to capture more nuanced trends.

5.3.3 Statistical Analyses

Regression-based supervised machine learning approaches were used to explore the relationships between dependent variables and several independent variables. A series of different techniques were employed to increase the robustness of explanatory findings. Methods used in this work include Multivariate Adaptive Regression Spline (MARS), Generalized Additive Model for Location, Scale, and Shape (GAMLSS), and Least Absolute Shrinkage and Selection Operator (LASSO). Generally, MARS was selected due to its ability to allow high degree interactions between independent variables to attempt to predict the dependent variable (Friedman, 1991), GAMLSS was selected due to its ability to deal with highly skewed and kurtotic distributions (Stasinopoulos and Rigby, 2007), and LASSO was selected due to its natural ability for variable selection and its greater prediction accuracy due to its use of an L1 penalty term (Ranstam and Cook, 2018). All three methods have been successfully used in a variety of groundwater-based applications, inclusive of the dependent variables of interest in this study (White et al., 2023b).

Based on previous research indicating the significance of seasonality as a driver of *E. coli* transport in the subsurface (e.g., Atherholt et al., 2017; White et al., 2021) and recognizing that variables linked to seasonality (such as snowmelt and rainfall) should only be considered in models for seasons in which these types of precipitation occur, explanatory regression models were developed seasonally (i.e., “summer”, “shoulder”, “winter”, “all seasons”). This was accomplished through a two-stage process, with the first stage being exploratory in nature and examining diverse physical and human dependent and independent variable sets. The second stage involved building informed, seasonal-based coupled-systems explanatory models, utilizing the findings from the first stage (Figure 5-3).

Five dependent variables were employed in the first stage, including *E. coli* presence (including non-detects), *E. coli* severity (restricted to adverse *E. coli* observations only), testing hotspot, testing coldspot, and testing frequency. With respect to the *E. coli* dependent variables, presence indicates the likelihood of *E. coli* occurrence in a well, while severity indicates the level of *E. coli* contamination in a well. The *E. coli*-based dependent variables were utilized to construct exploratory models based on well characteristics, location, and weather. The testing hotspot, testing coldspot, and testing frequency dependent variables were utilized to develop testing drop-off location-based exploratory models. Figure 5-3 shows the variable subsets included in each exploratory model.

In the second stage, the only dependent variables explored were the *E. coli*-based variables (i.e., presence and severity). Important independent variables identified in the

first stage (i.e., exploratory models), along with independent variables identified as important in previous studies (White et al., 2023a, 2023b, 2021), were then used to develop the seasonally driven informed coupled-systems models (Figure 5-3). As mentioned previously, these models were developed seasonally to capture the relevant form(s) of precipitation (i.e., rainfall, snowmelt).

MARS, GAMLSS, and LASSO regression were used for variable selection in both the exploratory and explanatory coupled-systems stages. Once important variables were identified, confounding variables were removed, and model parameters were fine-tuned (i.e., MARS tuned via degree of interactions and iterations; LASSO tuned via shrinkage penalty or lambda value; and GAMLS tuned via fitting families), training and testing were conducted on 10 unique data splits (80% training and 20% testing) (White et al., 2023b). Independent variables were deemed insignificant if two of three regression methods agreed on its removal. Model performances were compared using the coefficient of determination (R^2) and root mean squared error.

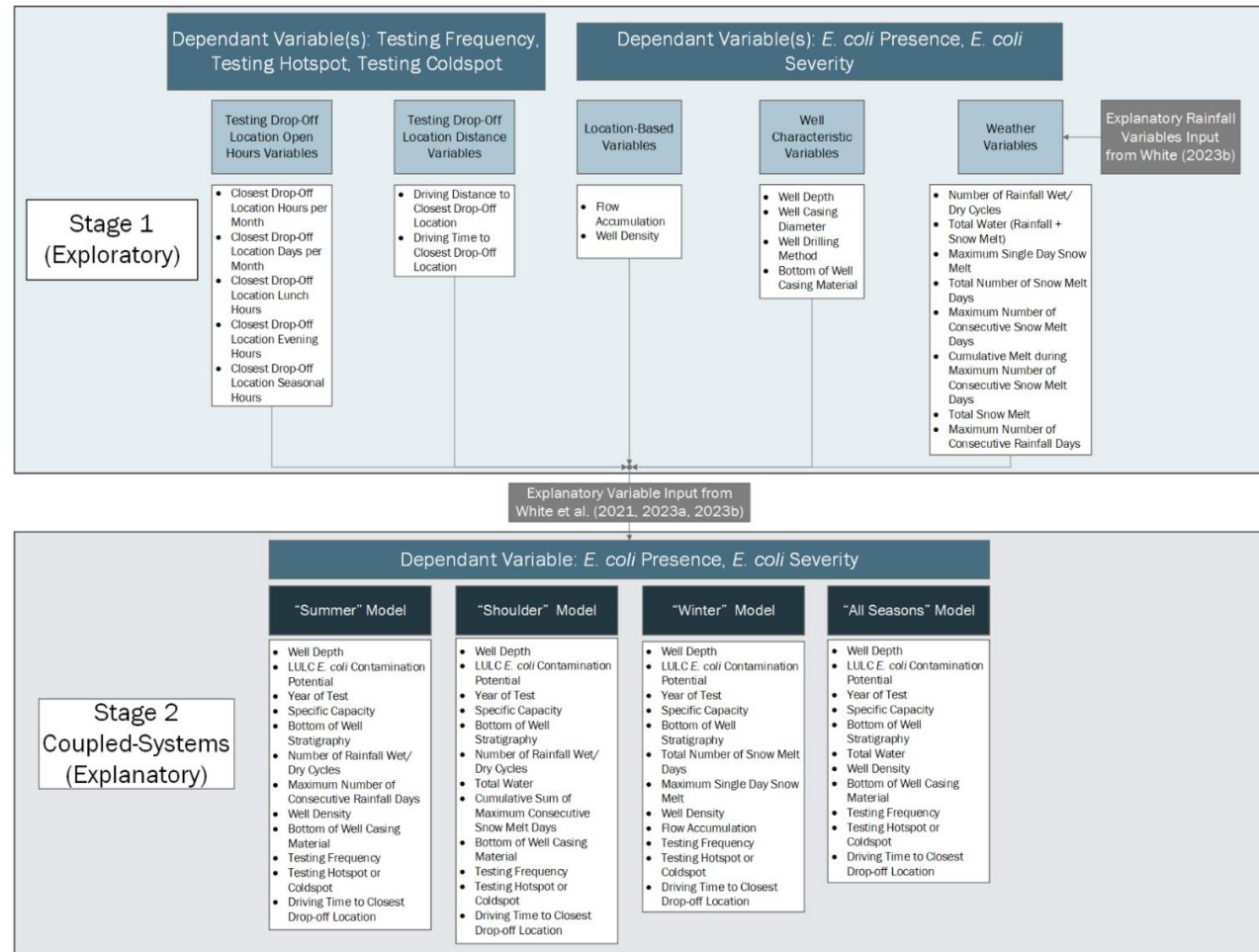


Figure 5-3: Overview of dependent and independent variable subsets explored in this work.

5.4 RESULTS AND DISCUSSION

The two-stage analysis employed utilized the exploratory analyses (Section 5.4.1) for variable selection which strengthened findings by informing variable choice in each of the seasonal (i.e., “summer”, “shoulder”, “winter”, “all seasons”) explanatory coupled-systems analyses (Section 5.4.2) (Figure 5-3).

5.4.1 Exploratory Analyses

The exploratory analyses were initially undertaken using the weather-based variables, followed by those using the well characteristic, location-based, and human behaviour variable sets (Figure 5-4).

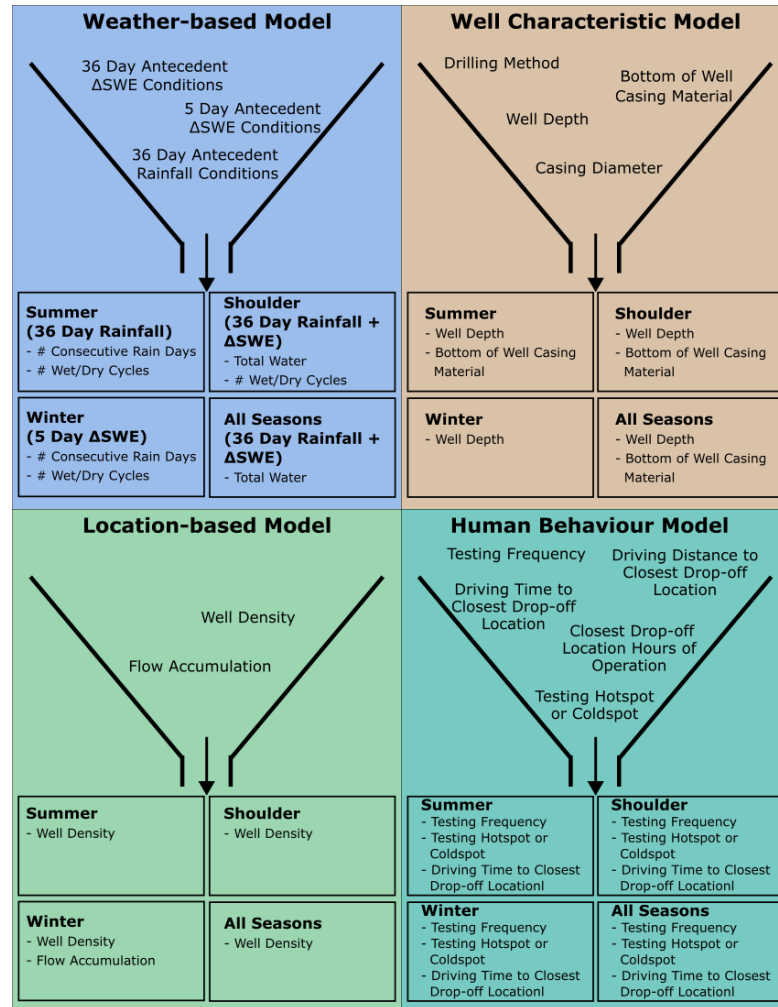


Figure 5-4: Summary of exploratory variable sets, identifying individual variables explored and variables that were deemed important in the informed seasonal coupled-systems models.

5.4.1.1 Weather-based Model

The current analyses are constrained to the shoulder and winter seasons where snow melt occurs, as summer precipitation events have been thoroughly explored in White et al. (2023b). The key findings from White et al. (2023b) are brought straight into the explanatory model (Section 5.4.2). Specific findings from these previous analyses include:

- a 36-day antecedent period that was most explanatory for contamination events;
- the

number of consecutive wet days and number of wet/dry cycles (Figure 5-4); iii) zero wet days in the 36-day antecedent period that significantly reduce the likelihood of *E. coli* contamination; and iv) an increased number of consecutive wet days or wet/dry cycles that result in an increased likelihood of *E. coli* presence, but at lower concentrations, in wells.

Snowmelt conditions using the 5- and 36-day antecedent periods were explored in separate “shoulder” and “winter” models. For the shoulder seasons, models using the dependent variables “all *E. coli* observations” and “adverse *E. coli* observations” found the 36-day antecedent period variable outperformed the 5-day antecedent variable in GAMLSS and LASSO. This is likely due to the fact that there is mixed precipitation during the shoulder seasons, combined with snowpack melt and flow pathways due to warmer air temperatures and thawing soil matrix. Thus, the 36-day antecedent period variable was employed in the remaining shoulder season exploration models. Ultimately, the best variable combination (all variables deemed important by GAMLSS, LASSO, and MARS) carried through to the explanatory model was total water equivalent (rainfall + snow melt water equivalent), number of rainfall wet/dry cycles, and cumulative melt during maximum number of consecutive snow melt days (Figure 5-4). The models showed that increasing amounts of total water equivalent up to a value of ~157mm results in increased *E. coli* concentrations; however, *E. coli* concentrations plateaued past this threshold. Increasing both rainfall wet/dry cycles and cumulative melt during maximum number of consecutive snow melt days resulted in decreases in *E. coli* concentrations. This relationship is likely

due to the fact that more pulses and larger volumes of water will dilute the *E. coli* in the system.

For the “winter” models using the dependent variables “all *E. coli* observations” and “adverse *E. coli* observations”, the 5-day antecedent period variable outperformed the 36-day with respect to explanatory power in both GAMLSS and LASSO. This is likely due to the presence of frozen ground, which would decrease infiltration, and therefore *E. coli* transport opportunities. Winter temperatures in Ontario are typically below freezing, with occasional melt events (OASDI, 2023). When melt events occur, their duration is typically short. Thus, any *E. coli* that is mobilised may quickly travel to and contaminate a well (i.e., within 5 days) or re-freeze, becoming immobilised again. This could also be associated with increased inactivation or die-off rates of *E. coli* in the colder winter temperatures, with longer lag times presenting less of a concern. Thus, the 5-day antecedent period variable was employed in the remaining “winter” exploration models. The best variable combination that was carried through to the exploratory model was determined to include maximum single day snow melt (GAMLSS, LASSO) and total number of snow melt days (GAMLSS, LASSO, MARS) (Figure 5-4). Increasing maximum single day snow melt was found to increase *E. coli* concentrations, supporting the hypothesis that snow melt is a transport mechanism for *E. coli* (Staley et al., 2017). While increasing the total number of snow melt days increases *E. coli* presence likelihood, it decreases *E. coli* concentration, likely due to the dilution effect (White et al., 2023a).

An exploratory “all seasons” model was not developed using weather-based variables, as the only variable that is physically relevant throughout the year is total water equivalent. As such, total water equivalent was the only weather-based variable carried forward to the explanatory stage 2 coupled-systems “all seasons” model.

5.4.1.2 *Well Characteristics Model*

Drilling method and bottom of well casing material were first explored in an “all seasons” model to identify meaningful categories for these variables (Figure 5-4). As well, drilling methods are typically associated with particular ranges for both depth and diameter, it is likely that these variables will be confounding. As such, exploratory analyses only ever included one of these variables in the analyses at any given time. First, regression models were used to group similar drilling methods based on their association with *E. coli* concentrations. Four categories emerged, with category 1 being associated with the highest potential for *E. coli* contamination and category 4 with the lowest (Table 5-2). However, these findings do not align with the current understanding in the literature. For example, both driving and jetting drilling methods were identified as being associated with moderate *E. coli* contamination potential, yet these methods are typically used to drill relatively shallow wells. This contradicts the literature, which typically associates shallow wells with higher *E. coli* contamination potential (e.g., Allen et al., 2019; Hynds et al., 2014, Kreutzweiser et al., 2010). Since the well drilling method categories do not align with the current physical understanding of the system, well depth and casing diameter were carried

forward as independent well characteristic variables in the exploratory well characteristic models.

Similarly, regression analyses were used to categorize bottom of well casing materials based on their association with *E. coli* contamination potential. Bottom of well casings were grouped into four categories, category 1 being associated with the highest potential for *E. coli* contamination and category 4 with the lowest (Table 5-2). The very high *E. coli* contamination potential class (category 1, n = 47,199) includes casings made of fiberglass, concrete, and plastic. This can be explained by risk of hydraulic collapse, cracking, and difficulty in sealing joints, increasing potential for contaminant transport highways (Ontario Ministry of Environment Conservation and Parks, 2019; Rauf and Amani, 2017). The high *E. coli* contamination potential class (category 2, n = 384,157) consists of open hole (i.e., no casing), explained by their presence in fractured rock (or other high yield material), which again represents potential contaminant transport highways (White et al., 2021). The moderate *E. coli* contamination potential class (category 3, n = 332,860) includes galvanized steel and steel. Contamination ingress in these wells is associated with their corrosion potential in certain environments that increase the likelihood of unexpected groundwater – and therefore contaminant - ingress (Ontario Ministry of Environment Conservation and Parks, 2019). Finally, the low *E. coli* contamination potential class (category 4, n = 54) consists of stainless steel, which is less likely to corrode and crack, which decreases potential contamination. However, it can be

expensive, which could explain its low usage (Ontario Ministry of Environment Conservation and Parks, 2019).

Once relevant well characteristic variables were identified, individual seasonal models were explored (Figure 5-4). Based on findings, well depth was deemed important for “shoulder” (GAMLSS, MARS), “summer” (GAMLSS, LASSO), “winter” (GAMLSS, MARS), and “all seasons” models (GAMLSS, MARS), with a consistent trend of increasing well depth associated with decreasing *E. coli* concentration in wells. Bottom of well casing material was found to be important for “shoulder” (GAMLSS, LASSO), “summer” (LASSO, MARS), and “all seasons” (LASSO, MARS) models. Generally, very high (cracking potential) and high (no casing) well casing *E. coli* contamination potential classes were found to increase concentrations, while moderate and low classes were found to be protective. Well casing diameter was not deemed explanatory in any of the models.

5.4.1.3 *Location-based Model*

Well density, as a surrogate for aquifer vulnerability, and flow accumulation, as a surrogate for pooling or high overland flow zones, were explored (Figure 5-4). Well density was found to be explanatory in the “summer” (GAMLSS, LASSO, MARS), “winter” (GAMLSS, LASSO), and “all seasons” (GAMLSS, LASSO, MARS) models, where increasing density was associated with decreasing *E. coli* presence and concentrations. As noted above, well bores increase the risk of hydraulic connections between the aquifer and ground surface, and therefore the potential for aquifer contamination increases with the number of wells. Additionally, well clusters increase the likelihood of *E. coli* transport

between wells in the same aquifer (Niagara Peninsula Conservation Authority, 2011). As such, the relationship identified between density and *E. coli* concentration was counter to the expectation. The location of well clusters within lower contamination potential LULC classes was considered as a potential explanation. However, it was found that a majority of the category 3 and 4 well densities are located in LULC classes associated with very high and high *E. coli* contamination potentials (Table 5-3). As such, this does not explain the association between decreasing *E. coli* with increasing well density. Another potential explanation is that high well densities are located in communities that are heavily reliant on groundwater, which may have programs in place to support better well stewardship (e.g., education, source water protection, etc.). However, more data are required to explore this postulation, and thus the association between decreasing *E. coli* with increasing well density also requires further exploration.

Table 5-3: Comparison of well density categories and LULC *E. coli* contamination classes.

| | | LULC | | | |
|---------------------|-------------------------------------|-----------|--------|----------|-------|
| | | Very High | High | Moderate | Low |
| Well Density | Category 1 (0-6 wells) | 12.79% | 5.82% | 10.59% | 2.50% |
| | Category 2 (6-12 wells) | 14.77% | 9.48% | 6.31% | 2.51% |
| | Category 3 (12-24 wells) | 10.55% | 11.16% | 3.65% | 1.87% |
| | Category 4 (24+ wells) | 1.58% | 5.35% | 0.72% | 0.34% |

Flow accumulation was found to only be explanatory in “winter” models (GAMLSS, LASSO, MARS). Increasing flow accumulation was associated with decreased *E. coli* concentrations up to values of 10 (i.e., 10 cells flowing into target cell) (categories 1 and 2); however, *E. coli* concentrations do not decrease any further for flow accumulation values beyond 10 (categories 3 and 4). The pooling of water in the winter season protects against surface water ingress through the freezing of saturated ground, reducing permeability and therefore infiltration (Niu and Yang, 2006). The threshold of 10 likely represents a level of saturation, which when frozen, reduces permeability to near zero. Flow accumulation thresholds have not, to the authors’ knowledge, been explored in the context of well contamination before. The proposed threshold of 10, identified in these analyses, supports literature suggesting that some pooling of water is required in the winter for protective attributes to occur.

5.4.1.4 Human Behaviour Model

Human testing behaviours are important in understanding well user’s knowledge, attitudes, and practices which translates into well condition and treatment practices (Di Pelino et al., 2019; Lavallee et al., 2021). It has already been identified that, of the wells tested at least once in the 8-year study period, only two percent of wells were tested at or more than the recommended frequency of twice per year (White et al., 2021). To better understand convenience-based barriers that may be preventing well users from testing consistently, driving time, driving distance, and hours of operation (including days per month open, hours per month open, occurrence of seasonal hours (i.e., closed for part of

the year), lunch closures, and evening availability (after 6pm) for the closest drop-off location were explored.

The first set of models explored driving time and distance independently. As driving time and driving distance are highly correlated, models were created to explore each of these variables independently. It was determined that driving time outperformed driving distance when explaining testing frequency and testing hotspots/coldspots (GAMLSS, MARS). The next set of models explored the driving time and hours of operation for the nearest testing drop-off location. Generally, regression models did not find consistent or strong (i.e., small coefficient) trends for the impact of hours of operation on testing behaviour. However, one notable finding is that evening hours of operation were associated with increased likelihood to test (from 0 tests to 1 test) but did not drive testing frequency (more than 1 test). This may mean that evening availability increases convenience for one-off testing but does not incentivise consistent testing. Another consistent trend is that seasonal hours (e.g., March to October open every Tuesday, November to February open last Tuesday of every month) were associated with decreased testing frequency and increased testing coldspots. This observation suggests that it is better for a drop-off location to have consistent, predictable hours year-round than improved hours for part of the year.

Regression models found that of the drop-off location convenience-related variables, the largest driver of testing behaviour was the driving time to the closest drop-off location. In LASSO and GAMLSS models there was a consistent trend of increased driving time associated with decreasing testing frequency and increasing coldspot confidence. Further,

when exploring the impact to testing coldspots, MARS models identified a threshold around the 70-minute point (one way), beyond which coldspot confidence started to decrease slightly. The identification of this threshold prompted additional exploration of the data to uncover potential reasons behind this trend.

Bivariate analyses were conducted comparing discrete bins of minimum driving times with testing frequency, hotspot occurrence, coldspot occurrence, day of week the sample was collected, as well as previously identified variables including the first test result status, year of test, and month of test (White et al., 2021). Analyses of testing frequency, and hot or coldspot occurrences based on minimum driving distance thresholds identified a change in testing behaviour around the 45-50 minute mark (Figure 5-5). Specifically, increasing minimum driving time is associated with increased numbers of wells that have not tested during the study period up to 45-50 minutes travel time to a drop off location (Figure 5-5a), likely due to decreasing convenience. For travel times longer than 45-50 minutes the “no test” curve flattens, possibly representing a convenience versus habit threshold in peoples’ behaviour. Supporting trends are found in the hotspot (Figure 5-5b) and coldspot (Figure 5-5c) analyses. For testing hotspots, trends remain consistent between 0- and 40-minutes one-way minimum travel time, past which point hotspot confidence decreases until a plateau starts to emerge around the 50-minute mark. For testing coldspots, confidence begins to increase sooner, at the 30-minute mark, and similarly plateaus around the 50-minute mark. This strongly suggests that a well user’s perception of convenience is a drop-off location closer than 35-minutes (one way) driving time.

When considering driving time to the nearest drop-off location in the context of the first test result status, it was hypothesised that longer minimum driving times would be associated with a greater number of adverse (“unsafe” or “may be unsafe”) sample results, which would incentivise well users to go out of their way to test. Instead, the distribution of adverse test results was similar across all driving distances (Figure A.5-1).

Travel time to the nearest sample drop-off location was then examined in the context of the year, month, and day of week that the sample was collected. Up to the 80-minute minimum one-way travel time, data distributions remained similar, past such point a greater distribution of tests occurred in 2017 (Figure A.5-2), a greater distribution of tests occurred between June and August (Figure A.5-3), and a slightly greater distribution of tests occurred on Tuesdays (Figure A.5-4). Increased testing between June and August and on Tuesdays may be associated with the common use of cottages in the summer months. Additionally, an increase in tests on Tuesdays, accounting for approximately 25% of tests, may also be driven by the number of long weekends cottage residents take in the summer months, with users dropping their water samples off for testing on their way back to their primary residence.

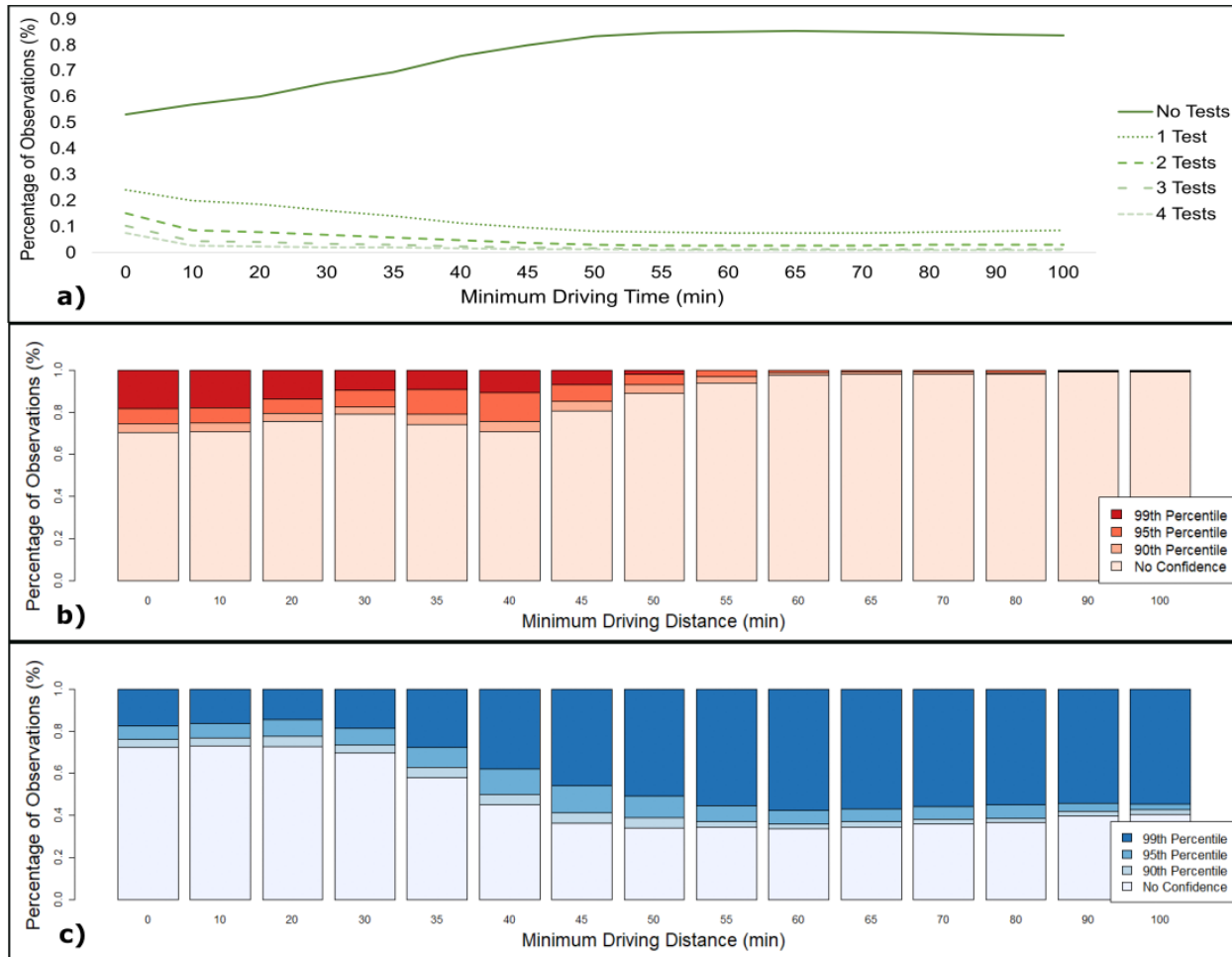


Figure 5-5: Summaries of trends between minimum driving times and testing frequency (a), testing hotspots (b), and testing coldspots (c).

5.4.2 Informed Coupled-systems Models

The informed coupled-systems models move towards a coupled-systems approach by combining the important human and physical variables identified in the exploratory variable sets (Section 5.4.1; Figure 5-4) with the goal of explaining *E. coli* presence and concentrations in private wells. This section is split into the different seasonal models (i.e., “summer”, “shoulder”, “winter”, “all seasons”) to capture the varying effects of location on fate and transport of *E. coli*.

5.4.2.1 “Summer” Model

For the “summer” model, variables that were deemed significant across all three regression methods (LASSO, GAMLSS, MARS) included LULC *E. coli* contamination potential and well density (Table 5-4; Table A.5-2). The most consistent variable is LULC *E. coli* contamination potential. Generally, *E. coli* presence and concentration decrease with decreasing *E. coli* contamination potential. More specifically, high *E. coli* concentrations are consistently associated with very high and high classes, decreasing for the transitions from high to moderate and moderate to low. This finding supports the concept that *E. coli* presence near a well is a strong driver for it to be present in the well. Further, there are key LULCs (i.e., agricultural/pastural, urban, aggregate mines, bare/openland) that require hypervigilance by well users in these locations.

Table 5-4: Summary of key trends identified from the informed coupled-systems regression analyses in the summer months.

| Variables Explored | Trends Identified | Level of Finding¹ |
|---------------------------|--------------------------|-------------------------------------|
|---------------------------|--------------------------|-------------------------------------|

| | | |
|--|--|--------------|
| LULC <i>E. coli</i> Contamination Potential | <ul style="list-style-type: none"> - Transitioning from Very High to Low results in decrease in <i>E. coli</i> presence and concentration - Very High and High have similar impact to <i>E. coli</i> | Primary |
| Well Density | <ul style="list-style-type: none"> - Categories 1, 2, and 3 increase <i>E. coli</i> concentration (Category 3 has highest increase to concentration) - Category 1 increases <i>E. coli</i> presence | Primary |
| Testing Frequency | <ul style="list-style-type: none"> - Increased testing frequency increases <i>E. coli</i> presence and decreases <i>E. coli</i> concentration | Secondary |
| Well Depth | <ul style="list-style-type: none"> - Increased well depth decreases <i>E. coli</i> presence and concentration | Secondary |
| Bottom of Well Stratigraphy | <ul style="list-style-type: none"> - Consolidated materials increase <i>E. coli</i> presence and decrease <i>E. coli</i> concentration | Secondary |
| Specific Capacity | NA | Inconclusive |
| Number of Rainfall Wet/Dry Cycles | NA | Inconclusive |
| Maximum Number of Consecutive Rainfall Days | NA | Inconclusive |
| Bottom of Well Casing Material | NA | Inconclusive |
| Testing Hotspot or Coldspot | NA | Inconclusive |
| Driving Time to Closest Drop-off Location | NA | Inconclusive |

¹ Primary represents significance in all regression models, secondary represents significance in two regression models, inconclusive represents significance in zero or one regression model and is not discussed due to lack of significance consensus

Well density was considered a surrogate for aquifer vulnerability, with higher density potentially representing greater aquifer vulnerability. While all models found well density explanatory, there was no clear trend in the association between density and *E. coli*

presence or concentration. Categories 1 ($n = 107,652$; <6 wells per 1km^2), 2 ($n = 94,609$; <12 wells per 1km^2), and 3 ($n = 100,408$; <24 wells per 1km^2) were all associated with increased *E. coli* presence and concentration in wells, with Category 3 being associated with the greatest increases in *E. coli* concentration. However, the highest well density, category 4 ($n = 97,860$; >24 wells per 1km^2), is associated with decreases in *E. coli* presence and concentration. This may represent a vulnerability threshold whereby an aquifer cannot become any more vulnerable once there is a density of 12+ wells per one km^2 . Category 1 (<6 wells per 1km^2) is the only well density category found to be associated with increased likelihood of *E. coli* presence. Another explanation could be that well users in an area with fewer wells are more isolated and may have less access to resources to improve well stewardship. A bivariate analysis comparing driving time to closest drop-off location and well density finds that there is a statistically significant weak inverse correlation (-0.26) between the two variables.

“Summer” models find that an increase in testing frequency is associated with an increase in presence of *E. coli* (MARS, LASSO). This likely results from the same population-testing trends observed during the COVID-19 pandemic – as number of overall tests increase, the number of positive tests will also increase (Chiu and Ndeffo-Mbah, 2021). Testing frequency is considered a surrogate measure for an individual’s KAP towards their well, with the general concept being that well users who test more frequently likely have a better understanding of the health risks posed by their well (Lavallee et al., 2023). Indeed, the “summer” models (MARS, LASSO) found that while *E. coli* presence

was found to increase with testing frequency, *E. coli* concentration was found to decrease, supporting the relationship between well testing frequency and well stewardship. Since it is likely that wells with higher testing frequencies were also tested in the shoulder and winter seasons, additional analyses were conducted to compare each unique season's testing frequency to explore the relationship between KAP and seasonal well users. When considering wells that were only tested once during the study period, 36.6% of these wells were tested in the summer, 32.9% in winter, and 29.7% in shoulder seasons. This may support a hypothesis that a higher proportion of summer well users are seasonal well users (e.g., cottage users) with lower KAPs towards their well. Wells tested in the shoulder and winter months were statistically significantly ($p < 0.0001$) more likely to be tested at higher frequencies than the summer months, suggesting more habitualised testing behaviour if wells are sampled outside the summer season. Annual testing over the study period would require a well to be tested at least eight times, while adhering to the recommended testing frequency of twice per year would require at least 16 tests during the study period. The analyses show that 16.8% and 15.5% of all wells with 8 or more tests included tests in the shoulder and winter seasons, respectively, while 5.5% and 5.2% of all wells with 16 or more tests included tests in the shoulder and winter seasons, respectively. In comparison, the number of wells with eight or more tests that included tests in the summer season fell to 11.8% and for 16 tests, it fell to 3.6%. These trends highlight the need for further KAP-based research into the decision-making processes driving well user testing.

Other trends held from previous investigations include a decrease in *E. coli* as well depth increases, and an increase in *E. coli* presence but decrease in concentration in wells finished in consolidated materials (Table 5-4) (White et al., 2021).

5.4.2.2 “Shoulder” Model

In the “shoulder” model, LULC *E. coli* contamination potential and bottom of well stratigraphy are deemed significant in all three regression models (Table 5-5). The same trend is found for LULC *E. coli* contamination potential as the “summer” model; there is an association between decreases in *E. coli* and the transition from very high to low contamination potential classes, while the very high and high classes have the same association with *E. coli* presence and concentration. Finally, there was found to be an association between increases in presence but decreases in concentration of *E. coli* with consolidated bottom of well stratigraphy. This supports previous findings that wells finished in consolidated materials have a higher likelihood of experiencing contamination events when compared to their unconsolidated counterparts (White et al., 2021).

Table 5-5: Summary of key trends identified from the informed coupled-systems regression analyses in the shoulder months.

| Variables Explored | Trends Identified | Level of Finding¹ |
|--|---|-------------------------------------|
| LULC <i>E. coli</i> Contamination Potential | <ul style="list-style-type: none"> - Transitioning from Very High to Low results in decreasing <i>E. coli</i> presence and concentration - Very High and High have similar impact to <i>E. coli</i> | Primary |
| Bottom of Well Stratigraphy | <ul style="list-style-type: none"> - Consolidated materials increase <i>E. coli</i> presence and decrease <i>E. coli</i> concentration | Primary |

| | | |
|---|--|--------------|
| Well Depth | - Increased well depth decreases <i>E. coli</i> presence and concentration | Secondary |
| Total Water | - Increased total water increases <i>E. coli</i> presence and concentration | Secondary |
| Cumulative Sum of Consecutive Snow Melt Days | - Increased cumulative sum decreases <i>E. coli</i> presence and concentration | Secondary |
| Driving Time to Closest Drop-off Location | NA | Inconclusive |
| Testing Frequency | NA | Inconclusive |
| Specific Capacity | NA | Inconclusive |
| Number of Rainfall Wet/Dry Cycles | NA | Inconclusive |
| Bottom of Well Casing Material | NA | Inconclusive |
| Testing Hotspot or Coldspot | NA | Inconclusive |

¹ Primary represents significance in all regression models, secondary represents significance in two regression models, inconclusive represents significance in zero or one regression model and is not discussed due to lack of significance consensus

Trends found in GAMLSS and LASSO identified a decrease in *E. coli* in association with increasing well depth, increased *E. coli* presence and concentration associated with increasing total water equivalent (i.e., rainfall + snow melt), and decreased *E. coli* presence and concentration associated with increasing cumulative melt during maximum consecutive melt day periods (Table 5-5; Table A.5-3).

5.4.2.3 “Winter” Model

For the “winter” model, there are no variables deemed significant in all three regression models. Variables deemed significant in two regression methods include

minimum driving time to nearest drop-off location, LULC *E. coli* contamination potential, testing frequency, and number of melt days (Table 5-6; Table A.5-4). Similar trends to both the “summer” and “shoulder” models were found for LULC *E. coli* contamination potential and testing frequency. Further, an increase in the number of melt days is found to increase the presence of *E. coli* but decrease the concentration. This supports findings in literature on the impact of the concentration and dilution effect (White et al., 2023b).

As driving time increases for sampling, *E. coli* presence and concentration decreases. Enhanced stewardship associated with habitual testing, is likely to explain the lower *E. coli* presence and concentration. This raises the question about the status of the relationship in each of the three unique seasons (i.e., summer, shoulder, winter), which is explored via a bivariate analysis. The bivariate analysis shows that of the three seasons, winter has the lowest *E. coli* concentration regardless of driving time, and that the concentration decreases with increasing driving time. This may be associated with a number of factors including a naturally lower presence and concentration of *E. coli* during the winter months, and an increase in the proportion of full time well users (due to the end of cottage season), representing a shift in user’s KAP – in particular whether they test their well out of habit or convenience.

To further explore the concept of habitual versus convenience testing across unique seasons, a bivariate analysis was conducted comparing driving time to closest drop-off location with testing frequency for wells that were tested at appropriate frequencies (i.e., ≥ 2 tests per year, ≥ 16 test over 8-year study period). Relationships presented as an

exponential decay function where slightly higher decay rates occurred in the summer and shoulder seasons (-0.383 ± 0.048 and -0.386 ± 0.052 , respectively) versus the winter season (-0.357 ± 0.054) as driving time increased. While the lower winter decay rate may indicate that well users test their wells out of habit during these months (i.e., driving time has less impact on testing), the seasonal slopes are not statistically significantly different. Increased minimum driving time may also be associated with a decrease in community involvement with well water testing programs and thus the potential for less information dissemination regarding health risks.

These findings, combined with the fact that wells tested during shoulder and winter seasons are more likely to be tested at a higher frequency, suggest that perhaps full time well users have more habitual testing patterns; however, more research is needed to strengthen this relationship.

Table 5-6: Summary of key trends identified from the informed coupled-systems regression analyses in the winter months.

| Variables Explored | Trends Identified | Level of Finding¹ |
|--|---|-------------------------------------|
| LULC <i>E. coli</i> Contamination Potential | <ul style="list-style-type: none"> - Transitioning from Very High to Low results in decreasing <i>E. coli</i> presence and concentration - Very High and High have similar impact to <i>E. coli</i> | Secondary |
| Driving Time to Closest Drop-off Location | <ul style="list-style-type: none"> - Increased driving time decreases <i>E. coli</i> presence and concentration | Secondary |
| Testing Frequency | <ul style="list-style-type: none"> - Increased testing frequency increases <i>E. coli</i> presence and decreases <i>E. coli</i> concentration | Secondary |

| | | |
|---------------------------------------|---|--------------|
| Total Number of Snow Melt Days | - Increased number of days increases <i>E. coli</i> presence and decreases <i>E. coli</i> concentration | Secondary |
| Bottom of Well Stratigraphy | NA | Inconclusive |
| Well Depth | NA | Inconclusive |
| Specific Capacity | NA | Inconclusive |
| Well Density | NA | Inconclusive |
| Maximum Single Day Snow Melt | NA | Inconclusive |
| Flow Accumulation | NA | Inconclusive |
| Testing Hotspot or Coldspot | NA | Inconclusive |

¹ Primary represents significance in all regression models, secondary represents significance in two regression models, inconclusive represents significance in zero or one regression model and is not discussed due to lack of significance consensus

5.4.2.4 “All Seasons” Model

When considering all seasons, variables deemed significant in all three regression models include LULC *E. coli* contamination potential and well density. The relationships with *E. coli* presence and concentration were the same as previous models (Table 5-7; Table A.5-5). Further, secondary findings, including driving time to closest drop-off location, testing frequency, total water, bottom of well stratigraphy, and well depth, also reflect trends previously identified in the “summer”, “shoulder”, and/or “winter” models. This absence of unique findings in the “all seasons” model, when compared to the “summer”, “shoulder”, and/or “winter” models, supports the value of moving towards the more nuanced geographically-driven seasonal models. However, the trends in this model cannot be completely discounted as they confirm that the general drivers of contamination are human influenced, namely where a well is located and when users are testing.

Table 5-7: Summary of key trends identified from the informed coupled-systems regression analyses in all months.

| Variables Explored | Trends Identified | Level of Finding¹ |
|--|--|-------------------------------------|
| LULC <i>E. coli</i> Contamination Potential | <ul style="list-style-type: none"> - Transitioning from Very High to Low results in decreasing <i>E. coli</i> presence and concentration - Very High and High have similar impact to <i>E. coli</i> | Primary |
| Well Density | <ul style="list-style-type: none"> - Categories 1, 2, and 3 increased <i>E. coli</i> concentration (Category 3 has highest increase to concentration) - Category 1 increases <i>E. coli</i> presence | Primary |
| Driving Time to Closest Drop-off Location | <ul style="list-style-type: none"> - Increased driving time decreases <i>E. coli</i> presence and concentration | Secondary |
| Testing Frequency | <ul style="list-style-type: none"> - Increased testing frequency increases <i>E. coli</i> presence and decreases <i>E. coli</i> concentration | Secondary |
| Total Water | <ul style="list-style-type: none"> - Increased total water increases <i>E. coli</i> presence and decreases <i>E. coli</i> concentration | Secondary |
| Bottom of Well Stratigraphy | <ul style="list-style-type: none"> - Consolidated materials decreases <i>E. coli</i> concentration | Secondary |
| Well Depth | <ul style="list-style-type: none"> - Increased well depth decreases <i>E. coli</i> presence and concentration | Secondary |
| Specific Capacity | NA | Inconclusive |
| Bottom of Well Casing Material | NA | Inconclusive |
| Testing Hotspot or Coldspot | NA | Inconclusive |

¹ Primary represents significance in all regression models, secondary represents significance in two regression models, inconclusive represents significance in zero or one regression model and is not discussed due to lack of significance consensus

5.5 STUDY LIMITATIONS

The fact that borehole logs are hand recorded in the field and later transcribed into an online database is a serious limitation of the WWIS database as it creates the opportunity for data entry errors, especially as some records date back to the 1910's (e.g., White et al., 2021). This was addressed through the removal of outliers that were outside the range of realistic values e.g., casing diameters less than zero. There is no indication that these outliers are systematic.

5.6 CONCLUSION

This study has introduced geographically driven seasonal delineations to explore *E. coli* presence and concentration in private wells in Ontario. Further, variables representing both the physical system and human behaviour were integrated into the data driven models as an initial step towards a coupled-systems approach, recognizing that human and natural systems interact. Key findings from the seasonal models include:

- Human behaviour-based variables appeared in most models, validating the need to use a coupled-systems approach;
- Driving time thresholds clearly exist, suggesting that some people test their water because the sample drop off location is convenient (<35 minutes away) and other people who test their water out of habit, and presumably awareness, rather than convenience (>50 minutes away);
- Drive time to closest drop-off location is relevant in the “winter” models. Increased distances are associated with decreased *E. coli* presence and

concentration which likely reflects seasonal contamination potential as well as more habitual rather than seasonal testers;

- LULC *E. coli* contamination potential has the greatest and most robust association with *E. coli* presence and concentration for all models, indicating that wells located in a very high or high contamination potential class are at a similar likelihood of experiencing contamination while those located in moderate and low classes are increasingly more protected from contamination;
- While increased testing frequency is associated with increased *E. coli* presence in the “summer” and “winter” models, it is also associated with decreased *E. coli* concentrations;
- Weather-based variables are only deemed significant in the “shoulder” and “winter” models, clearly demonstrating the concentration-dilution effect of melt days in the winter and shoulder seasons that were associated with decreased *E. coli* presence and concentration, versus the combination of rainfall and snow melt in the shoulder seasons that increased *E. coli*; and,
- Well characteristics emerged as significant in the “summer” and “shoulder” models. In both models, consistent trends are identified as increased well depth associated with decreasing *E. coli* contamination, and consolidated bottom of well stratigraphy associated with an increased likelihood of *E. coli* contamination.

This study demonstrates that applying regression-based machine learning techniques to a large, composite dataset, offers an innovative way forward to improve knowledge and understanding of private well contamination and stewardship for wells users and policy makers. This approach to a human-centered environmental problem will have a wide array of applications both geographically and conceptually with other similar environmental challenges.

5.7 WORKS CITED

- Allen, V., Edelsward, S., Murphy, A., Majury, A., Maki, A., 2019. Public Health Inspector's Guide to Environmental Microbiology Laboratory Testing.
- Atherholt, T.B., Procopio, N.A., Goodrow, S.M., 2017. Seasonality of Coliform Bacteria Detection Rates in New Jersey Domestic Wells. *Groundwater* 55, 346–361. <https://doi.org/10.1111/gwat.12482>
- Boisvert, E., Cotteret, G., 2021. Geoscience: Groundwater and aquifers [WWW Document]. <https://www.nrcan.gc.ca/earth-sciences/earth-sciences-resources/geoscience-groundwater-and-aquifers/10909>.
- Chiu, W.A., Ndeffo-Mbah, M.L., 2021. Using test positivity and reported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLoS Comput Biol* 17, 1–19. <https://doi.org/10.1371/journal.pcbi.1009374>
- De Roos, A.J., Kondo, M.C., Robinson, L.F., Rai, A., Ryan, M., Haas, C.N., Lojo, J., Fagliano, J.A., 2020. Heavy precipitation, drinking water source, and acute gastrointestinal illness in Philadelphia, 2015-2017. *PLoS One* 15, e0229258. <https://doi.org/10.1371/journal.pone.0229258>
- Di Pelino, S., Schuster-Wallace, C., Hynds, P.D., Dickson-Anderson, S.E., Majury, A., 2019. A coupled-systems framework for reducing health risks associated with private drinking water wells. *Canadian Water Resources Journal* 44, 280–290. <https://doi.org/10.1080/07011784.2019.1581663>

DMTI Spatial Inc., 2021. Network Data Set.

ESRI, 2022. How Kernel Density works—ArcGIS Pro | Documentation [WWW Document]. URL <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-kernel-density-works.htm> (accessed 8.15.22).

Friedman, J., 1991. Multivariate Adaptive Regression Splines. *Ann Stat* 19.

Hynds, P.D., Gill, L.W., Misstear, B.D., 2014. A Quantitative Risk Assessment of Verotoxigenic *E. coli* (VTEC) in Private Groundwater Sources in the Republic of Ireland. *Human and Ecological Risk Assessment: An International Journal* 206, 1446–1468. <https://doi.org/10.1080/10807039.2013.862065>

Invik, J., Barkema, H.W., Massolo, A., Neumann, N.F., Cey, E., Checkley, S., 2019. *Escherichia coli* contamination of rural well water in Alberta, Canada is associated with soil properties, density of livestock and precipitation. *Canadian Water Resources Journal* 44, 248–262. <https://doi.org/10.1080/07011784.2019.1595157>

Jabbar, F.K., Grote, K., Tucker, R.E., 2019. A novel approach for assessing watershed susceptibility using weighted overlay and analytical hierarchy process (AHP) methodology: a case study in Eagle Creek Watershed, USA. *Environmental Science and Pollution Research* 26, 31981–31997. <https://doi.org/10.1007/s11356-019-06355-9>

Kraay, A.N.M., Man, O., Levy, M.C., Levy, K., Ionides, E., Eisenberg, J.N.S., 2020. Understanding the impact of rainfall on diarrhea: Testing the concentration-dilution

hypothesis using a systematic review and meta-analysis. *Environ Health Perspect.*
<https://doi.org/10.1289/EHP6181>

Kreutzwiser, R., de Loë, R.C., Imgrund, K., 2010. Out of Sight, Out of Mind: Private Water Well Stewardship in Ontario. Report on the Findings of the Ontario Household Water Well Owner Survey 2008, Water Policy and Governance Group, University of Waterloo, Waterloo, ON.

Latchmore, T., Hynds, P., Brown, S., Schuster-Wallace, C., Dickson-Anderson, Sarah McDermott, K., Majury, A., 2020. Analysis of a Large Spatiotemporal Groundwater Quality Dataset, Ontario 2010 - 2017: Informing Human Health Risk Assessment and Testing Guidance for Private Drinking Water Wells.

Lavallee, S., Hynds, P.D., Brown, R.S., Majury, A., 2023. Classification of sub-populations for quantitative risk assessment based on awareness and perception: A cross-sectional population study of private well users in Ontario. *Science of the Total Environment* 857, 159677. <https://doi.org/10.1016/j.scitotenv.2022.159677>

Lavallee, S., Latchmore, T., Hynds, P.D., Brown, R.S., Schuster-Wallace, C., Anderson, S.D., Majury, A., 2021. Drinking Water Consumption Patterns among Private Well Users in Ontario: Implications for Exposure Assessment of Waterborne Infection. *Risk Analysis* 41, 1890–1910. <https://doi.org/10.1111/risa.13676>

- McKenney, D.W., Pedlar, J.H., Lawrence, K., Papadopol, P., Campbell, K., Hutchinson, M.F., 2014. Change and evolution in the plant hardiness zones of Canada. *Bioscience* 64, 341–350. <https://doi.org/10.1093/biosci/biu016>
- Michel, P., Wilson, J.B., Martin, S.W., Clarke, R.C., McEwen, S.A., Gyles, C.L., 1999. Temporal and geographical distributions of reported cases of *Escherichia coli* O157:H7 infection in Ontario. *Epidemiol Infect* 122, 193–200. <https://doi.org/10.1017/S0950268899002083>
- Murphy, H.M., Thomas, M.K., Schmidt, P.J., Medeiros, D.T., McFadyen, S., Pintar, K.D.M., 2016. Estimating the burden of acute gastrointestinal illness due to *Giardia*, *Cryptosporidium*, *Campylobacter*, *E. coli* O157 and norovirus associated with private wells and small water systems in Canada. *Epidemiol Infect* 144, 1355–1370. <https://doi.org/10.1017/S0950268815002071>
- Namugize, J.N., Jewitt, G., Graham, M., 2018. Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. *Physics and Chemistry of the Earth* 105, 247–264. <https://doi.org/10.1016/j.pce.2018.03.013>
- Niagara Peninsula Conservation Authority, 2011. Groundwater Vulnerability Analysis: Niagara Peninsula Source Protection Area. Welland.
- Niu, G.Y., Yang, Z.L., 2006. Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *J Hydrometeorol* 7, 937–952. <https://doi.org/10.1175/JHM538.1>

OASDI, 2023. Ontario's Climate and Seasons [WWW Document]. URL <https://www.oasdi.ca/living-in-ontario/weather-and-seasons/> (accessed 4.23.23).

O'Dwyer, J., Hynds, P.D., Byrne, K.A., Ryan, M.P., Adley, C.C., 2018. Development of a hierarchical model for predicting microbiological contamination of private groundwater supplies in a geologically heterogeneous region. *Environmental Pollution* 237, 329–338. <https://doi.org/10.1016/j.envpol.2018.02.052>

OMAFRA, 2020. Climate Zones and Planting Dates for Vegetables in Ontario [WWW Document]. Vegetable Crops. URL <http://www.omafra.gov.on.ca/english/crops/facts/climzoneveg.htm> (accessed 2.15.21).

Ontario Ministry of Environment Conservation and Parks, 2020. Well records - WWIS - Microsoft Access - Ontario Data Catalogue.

Ontario Ministry of Environment Conservation and Parks, 2019. Water Supply Wells: Requirements and Best Practices.

Ontario Ministry of Natural Resources and Forestry, 2019. Provincial Digital Elevation Model (PDEM). Data [WWW Document]. Ontario Ministry of Natural Resources and Forestry - Provincial Mapping Unit. URL <https://geohub.lno.gov.on.ca/datasets>

Paule-Mercado, M.A., Ventura, J.S., Memon, S.A., Jahng, D., Kang, J.H., Lee, C.H., 2016. Monitoring and predicting the fecal indicator bacteria concentrations from

- agricultural, mixed land use and urban stormwater runoff. *Science of the Total Environment* 550, 1171–1181. <https://doi.org/10.1016/j.scitotenv.2016.01.026>
- PlantMaps, 2022. Map of Ontario First and Last Frost Dates [WWW Document].
- Qayyum, S., Hynds, P., Richardson, H., McDermott, K., Majury, A., 2020. A geostatistical study of socioeconomic status (SES), rurality, seasonality and index test results as drivers of free private groundwater testing in southern Ontario, 2012–2016. *Science of the Total Environment* 717. <https://doi.org/10.1016/j.scitotenv.2020.137188>
- Qin, C., Zhu, A. -X., Pei, T., Li, B., Zhou, C., Yang, L., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *International Journal of Geographical Information Science* 21, 443–458. <https://doi.org/10.1080/13658810601073240>
- Ranstam, J., Cook, J.A., 2018. Statistical nugget LASSO regression. <https://doi.org/10.1002/bjs.10895>
- Rauf, A.A., Amani, M., 2017. An Update on the Use of Fiberglass Casing and Tubing in Oil and Gas Wells. *International Journal of Petroleum and Petrochemical Engineering* 3, 43–53. <https://doi.org/10.20431/2454-7980.0304004>
- Reynolds, C., Checkley, S., Chui, L., Otto, S., Neumann, N.F., 2020. Evaluating the risks associated with shiga-toxin-producing escherichia coli (Stec) in private well waters in Canada. *Can J Microbiol* 66, 337–350. <https://doi.org/10.1139/cjm-2019-0329>
- Rivera, A., 2017. The state of ground water in Canada. Ground Water Canada.

- Staley, Z.R., He, D.D., Edge, T.A., 2017. Persistence of fecal contamination and pathogenic *Escherichia coli* O157:H7 in snow and snowmelt. *J Great Lakes Res* 43, 248–254. <https://doi.org/10.1016/j.jglr.2017.01.006>
- Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw* 23, 1–46. <https://doi.org/10.18637/jss.v023.i07>
- Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S., Wilson, B.E., 2020. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. Data set. [WWW Document]. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. URL <https://doi.org/10.3334/ORNLDAAAC/1840>
- UN Water, 2015. Water for a Sustainable World, The United Nations World Water Development Report.
- White, K., Dickson-Anderson, S., Majury, A., McDermott, K., Hynds, P., Brown, R.S., Schuster-Wallace, C., 2021. Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset. *Water Res* 197, 117089. <https://doi.org/10.1016/j.watres.2021.117089>

White, K., Schuster-Wallace, C., Dickson-Anderson, S., 2023a. Converting Land Use – Land Cover to *E. coli* Contamination Potential Classes for Groundwater Wells: Utilizing a Large Ontario-based Dataset. Manuscript submitted for publication.

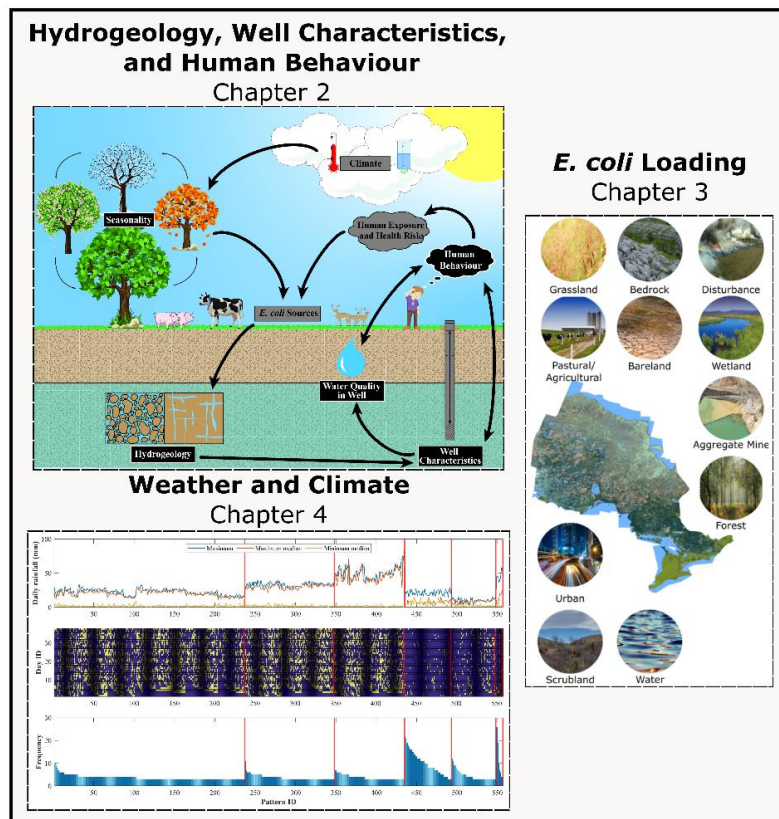
White, K., Yosri, A., Dickson-Anderson, S., Schuster-Wallace, C., 2023b. Exploring the role of rainfall intermittency on *E. coli* contamination events in private wells: an Ontario case study. Manuscript to be submitted for publication.

Whitman, R.L., Przybyla-Kelly, K., Shively, D.A., Nevers, M.B., Byappanahalli, M.N., 2008. Sunlight, season, snowmelt, storm, and source affect *E. coli* populations in an artificially ponded stream. Science of the Total Environment 390, 448–455.
<https://doi.org/10.1016/j.scitotenv.2007.10.014>

CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS

6.1 CONCLUSIONS

This thesis used data-driven methods to identify trends and relationships between the physical world, human behaviour, and *E. coli* presence and concentration in private wells. The key contributions of this thesis include an initial step toward the creation of a user-friendly predictive tool through the identification of relative variable importance and the creation, exploration, and trend identification of physical and human behaviour variables in the context of *E. coli* contamination potential (Chapters 2, 3, 4, and 5); the proven importance of meaningful data subsetting in the context of *E. coli* contamination potential with the innovative introduction of a new method of identifying seasonal delineations (Chapter 5); and a proof of concept regarding the requirement of a coupled-systems approach identifying the presence of feedback loops impacting *E. coli* presence in wells. It underscores the need to consider both the human and physical systems together, including the feedback between them, to have a good understanding of contaminant drivers and appropriate mitigation and intervention strategies.



**Private Well *E. coli* Contamination -
Identified Primary and Secondary Explanatory Variables**

Chapter 5

| | Summer | Spring/Autumn | Winter |
|-------------------|--|--|---|
| Secondary/Primary | - Land Use-Land Cover (LULC) <i>E. coli</i> Contamination Potential - Well Density | - LULC <i>E. coli</i> Contamination Potential - Bottom of Well Stratigraphy | |
| | - Testing Frequency - Well Depth - Bottom of Well Stratigraphy | - Well Depth - Total Water - Cumulative Sum of Consecutive Snow Melt Day | - LULC <i>E. coli</i> Contamination - Driving Time to Closest Drop-Off Location - Testing Frequency - Total Number of Snow Melt Days |

- Introduction of a geographically-driven, seasonal-based transport model for *E. coli*, utilizing a data-driven machine learning approach
- Demonstrated ability and need to incorporate a coupled-systems approach to private well water contamination prediction tools

Figure 6-1: Summary of work completed in this thesis through the exploration of the key drivers of *E. coli* contamination in private wells. Primary variables represent those with consensus across three regression models (Chapter 5) and secondary variables represent those with consensus across two regression models (Chapter 5).

First, an exploratory data analysis and proof of methods concept was conducted on a large private well dataset, exploring the relationships between microbiological, hydrogeological, well characteristic, and testing behaviour variables utilizing supervised machine learning techniques. Second, an innovative method was developed to explore land use-land cover as a driver of *E. coli* contamination in private wells. Third, wet-dry patterning was explored as a driver of *E. coli* contamination events, which moves beyond the current standard of using rainfall lag periods. Finally, moving toward a coupled-systems approach, geographically-driven seasonal-based models were developed based on the key findings from Chapters 2-4. The main conclusions of this work are summarized in the following sections.

6.1.1 Exploration of private well dataset utilizing innovative machine learning approaches

- The supervised machine learning techniques, GAMLSS and ARA, were shown to be valuable in exploring and explaining the drivers of *E. coli* contamination in private wells as a function of hydrogeologic and well and aquifer characteristic variables. However, the importance of coupling machine learning driven methods with disciplinary expertise was also identified.
- Latitude was determined to be a stronger driver of *E. coli* contamination when compared to typical seasonal delineations. This suggests that there is an opportunity for improved understanding of seasons in Ontario.

- Important physical drivers of *E. coli* contamination included: well depth, where deeper wells were more protective from contamination, but did not eliminate *E. coli* completely; and bottom of well stratigraphy, where wells completed in sedimentary and igneous formations experienced more *E. coli* contamination events compared to metamorphic and unconsolidated formations.
- Important human behaviours pertaining to testing frequency included initial test message received impacting testing patterns; and timing of testing resulting in a decoupling between peak testing and peak *E. coli* presence in later study years.

6.1.2 Introduction of data-driven LULC *E. coli* contamination potential class mapping

- Regression analyses were used in conjunction with literature-based confidence ratings to assess the impact of LULC on *E. coli* contamination potential for wells and derive an easy-to-use geospatial raster dataset.
- Very high (i.e., pastoral/agricultural), high (i.e., urban, aggregate mines, open/barelands), moderate (i.e., forest, bedrock, disturbance, scrubland), and low (i.e., water, grasslands, wetlands) *E. coli* contamination potential classes for wells in Ontario were identified.
- Derived raster dataset was utilized to identify policy and regulatory implications including specific well characteristics (i.e., well depth) based on geographic location, and a proposed expansion of the definition of vulnerable areas under the *Clean Water Act, 2006* to encompass very high and high LULC *E. coli* contamination potential classes.

6.1.3 Exploration of wet-dry patterning as a driver of *E. coli* contamination events, moving beyond current standard rainfall lag periods

- The supervised machine learning technique random forest-ensembled decision tree-based classification finds the optimal lag time to explain an *E. coli* contamination event in private wells is 36 days.
- Contamination events are significantly less likely to occur if there is no rainfall for 36 days, strengthening the argument of rainfall being a driver of contamination.
- An increase in number of wet days ($>0.1\text{mm}$) or wet/dry cycles increases the likelihood of *E. coli* presence in a well while decreasing the concentration, supporting the importance of the concentration/dilution effect.

6.1.4 Development of geographically-driven, seasonally-based models, moving toward a coupled-systems approach, to explain *E. coli* contamination in private wells

- The developed models were found to be valuable in beginning to explain seasonally-based trends in *E. coli* presence and contamination in private wells.
- Human behaviour-based variables appear in most models, validating the value of utilizing a coupled-systems approach.
- “Summer” models found that the primary explanatory variables include LULC *E. coli* contamination potential and well density, while secondary explanatory variables include testing frequency, well depth, and bottom of well stratigraphy.
- “Shoulder” models found that the primary explanatory variables include LULC *E. coli* contamination potential and bottom of well stratigraphy, while secondary explanatory

variables include well depth, total water, and cumulative sum of consecutive snow melt days.

- “Winter” models found that while there were no primary explanatory variables, secondary explanatory variables include LULC *E. coli* contamination potential, driving time to closest drop-off location, testing frequency, and total number of snow melt days.

6.2 STUDY LIMITATIONS

This work exhibits three primary areas of study limitations: data limitations, variables used, and limitations of machine learning. As mentioned throughout, the WWIS data are derived from borehole logs that were initially handwritten and later transcribed into an online database. This transcription process introduced errors into the dataset, as evidenced by unrealistic values for certain variables (e.g., casing diameters less than zero). Unrealistic values were removed from the analyses. There was no indication that these errors were systematic. Additionally, the microbiological testing in the WWTD utilized the membrane filtration method, using differential coliform agar for the simultaneous detection and enumeration of total coliforms and *E. coli*. Adverse results were considered presumptive after 18-24 hours, potentially resulting in false positives or negatives in the dataset. Identifying error values to account for the possibility of false positives or negatives was not feasible because results were not confirmed further using methods such as MALDI-TOF, DNA-based, or biochemical testing. Additionally, colony-forming units are counted by lab technicians, which could introduce some additional human error. Given these data

were captured by PHO and sufficient information was not supplied to calculate these potential errors, all coliform and *E. coli* results are assumed to represent the true count. Finally, some dataset dates don't align with the primary WWIS dataset time period (2010-2017). As an example, the LULC data employed in this study were reported to have been collected between 2000 and 2015. Since there is no means to verify the specific LULC conditions on the day of each well water observation, it was assumed that the LULC data remained accurate for the period corresponding to the *E. coli* results.

Regarding the independent variables used in the analyses, there are always opportunities to include more, different, or more specific information. Some examples include: the addition of potential point source pollution features, such as septic tank locations, to the analyses; expansion of the LULC variable to encompass more nuanced land use categories integrating tilling practices or distinguishing urban areas in residential versus commercial zones, all of which offer differing levels of contaminant loading and transport characteristics. Another possibility is identifying a new way to incorporate all stratigraphic information from a given well to determine the impact on *E. coli* contamination potential by uncovering trends in transport mechanisms. Continuously exploring and integrating independent variables will help transition the exploratory model findings in this work toward an eventual predictive model that can be utilized by private well users and policy makers.

Finally, the use of machine learning techniques offers its own set of potential limitations. One limitation, impacted by data availability, is that machine learning models

are only as good as the data given to it during the training and validation processes. If the data is too aggregated (not capturing the nuances aimed to depict), or misrepresented (not capturing all potential results proportionately), the model may not have the appropriate information to improve the model's performance or may create a model too general for the goal. Another limitation includes the need for computational resources required. While this work was able to utilize some computationally intensive regression models, there are other deep learning models, such as ANN, that may further improve the results of this work. The final limitation that will be discussed here is the proven requirement of subject matter expertise when using machine learning approaches. It is imperative that subject matter experts remain hypervigilant while interpreting the outputs to ensure that correlations and relationships derived by machine learning models are mechanistic and based on a solid understanding of the physical system.

6.3 RECOMMENDATIONS FOR FUTURE RESEARCH

The research presented in this thesis is aimed towards the creation of a predictive *E. coli* transport model for private wells in Ontario. The results identified key drivers of *E. coli* contamination which would not have been identified without utilizing innovative methods and a large geospatial-temporal dataset containing key microbiological variables. However, this work is an initial phase of a predictive transport model, as such there are several recommendations for future work to continue this goal.

With improved understanding of unique variables (e.g., well depth, LULC, driving time to closest drop-off location, etc.), programs such as Maxent, which utilizes a machine

learning approach called maximum entropy modelling, could be used to create a model using a mapping-based approach (Phillips et al., 2017). This technique would involve creating individual geospatial maps of each driver of *E. coli* presence and concentration, layering them on top of one another, and creating an output that can be interpreted as predicted probability of presence in a geospatial format. In addition to variables explored through chapters 2 to 5, supplementary variables assessing well users KAPs and *E. coli* fate could be further explored to improve the accuracy and holistic nature of models. Once complete, model outputs could be combined with quantitative microbial risk assessment (QMRA) profiles to better inform well users how *E. coli* presence and concentration in wells could directly impact their health.

This work has continued to identify how, due to the large geographic expanse of Ontario, vastly different conditions are when considering seasonality and weather, human behaviour, hydrogeology, and policy. While trends uncovered in this work aim to begin parsing out the relative importance of variables based on seasonality, finding that these subsets improve performance, there are many additional small-scale subsets that can be explored. Exploring subset case-studies based on LULC classes, bottom of well stratigraphy formations, or latitude-based geographic locations could be excellent next steps towards smaller-scale predictive models.

This work has demonstrated that a coupled-systems approach advances a holistic understanding of the system by acknowledging and modeling the interactions and impacts the physical world has on the human world, and vice versa. While this work begins to

evaluate human behaviour alongside physical variables, the next step is to include the critical concept of feedback loops in the models. The concept of feedback loops aim to identify the relationship between variables whereby two variables may have a positive (e.g., as variable one increases so does variable 2) or negative (e.g., as variable one increases, variable 2 decreases) feedback loop. This integration will further improve policy and well stewardship recommendations as more complex impacts between variables can be captured.

6.4 WORKS CITED

Phillips, S., Anderson, R., Dudik, M., Schapire, R., Blair, M., 2017. Opening the black box: an open-source release of Maxent. *Journal of Ecography* 40. <https://doi.org/10.1111/ecog.03049>

CHAPTER A APPENDICES

A.2 CHAPTER 2 APPENDICES

A.2.1 Data Processing

A.2.1.1 *E. coli* Enumeration

The province of Ontario provides free testing of private drinking water wells for bacterial indicators of contamination (i.e. total coliforms and *E. coli*). Provincial laboratories in Ontario use the membrane filtration (MF) method using differential coliform (DC) agar for the simultaneous detection and enumeration of total coliforms and *E. coli*, where results are reported as colony forming units (CFU)/100 mL (Ontario Agency for Health Protection and Promotion, 2019). All positive (i.e., adverse; > 0 CFU/100 mL) results are considered presumptive after 18-24 hrs (i.e., occasional false negative and positive readings can occur) as they are not confirmed further (using e.g., MALDI-TOF, DNA-based, or biochemical testing). Given that there is insufficient information to determine the true uncertainty within this particular dataset, the numerical count of *E. coli* in a sample is assumed to represent the true count. Adverse results were categorized as follows: Category 1: 1-10 CFU/100 mL; Category 2: 11-50 CFU/100 mL; Category 3: 51+ CFU/100 mL (Bain et al., 2014). Observations were excluded from analyses employing *E. coli* contamination levels if the sample status indicated a flagged issue (Table A.2-3), as they provide no contaminant-specific information. This resulted in a total of 789,072 observations from 251,422 unique wells for the analyses. Sample flags that did not return

a numerical value (“no result”) for *E. coli* were re-classified according to cause (Table A.2-3).

To explore testing behaviour, sample results were also re-classified according to the message returned to users (Public Health Ontario, 2020):

1. “no significant evidence” - no evidence of *E. coli* and Total Coliforms are ≤ 5 CFU/100mL (n = 580,202);
2. “no result” - observations are not available due to PI, CT, AO, IS, and US (Table A.2-3) (n = 42,014) (only used to assess impact of this message on testing behaviour);
3. “may be unsafe” - no evidence of *E. coli*, but more than five total coliforms are present (n= 103,694) (only used to assess impact of this message on testing behaviour); and,
4. “unsafe to drink” - observations contain *E. coli* (Category 1-3) or a sample is overgrown, either with or without evidence of total coliforms or *E. coli* (Status 22 and 4 in Table A.2-3, respectively) the presence of *E. coli* cannot be guaranteed or quantified, but also cannot be ruled out (n = 63,163).

A.2.1.2 Data Classification

Bedrock wells were classified based on the rock type in which they were completed (i.e., the rock type located at the bottom of the well). Major categories were metamorphic (n = 7,089, 1.4%), sedimentary (n = 366,620, 70.4%), and igneous (n = 146,842, 28.2%) (Table A.2-1). All wells were also classified according to the permeability of the geological formation in which they were completed, i.e., low (n = 148,880, 33.2%), medium (n = 254,441, 56.8%), and high (n = 44,568, 10.0%) (Table A.2-2) (Freeze and Cherry, 1978).

Well depths were reclassified into shallow/moderate (< 12.5 m), moderate (Moderate1, $12.5 \text{ m} \leq x < 18.3 \text{ m}$; Moderate2, $18.3 \text{ m} \leq x < 24.4 \text{ m}$; Moderate3, $24.4 \text{ m} \leq x < 31.1 \text{ m}$; Moderate4, $31.1 \text{ m} \leq x < 41.8 \text{ m}$; Moderate5, $41.8 \text{ m} \leq x < 61 \text{ m}$), and deep ($> 61 \text{ m}$). Categories were determined to distribute the data evenly and maintain some consistency with the Ontario regulations pertaining to new well construction (Ontario Ministry of Environment Conservation and Parks, 2019).

Specific well capacity (SC) was calculated as (Freeze and Cherry, 1978):

$$SC = \frac{Q}{S} \quad [\text{Eq. A.2.1}]$$

where SC is the specific capacity ($\text{L}^3 \cdot \text{T}^{-1} \cdot \text{L}^{-1}$); Q is the pumping rate ($\text{L}^3 \cdot \text{T}^{-1}$); and, S is the drawdown (L). Once calculated, data were cleaned to remove negative and undefined specific capacities. Equal data bins were used to classify the remaining data into the following ranges: $0 \text{ gpm/m} < x \leq 3.3 \text{ gpm/m}$, $3.3 \text{ gpm/m} < x \leq 16.4 \text{ gpm/m}$, and $> 16.4 \text{ gpm/m}$.

A.2.1.3 User Testing Recommendations

Recommendations for user testing vary across Canada, falling between one and three times per year, as discussed by Maier et al. (2014). Currently, there is no Ontario-specific guideline, thus the Health Canada guideline of testing “at least two times per year” has been adopted for the purpose of this work (Health Canada, 2020). Testing frequencies in this study were therefore classified according to whether an individual well was tested less than or at least 16 times over the eight-year dataset period. Wells were categorized as

testing on aggregate less than two times per year (< 16 samples) or meeting or exceeding two times per year (≥ 16 samples).

A.2.2 Statistical Analyses

A.2.2.1 Seasonal Analyses

Regression analyses conducted on seasonality involved assessing *E. coli* concentration as the dependent variable and latitude and longitude coordinates as independent variables, adding three-month season-based delineations (winter starting in either November, December, or January), month of observation, and year of observation in three separate models.

A.2.2.2 Well Characteristics

To determine the probability of contamination given well depth, the probability of each depth subcategory resulting in a specific contamination concentration was determined as follows:

$$P_{D,C} = \frac{n_{D,C}}{n_D} \quad [\text{Eq. A.2.2}]$$

where $P_{D,C}$ is the probability of contamination at a given depth subcategory (D) and *E. coli* concentration (C); $n_{D,C}$ is the number of observations at a given D and C ; and, n_D is the number of observations at a given D . Analyses exploring the probability of contamination given geological formation were conducted in the same manner, replacing D with the respective variable of interest in Eq. A.2.2 Two-tailed hypothesis tests were subsequently used to assess the statistical significance at several levels.

A.2.2.3 Well Sampling

The total number of tests were standardized as follows:

$$P_{T,i,n} = \frac{Freq_{i,n}}{T_i} \times 100\% \quad [\text{Eq. A.2.3}]$$

where i is the index test status (“no significant evidence”, “no result”, “may be unsafe”, “unsafe to drink”); n is the number of tests conducted on an individual well; $P_{T,i,n}$ is the percentage of total individual wells with an initial test status, i , and number of tests, n ; $Freq_{i,n}$ is the total number of individual wells with initial test status, i , and number of tests, n ; and, T_i is the total number of individual wells with initial test status, i .

Decay curves were created to compare decay rates across initial test statuses as follows:

$$y(n_i) = y_f + (y_o - y_f)e^{-\exp(\log \alpha)n_i} \quad [\text{Eq. A.2.4}]$$

where i is the index test status (“no significant evidence”, “no result”, “may be unsafe to drink”, “unsafe to drink”); n_i is the number of tests conducted on an individual well with a status i ; $y(n_i)$ is the percentage of total individual wells at a number of test, n , with an initial test status, i ; y_o is the function’s starting value, y_f is the value the function decays towards; and α is the decay rate of the function.

Adverse test results for each month were standardized with respect to year as follows:

$$AR_{m,y,a} = \frac{AT_{m,y,a}}{AT_{total,y,a}} \times 100\% \quad [\text{Eq. A.2.5}]$$

where a is the adverse *E. coli* sample concentration category (Category 1, 1-10 CFU/100mL; Category 2, 11-50 CFU/100mL; Category 3, 51+ CFU/100mL); $AT_{m,y,a}$ is the number of adverse tests in a given month, m , year, y , and concentration category, a ; $AT_{total,y,a}$ is the total number of adverse tests in the target year, y and concentration category, a ; and, $AR_{m,y,a}$ is the percentage of adverse results in the target month, year, and adverse concentration category.

A.2.2 Table and Figures

Table A.2-1: Classification of bedrock types.

| Metamorphic (7,089 obs) | Sedimentary (366,620 obs) | Igneous (146,842 obs) |
|--------------------------------|----------------------------------|------------------------------|
| Gneiss (GNIS) (127 obs) | Chert (CHRT) (121 obs) | Basalt (BSLT) (290 obs) |
| Schist (SHST) (66 obs) | Conglomerate (CONG) (128 obs) | Granite (GRNT) (146,522 obs) |
| Slate (SLTE) (1773 obs) | Dolomite (DLMT) (3,812 obs) | |
| Feldspar (FLDS) (64 obs) | Flint (FLNT) (218 obs) | |
| Greenstone (GRSN) (1,838 obs) | Gypsum (GYPS) (27 obs) | |
| Marble (MRBL) (660 obs) | Limestone (LMSN) (273,995 obs) | |
| Greywacke (GRWK) (86 obs) | Iron Formation (IRFM) (24 obs) | |
| Quartzite (QRTZ) (2,341 obs) | Marl (MARL) (75 obs) | |
| Soapstone (SPST) (134 obs) | Quartz (QTZ) (5,479 obs) | |
| | Shale (SHLE) (50,435 obs) | |
| | Sandstone (SNDS) (32,306 obs) | |

Table A.2-2: Classification of permeability.

| Low Permeability Formations | Medium Permeability Formations | High Permeability Formations |
|------------------------------------|---------------------------------------|-------------------------------------|
| Clay (53,102 obs) | Coarse Sand (18,331 obs) | Boulders (4248 obs) |
| Clayey (6 obs) | Gravel (97,787 obs) | Coarse-Grained (1171 obs) |
| Cemented (1538 obs) | Gravelly (102 obs) | Coarse Gravel (7073 obs) |
| Dense (4816 obs) | Loose (16,420 obs) | Pea Gravel (222 obs) |
| Dry (60 obs) | Medium-Grained (7712 obs) | Porous (4851 obs) |
| Fine-Grained (622 obs) | Medium-Gravel (985 obs) | Stoney (316 obs) |
| Fine-Gravel (2127 obs) | Medium-Sand (29,676 obs) | Stones (20,959 obs) |
| Fine-Sand (18,535 obs) | Quicksand (1517 obs) | Fractured (4757 obs) |
| Hardpan (3967 obs) | Sand (81,911 obs) | Fractured Rock (971 obs) |
| Packed (5738 obs) | | |
| Silt (11,587 obs) | | |
| Silty (1068 obs) | | |
| Rock (45,714 obs) | | |

Table A.2-3: NA sample status explanations.

| Sub classification | Status | Interpretation |
|---------------------------|---------------|---|
| Process Issue (PI) | 7 | The sample was too old |
| | 8 | The bottle was received broken or damaged |
| | 11 | The requisition was received separated from the sample bottle |
| | 12 | The sample was not collected in the proper bottle |
| | 13 | The sample was received frozen |
| | 14 | The sample was collected from a hot water tap |
| | 15 | Insufficient sample was submitted |
| | 17 | The sample leaked in transit |
| | 20 | Insufficient information was supplied on the sample |
| | 21 | Not tested for some other reason |
| | 24 | Sample received was very warm |
| | 27 | Unique identifier missing |
| | 28 | Outdated collection kit received |
| | 30 | Laboratory error |
| NDOGN | 4 | No Data: Overgrown with non-target - May be unsafe to drink |
| CT | 6 | The client requested chemical testing |
| AO | 10 | The appearance or odour makes the sample unacceptable as drinking water |
| IS | 26 | Interfering substances in the sample |
| US | 29 | Unauthorized submitter |
| NDOGT | 22 | No data: overgrown with target – unsafe to drink |
| OMIT | 5 | Sample taken from an unprotected source, such as a lake or river |
| | 18 | Sample was collected from a source located outside of Ontario |
| | 19 | Sample was collected from a municipal water supply |
| | 0 | Not yet tested |
| | 25 | Bottled water submitted |

Table A.2-4: Variable descriptions for regression analyses.

| Variables | Data Type | Data Range | Model of Interest |
|------------------------------|------------------|--|---|
| <i>E. coli</i> | Discrete | 0-81 CFU/100mL | seasonality, hydrogeological, informed physical |
| Longitude | Discrete | In 0.5-degree increments: -85.5 - -75.0 | seasonality, informed physical |
| Latitude | Discrete | In 0.5-degree increments: 42.5 - 49.0 | seasonality, informed physical |
| Seasonal Delineations | Categorical | Season 1: 1: Jan-Mar 2: Apr-Jun 3: Jul-Sept 4: Oct-Dec Season 2: 1: Feb-Apr 2: May-July 3: Aug-Oct 4: Nov-Jan Season 3: 1: Mar-May 2: Jun-Aug 3: Sept-Nov 4: Dec-Feb | seasonality |
| Month | Categorical | January - December | seasonality, informed physical, testing practices |
| Year | Categorical | 2010 - 2017 | seasonality, informed physical, testing practices |
| Bottom Stratigraphy | Categorical | Consolidated Unconsolidated | hydrogeological, informed physical |

| | | | |
|----------------------------------|-------------|--|------------------------------------|
| | Categorical | See Table A.2-10 for full list of levels | hydrogeological |
| Specific Capacity | Continuous | 0.001 - 1919 gpm/m | hydrogeological, informed physical |
| Well Depth | Continuous | 1 - 760 m | hydrogeological |
| | Categorical | ShallowModerate (< 12.5 m) | hydrogeological, informed physical |
| | | Moderate 1 (12.5 m ≤ x < 18.3 m) | |
| | | Moderate 2 (18.3 m ≤ x < 24.4 m) | |
| | | Moderate 3 (24.4 m ≤ x < 31.1 m) | |
| | | Moderate 4 (31.1 m ≤ x < 41.8 m) | |
| | | Moderate 5 (41.8 m ≤ x < 61 m) | |
| | | Deep (≥ 61 m) | |
| Casing Diameter | Continuous | 0 - 190.5 m | hydrogeological |
| | Categorical | Narrow (< 5 cm) | hydrogeological |
| | | Moderate (5 cm ≤ x ≤ 61 cm) | |
| | | Large (> 61 cm) | |
| Year of Well Construction | Discrete | 1903 - 2017 | hydrogeological |
| Testing Frequency | Discrete | 1-446 | testing practices |
| Message Returned to User | Categorical | no significant evidence no result may be unsafe unsafe to drink | testing practices |

Table A.2-5: Statistical summary of seasonality, hydrogeological, well characteristics, and testing practices model variables used in regression analyses.

| Model | Variable | Min | 1 st Qu. | Median | Mean | 3 rd Qu. | Max | SD | p-value ¹ |
|---|--------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-----------------------|----------------------|
| Seasonality – Seasons Dependent Variable: <i>E. coli</i> Concentration Intercept: 8.26 ± 6.23 | Longitude | -2.93x10 ⁻² | -2.93x10 ⁻² | -2.72x10 ⁻² | -2.68x10 ⁻² | -2.57x10 ⁻² | -2.25x10 ⁻² | 2.64x10 ⁻³ | <0.01 |
| | Latitude | -1.68x10 ⁻¹ | -1.64x10 ⁻¹ | -1.60x10 ⁻¹ | -1.60x10 ⁻¹ | -1.59x10 ⁻¹ | -1.45x10 ⁻¹ | 6.71x10 ⁻³ | <0.01 |
| | Winter (1 - Jan) | -1.62 | -1.58 | -1.57 | -1.12 | -7.21x10 ⁻¹ | 6.36x10 ⁻² | 6.28 | 0.58 |
| | Spring (1 – April) | -1.75 | -1.74 | -1.72 | -1.27 | -8.87x10 ⁻¹ | -6.30x10 ⁻² | 6.28 | 0.55 |
| | Summer (1- July) | -1.75 | -1.73 | -1.71 | -1.27 | -8.80x10 ⁻¹ | -7.69x10 ⁻² | 6.28 | 0.55 |
| | Fall (1-Oct) | -1.89 | -1.87 | -1.86 | -1.41 | -1.01 | -2.34x10 ⁻¹ | 6.29 | 0.49 |
| Seasonality – Months Dependent Variable: <i>E. coli</i> Concentration Intercept: 9.04 ± 2.91x10⁻¹ | Longitude | -2.92x10 ⁻² | -2.60x10 ⁻² | -2.43x10 ⁻² | -2.45x10 ⁻² | -2.24x10 ⁻² | -2.10x10 ⁻² | 2.74x10 ⁻³ | <0.01 |
| | Latitude | -1.70x10 ⁻¹ | -1.65x10 ⁻¹ | -1.62x10 ⁻¹ | -1.61x10 ⁻¹ | -1.55x10 ⁻¹ | -1.52x10 ⁻¹ | 6.39x10 ⁻³ | <0.01 |
| | January | 1.19x10 ⁻¹ | 1.76x10 ⁻¹ | 2.04x10 ⁻¹ | 2.10x10 ⁻¹ | 2.25x10 ⁻¹ | 3.16x10 ⁻¹ | 5.75x10 ⁻² | 0.10 |
| | February | 5.15x10 ⁻² | 7.71x10 ⁻² | 1.28x10 ⁻¹ | 1.33x10 ⁻¹ | 1.71x10 ⁻¹ | 2.53x10 ⁻¹ | 6.67x10 ⁻² | 0.36 |
| | March | 1.81x10 ⁻¹ | 2.11x10 ⁻¹ | 2.63x10 ⁻¹ | 2.48x10 ⁻¹ | 2.77x10 ⁻¹ | 3.08x10 ⁻¹ | 4.49x10 ⁻² | 0.03 |
| | May | -1.23x10 ⁻¹ | -1.10x10 ⁻¹ | -6.62x10 ⁻² | -6.16x10 ⁻² | -1.03x10 ⁻² | 7.88x10 ⁻³ | 5.25x10 ⁻² | 0.55 |
| | June | 6.74x10 ⁻² | 8.73x10 ⁻² | 1.19x10 ⁻¹ | 1.21x10 ⁻¹ | 1.40x10 ⁻¹ | 1.97x10 ⁻¹ | 4.36x10 ⁻² | 0.20 |
| | July | 1.99x10 ⁻² | 6.12x10 ⁻² | 7.51x10 ⁻² | 8.19x10 ⁻² | 9.27x10 ⁻² | 1.59x10 ⁻¹ | 4.30x10 ⁻² | 0.37 |
| | August | -2.02x10 ⁻² | 2.68x10 ⁻² | 4.61x10 ⁻² | 6.56x10 ⁻² | 1.22x10 ⁻¹ | 1.48x10 ⁻¹ | 5.97x10 ⁻² | 0.53 |
| | September | -6.68x10 ⁻² | -3.26x10 ⁻² | -1.30x10 ⁻² | -1.34x10 ⁻² | -9.61x10 ⁻⁴ | 6.16x10 ⁻² | 3.41x10 ⁻² | 0.77 |

| | | | | | | | | | |
|------------------------------------|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-----------------------|-----------------|
| | October | -4.16x10 ⁻² | -1.36x10 ⁻² | 1.25x10 ⁻² | 1.66x10 ⁻² | 4.00x10 ⁻² | 9.12x10 ⁻² | 4.36x10 ⁻² | 0.70 |
| | November | -2.38x10 ⁻¹ | -1.84x10 ⁻¹ | -1.55x10 ⁻¹ | -1.63x10 ⁻¹ | -1.32x10 ⁻¹ | -9.71x10 ⁻² | 4.48x10 ⁻² | 0.12 |
| | December | -3.82x10 ⁻¹ | -2.89x10 ⁻¹ | -2.76x10 ⁻¹ | -2.72x10 ⁻¹ | -2.54x10 ⁻¹ | -1.79x10 ⁻¹ | 5.15x10 ⁻² | 0.03 |
| Seasonality – | Longitude | -3.19x10 ⁻² | -2.81x10 ⁻² | -2.55x10 ⁻² | -2.58x10 ⁻² | -2.34x10 ⁻² | -2.17x10 ⁻² | 3.63x10 ⁻² | <0.01 |
| Years | Latitude | -1.84x10 ⁻¹ | -1.77x10 ⁻¹ | -1.71x10 ⁻¹ | -1.73x10 ⁻¹ | -1.67x10 ⁻¹ | -1.66x10 ⁻¹ | 2.97x10 ⁻² | <0.01 |
| Dependent | 2011 | 7.00x10 ⁻² | 1.14x10 ⁻¹ | 1.39x10 ⁻¹ | 1.25x10 ⁻¹ | 1.42x10 ⁻¹ | 1.62x10 ⁻¹ | 3.30x10 ⁻³ | 0.04 |
| Variable: | 2012 | 8.37x10 ⁻² | 9.71x10 ⁻² | 1.48x10 ⁻¹ | 1.31x10 ⁻¹ | 1.58x10 ⁻¹ | 1.66x10 ⁻¹ | 6.52x10 ⁻³ | 0.06 |
| <i>E. coli</i> | 2013 | 2.72x10 ⁻² | 4.73x10 ⁻² | 8.60x10 ⁻² | 7.71x10 ⁻² | 1.04x10 ⁻¹ | 1.28x10 ⁻¹ | 3.26x10 ⁻² | 0.25 |
| Concentration | 2014 | -2.06x10 ⁻² | 9.53x10 ⁻³ | 1.65x10 ⁻² | 1.96x10 ⁻² | 2.97x10 ⁻² | 6.12x10 ⁻² | 3.38x10 ⁻² | 0.70 |
| Intercept: | 2015 | 8.54x10 ⁻³ | 3.26x10 ⁻² | 4.14x10 ⁻² | 3.84x10 ⁻² | 5.00x10 ⁻² | 5.43x10 ⁻² | 3.50x10 ⁻² | 0.55 |
| 9.54 ± 2.92x10⁻¹ | 2016 | -3.89x10 ⁻¹ | -3.54x10 ⁻¹ | -3.40x10 ⁻¹ | -3.33x10 ⁻¹ | -3.01x10 ⁻¹ | -2.81x10 ⁻¹ | 2.29x10 ⁻² | <0.01 |
| | 2017 | -8.22x10 ⁻¹ | -7.97x10 ⁻¹ | -7.70x10 ⁻¹ | -7.75x10 ⁻¹ | -7.59x10 ⁻¹ | -7.23x10 ⁻¹ | 1.50x10 ⁻² | <0.01 |
| Hydrogeological | | | | | | | | | |
| Dependent | Unconsolidated Bottom | | | | | | | | |
| Variable: | Stratigraphy | 1.05x10 ⁻¹ | 1.33x10 ⁻¹ | 1.36x10 ⁻¹ | 1.41x10 ⁻¹ | 1.52x10 ⁻¹ | 1.73x10 ⁻¹ | 1.91x10 ⁻² | <0.01 |
| <i>E. coli</i> | | | | | | | | | |
| Concentration | | | | | | | | | |
| Intercept: | Specific Capacity | -1.75x10 ⁻³ | 1.06x10 ⁻³ | 1.50x10 ⁻³ | 1.75x10 ⁻³ | 2.84x10 ⁻³ | 4.20x10 ⁻³ | 1.75x10 ⁻³ | 0.73 |
| 3.66 ± 3.12x10⁻² | | | | | | | | | |
| Well | Shallow/ModerateDepth | 1.37x10 ⁻¹ | 1.87x10 ⁻¹ | 2.01x10 ⁻¹ | 2.01x10 ⁻¹ | 2.25x10 ⁻¹ | 2.52x10 ⁻¹ | 3.77x10 ⁻² | 0.01 |
| Characteristics | Moderate1 Depth | 1.39x10 ⁻¹ | 1.87x10 ⁻¹ | 2.13x10 ⁻¹ | 2.01x10 ⁻¹ | 2.18x10 ⁻¹ | 2.41x10 ⁻¹ | 3.33x10 ⁻² | 0.01 |

| | | | | | | | | | |
|---|------------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-----------------------|-----------------|
| Dependent Variable: <i>E. coli</i> Concentration Intercept: $3.66 \pm 3.12 \times 10^{-2}$ | Moderate2 Depth | 8.10×10^{-2} | 1.24×10^{-1} | 1.51×10^{-1} | 1.44×10^{-1} | 1.68×10^{-1} | 1.78×10^{-1} | 3.10×10^{-2} | 0.04 |
| | Moderate3 Depth | 1.54×10^{-1} | 2.00×10^{-1} | 2.23×10^{-1} | 2.21×10^{-1} | 2.45×10^{-1} | 2.74×10^{-1} | 3.61×10^{-2} | <0.01 |
| | Moderate4 Depth | 5.03×10^{-2} | 9.82×10^{-2} | 1.29×10^{-1} | 1.21×10^{-1} | 1.50×10^{-1} | 1.73×10^{-1} | 3.86×10^{-2} | 0.11 |
| | Moderate5 Depth | 1.05×10^{-1} | 1.77×10^{-1} | 2.01×10^{-1} | 1.89×10^{-1} | 2.20×10^{-1} | 2.39×10^{-1} | 4.61×10^{-2} | 0.02 |
| Informed Physical Dependent Variable: <i>E. coli</i> Concentration Intercept: $9.14 \pm 7.17 \times 10^{-1}$ | Unconsolidated Bottom Stratigraphy | 3.94×10^{-2} | 6.03×10^{-2} | 6.84×10^{-2} | 6.76×10^{-2} | 7.27×10^{-2} | 9.25×10^{-2} | 1.71×10^{-2} | <0.01 |
| | Specific Capacity | -8.56×10^{-3} | -4.01×10^{-3} | -2.27×10^{-3} | -1.99×10^{-3} | 1.00×10^{-3} | 2.44×10^{-3} | 3.42×10^{-3} | 0.09 |
| | Shallow/ModerateDepth | 5.67×10^{-2} | 7.68×10^{-2} | 8.50×10^{-2} | 9.37×10^{-2} | 1.10×10^{-1} | 1.64×10^{-1} | 3.25×10^{-2} | 0.69 |
| | Moderate1 Depth | 1.36×10^{-1} | 1.41×10^{-1} | 1.54×10^{-1} | 1.62×10^{-1} | 1.69×10^{-1} | 2.10×10^{-1} | 2.71×10^{-2} | 0.21 |
| | Moderate2 Depth | 3.74×10^{-2} | 6.22×10^{-2} | 6.99×10^{-2} | 7.92×10^{-2} | 8.78×10^{-2} | 1.50×10^{-1} | 3.24×10^{-2} | 0.02 |
| | Moderate3 Depth | 1.30×10^{-1} | 1.50×10^{-1} | 1.67×10^{-1} | 1.75×10^{-1} | 2.08×10^{-1} | 2.14×10^{-1} | 3.20×10^{-2} | 0.24 |
| | Moderate4 Depth | 2.43×10^{-2} | 4.34×10^{-2} | 6.01×10^{-2} | 5.89×10^{-2} | 7.03×10^{-2} | 9.28×10^{-2} | 2.18×10^{-2} | 0.01 |
| | Moderate5 Depth | 1.23×10^{-2} | 6.63×10^{-2} | 8.24×10^{-2} | 7.86×10^{-2} | 8.70×10^{-2} | 1.44×10^{-1} | 3.36×10^{-2} | 0.38 |
| | January | 1.56×10^{-1} | 1.87×10^{-1} | 2.22×10^{-1} | 2.26×10^{-1} | 2.54×10^{-1} | 3.01×10^{-1} | 5.02×10^{-2} | 0.30 |
| | February | 1.59×10^{-1} | 1.71×10^{-1} | 1.97×10^{-1} | 2.15×10^{-1} | 2.67×10^{-1} | 2.97×10^{-1} | 5.40×10^{-2} | 0.07 |
| | March | 1.46×10^{-1} | 1.73×10^{-1} | 2.24×10^{-1} | 2.18×10^{-1} | 2.42×10^{-1} | 3.28×10^{-1} | 5.53×10^{-2} | 0.16 |
| | May | -2.51×10^{-1} | -2.42×10^{-1} | -2.20×10^{-1} | -2.06×10^{-1} | -1.77×10^{-1} | -1.12×10^{-1} | 4.59×10^{-2} | 0.09 |
| | June | 1.55×10^{-3} | 5.37×10^{-2} | 7.34×10^{-2} | 7.34×10^{-2} | 9.36×10^{-2} | 1.29×10^{-1} | 3.60×10^{-2} | 0.05 |
| | July | -3.12×10^{-2} | 1.78×10^{-3} | 1.74×10^{-2} | 2.63×10^{-2} | 4.52×10^{-2} | 1.22×10^{-1} | 4.75×10^{-2} | 0.45 |

| | | | | | | | | | |
|---|---|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-----------------------|-----------------|
| | August | -7.71x10 ⁻² | -5.90x10 ⁻² | -4.70x10 ⁻² | -2.75x10 ⁻² | 2.51x10 ⁻³ | 6.72x10 ⁻² | 5.10x10 ⁻² | 0.66 |
| | September | -1.03x10 ⁻¹ | -7.65x10 ⁻² | -6.73x10 ⁻² | -5.02x10 ⁻² | -3.99x10 ⁻² | 5.98x10 ⁻² | 4.65x10 ⁻² | 0.51 |
| | October | -1.41x10 ⁻¹ | -9.35x10 ⁻² | -5.36x10 ⁻² | -5.98x10 ⁻² | -3.56x10 ⁻² | 4.36x10 ⁻² | 5.73x10 ⁻² | 0.45 |
| | November | -2.39x10 ⁻¹ | -2.12x10 ⁻¹ | -1.55x10 ⁻¹ | -1.61x10 ⁻¹ | -1.21x10 ⁻¹ | -7.76x10 ⁻² | 5.61x10 ⁻² | 0.55 |
| | December | -4.55x10 ⁻¹ | -3.91x10 ⁻¹ | -3.29x10 ⁻¹ | -3.43x10 ⁻¹ | -2.98x10 ⁻¹ | -2.66x10 ⁻¹ | 6.08x10 ⁻² | 0.11 |
| | 2011 | 9.17x10 ⁻² | 1.03x10 ⁻¹ | 1.29x10 ⁻¹ | 1.34x10 ⁻¹ | 1.63x10 ⁻¹ | 1.87x10 ⁻¹ | 3.50x10 ⁻² | 0.01 |
| | 2012 | 1.31x10 ⁻¹ | 1.60x10 ⁻¹ | 1.89x10 ⁻¹ | 1.92x10 ⁻¹ | 2.14x10 ⁻¹ | 2.69x10 ⁻¹ | 4.19x10 ⁻² | 0.05 |
| | 2013 | 1.20x10 ⁻² | 5.52x10 ⁻² | 6.34x10 ⁻² | 6.88x10 ⁻² | 8.72x10 ⁻² | 1.33x10 ⁻¹ | 3.76x10 ⁻² | 0.02 |
| | 2014 | -1.11x10 ⁻² | 1.32x10 ⁻² | 3.29x10 ⁻² | 3.15x10 ⁻² | 5.32x10 ⁻² | 6.37x10 ⁻² | 2.66x10 ⁻² | 0.28 |
| | 2015 | 4.48x10 ⁻² | 5.95x10 ⁻² | 7.99x10 ⁻² | 7.69x10 ⁻² | 8.94x10 ⁻² | 1.07x10 ⁻¹ | 2.25x10 ⁻² | 0.66 |
| | 2016 | -4.73x10 ⁻¹ | -4.45x10 ⁻¹ | -4.29x10 ⁻¹ | -4.26x10 ⁻¹ | -4.13x10 ⁻¹ | -3.58x10 ⁻¹ | 3.46x10 ⁻² | 0.30 |
| | 2017 | -8.18x10 ⁻¹ | -8.12x10 ⁻¹ | -7.88x10 ⁻¹ | -7.83x10 ⁻¹ | -7.61x10 ⁻¹ | -7.26x10 ⁻¹ | 3.18x10 ⁻² | <0.01 |
| | Longitude | -3.26x10 ⁻² | -2.55x10 ⁻² | -2.24x10 ⁻² | -2.34x10 ⁻² | -2.01x10 ⁻² | -1.74x10 ⁻² | 4.99x10 ⁻³ | <0.01 |
| | Latitude | -1.87x10 ⁻¹ | -1.65x10 ⁻¹ | -1.59x10 ⁻¹ | -1.62x10 ⁻¹ | -1.57x10 ⁻¹ | -1.46x10 ⁻¹ | 1.24x10 ⁻² | <0.01 |
| Testing Practices Dependent Variable: Testing Frequency Intercept: 2.70 ± 3.66x10⁻³ | Message Received - No Significant Evidence ² | 5.03x10 ⁻² | 9.82x10 ⁻² | 1.29x10 ⁻¹ | 1.21x10 ⁻¹ | 1.50x10 ⁻¹ | 1.73x10 ⁻¹ | 3.97x10 ⁻³ | <0.01 |
| | Message Received - No Result ³ | 1.05x10 ⁻¹ | 1.77x10 ⁻¹ | 2.01x10 ⁻¹ | 1.89x10 ⁻¹ | 2.20x10 ⁻¹ | 2.39x10 ⁻¹ | 6.20x10 ⁻³ | <0.01 |
| | Message Received - Unsafe to Drink ⁴ | 4.01x10 ⁻³ | 1.50x10 ⁻² | 2.13x10 ⁻² | 1.95x10 ⁻² | 2.42x10 ⁻² | 3.06x10 ⁻² | 7.96x10 ⁻³ | 0.12 |
| | January | 2.41x10 ⁻¹ | 2.62x10 ⁻¹ | 2.66x10 ⁻¹ | 2.63x10 ⁻¹ | 2.70x10 ⁻¹ | 2.74x10 ⁻¹ | 1.09x10 ⁻² | <0.01 |

| | | | | | | | | |
|-----------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|-----------------------|-----------------|
| February | 2.22x10 ⁻¹ | 2.27x10 ⁻¹ | 2.32x10 ⁻¹ | 2.33x10 ⁻¹ | 2.41x10 ⁻¹ | 2.44x10 ⁻¹ | 8.42x10 ⁻³ | <0.01 |
| March | 3.74x10 ⁻³ | 1.22x10 ⁻² | 1.27x10 ⁻² | 1.31x10 ⁻² | 1.65x10 ⁻² | 2.23x10 ⁻² | 5.54x10 ⁻³ | 0.27 |
| May | 8.27x10 ⁻² | 9.27x10 ⁻² | 9.60x10 ⁻² | 9.54x10 ⁻² | 1.00x10 ⁻¹ | 1.05x10 ⁻¹ | 6.66x10 ⁻³ | <0.01 |
| June | 9.36x10 ⁻² | 9.97x10 ⁻² | 1.03x10 ⁻¹ | 1.03x10 ⁻¹ | 1.06x10 ⁻¹ | 1.13x10 ⁻¹ | 6.17x10 ⁻³ | <0.01 |
| July | 1.14x10 ⁻¹ | 1.17x10 ⁻¹ | 1.18x10 ⁻¹ | 1.19x10 ⁻¹ | 1.21x10 ⁻¹ | 1.29x10 ⁻¹ | 4.43x10 ⁻³ | <0.01 |
| August | 1.10x10 ⁻¹ | 1.15x10 ⁻¹ | 1.18x10 ⁻¹ | 1.19x10 ⁻¹ | 1.22x10 ⁻¹ | 1.30x10 ⁻¹ | 6.09x10 ⁻³ | <0.01 |
| September | 6.73x10 ⁻² | 7.27x10 ⁻² | 7.46x10 ⁻² | 7.60x10 ⁻² | 7.97x10 ⁻² | 8.42x10 ⁻² | 5.37x10 ⁻³ | <0.01 |
| October | 2.98x10 ⁻² | 4.02x10 ⁻² | 4.15x10 ⁻² | 4.12x10 ⁻² | 4.41x10 ⁻² | 4.98x10 ⁻² | 5.96x10 ⁻³ | <0.01 |
| November | 1.45x10 ⁻² | 1.92x10 ⁻² | 2.23x10 ⁻² | 2.37x10 ⁻² | 2.87x10 ⁻² | 3.63x10 ⁻² | 6.98x10 ⁻³ | 0.05 |
| December | 1.79x10 ⁻² | 2.98x10 ⁻² | 3.55x10 ⁻² | 3.31x10 ⁻² | 3.95x10 ⁻² | 4.25x10 ⁻² | 8.45x10 ⁻³ | 0.03 |
| 2011 | 3.21x10 ⁻² | 3.52x10 ⁻² | 3.75x10 ⁻² | 3.69x10 ⁻² | 3.84x10 ⁻² | 4.23x10 ⁻² | 3.17x10 ⁻³ | <0.01 |
| 2012 | -1.60x10 ⁻² | -9.08x10 ⁻³ | -5.56x10 ⁻³ | -6.68x10 ⁻³ | -3.86x10 ⁻³ | 2.66x10 ⁻⁴ | 5.00x10 ⁻³ | 0.51 |
| 2013 | 8.23x10 ⁻³ | 1.41x10 ⁻² | 1.58x10 ⁻² | 1.59x10 ⁻² | 1.78x10 ⁻² | 2.22x10 ⁻² | 3.90x10 ⁻³ | 0.11 |
| 2014 | -1.03x10 ⁻² | -6.26x10 ⁻³ | -5.18x10 ⁻³ | -4.99x10 ⁻³ | -3.66x10 ⁻³ | 3.88x10 ⁻⁵ | 3.28x10 ⁻³ | <0.01 |
| 2015 | -1.52x10 ⁻¹ | -1.49x10 ⁻¹ | -1.48x10 ⁻¹ | -1.48x10 ⁻¹ | -1.47x10 ⁻¹ | -1.44x10 ⁻¹ | 2.23x10 ⁻³ | <0.01 |
| 2016 | -2.34x10 ⁻¹ | -2.30x10 ⁻¹ | -2.27x10 ⁻¹ | -2.28x10 ⁻¹ | -2.26x10 ⁻¹ | -2.23x10 ⁻¹ | 3.22x10 ⁻³ | <0.01 |
| 2017 | -9.93x10 ⁻¹ | -9.86x10 ⁻¹ | -9.85x10 ⁻¹ | -9.85x10 ⁻¹ | -9.82x10 ⁻¹ | -9.80x10 ⁻¹ | 3.84x10 ⁻³ | <0.01 |

¹Bolded p-values represent significant variables

²”no significant evidence, no evidence of *E. coli* and Total Coliforms are ≤ 5 CFU/100mL

³“no result” provided due “to process issues, chemical testing, appearance or odour, interfering substances, or unauthorized submitter

⁴“unsafe to drink”, observations contain *E. coli* (Category 1-3) or a sample is overgrown, either with or without evidence of total coliforms or *E. coli*, the presence of *E. coli* cannot be guaranteed or quantified, but also cannot be ruled out

Table A.2-6: Statistical significance testing for bottom stratigraphy analysis comparing unconsolidated and consolidated categorization. Statistical significance is set to $p \leq 0.05$.

| Category | Test-Statistic | p-value |
|------------------------------|----------------|----------|
| non-detect | 23.90 | < 0.0001 |
| 1-10 CFU/100mL (Category 1) | -22.83 | < 0.0001 |
| 11-50 CFU/100mL (Category 2) | -7.91 | < 0.0001 |
| 51+ CFU/100mL (Category 3) | -2.60 | 0.0093 |

Table A.2-7: Top ten “most interesting” rules based on an association rule analysis including adverse *E. coli* concentrations and rock type. Where Category 1 is 1-10 CFU/100mL, Category 2 is 11-50 CFU/100mL, Category 3 is 51+ CFU/100mL, and water is “Unsafe to Drink (target)”.

| Rule | LHS | RHS | St. Lift | Support | Confidence | Lift |
|------|--|-------------------|----------|---------|------------|------|
| 1 | Igneous = No, Sedimentary = No | Metamorphic = Yes | 1.00 | 0.01 | 1.00 | 77.4 |
| 2 | Igneous = No, Metamorphic = No | Sedimentary = Yes | 1.00 | 0.70 | 1.00 | 1.4 |
| 3 | <i>E. coli</i> = Category 3, Igneous = No, Metamorphic = No | Sedimentary = Yes | 1.00 | 0.10 | 1.00 | 1.4 |
| 4 | <i>E. coli</i> = Category 1, Igneous = No, Metamorphic = No | Sedimentary = Yes | 1.00 | 0.10 | 1.00 | 1.4 |
| 5 | <i>E. coli</i> = Category 2, Igneous = No, Metamorphic = No | Sedimentary = Yes | 1.00 | 0.10 | 1.00 | 1.4 |
| 6 | <i>E. coli</i> = “unsafe to drink (target), Igneous = No, Metamorphic = No | Sedimentary = Yes | 1.00 | 0.21 | 1.00 | 1.4 |
| 7 | Sedimentary = No, Metamorphic = No | Igneous = Yes | 1.00 | 0.28 | 1.00 | 3.4 |
| 8 | <i>E. coli</i> = Category 3, Metamorphic = No, Sedimentary = No | Igneous = Yes | 1.00 | 0.03 | 1.00 | 3.4 |
| 9 | <i>E. coli</i> = Category 1, Metamorphic = No, Sedimentary = No | Igneous = Yes | 1.00 | 0.04 | 1.00 | 3.4 |
| 10 | <i>E. coli</i> = Category 2, Metamorphic = No, Sedimentary = No | Igneous = Yes | 1.00 | 0.04 | 1.00 | 3.5 |

Table A.2-8: Statistical significance testing for well depth analysis comparing shallow and deep wells. Statistical significance is set to $p \leq 0.05$. Where non-detects are defined as 0 CFU/100mL, category 1 as 1-10 CFU/100mL, category 2 as 11-50 CFU/100mL, and category 3 as 51+ CFU/100mL.

| Category | Test-Statistic | p-value |
|------------------------------|----------------|----------|
| non-detect | -8.56 | < 0.0001 |
| 1-10 CFU/100mL (Category 1) | 6.97 | < 0.0001 |
| 11-50 CFU/100mL (Category 2) | 3.50 | 0.0005 |
| 51+ CFU/100mL (Category 3) | 3.74 | 0.0002 |

Table A.2-9: Summary of user testing decay curves based on first test message received. Two decay equations are summarized per first test message received, curve characteristics for total tests 1 through 15, and curve characteristics for total tests 16+.

| | n | Decay Equation | y_f | y_o | log alpha | Standard Error | Alpha (decay rate) | Resid. Sum of Squares |
|-------------------------|---------|------------------|-----------------------|-----------------------|------------------------|-----------------------|-----------------------|-----------------------|
| No Significant Evidence | 183,608 | Under 16 Tests | 2.62×10^{-3} | 1.39 | -3.70×10^{-2} | 2.67×10^{-2} | 9.64×10^{-1} | 2.13×10^{-3} |
| | | 16 or More Tests | 2.38×10^{-5} | 1.43×10^{-2} | -2.13 | 1.73×10^{-2} | 1.18×10^{-1} | 1.81×10^{-7} |
| No Result | 15,816 | Under 16 Tests | 4.21×10^{-4} | 5.11×10^{-1} | -8.60×10^{-1} | 3.10×10^{-2} | 4.23×10^{-1} | 2.76×10^{-3} |
| | | 16 or More Tests | 1.24×10^{-4} | 3.33×10^{-2} | -1.83 | 8.41×10^{-2} | 1.61×10^{-1} | 1.54×10^{-6} |
| May Be Unsafe | 31,656 | Under 16 Tests | 4.82×10^{-4} | 4.74×10^{-1} | -9.14×10^{-1} | 1.26×10^{-2} | 4.01×10^{-1} | 3.68×10^{-8} |
| | | 16 or More Tests | 5.24×10^{-5} | 3.31×10^{-2} | -1.89 | 4.09×10^{-2} | 1.51×10^{-1} | 6.71×10^{-7} |
| Unsafe to Drink | 20,341 | Under 16 Tests | 4.10×10^{-4} | 4.80×10^{-1} | -9.14×10^{-1} | 2.09×10^{-2} | 4.01×10^{-1} | 1.17×10^{-3} |
| | | 16 or More Tests | 4.61×10^{-5} | 1.68×10^{-2} | -2.10 | 8.14×10^{-2} | 1.22×10^{-1} | 1.37×10^{-6} |

Table A.2-10: Exhaustive list of categorical bottom stratigraphy variable levels. Green represents present, red represents absent.

| Label | Metamorphic | Igneous | Sedimentary | Fine | Medium | Coarse |
|---------|-------------|---------|-------------|------|--------|--------|
| FFFFFF | | | | | | |
| FFFFFFT | | | | | | |
| FFFFFTF | | | | | | |
| FFFFFTT | | | | | | |
| FFFTFF | | | | | | |
| FFFTFT | | | | | | |
| FFFTTF | | | | | | |
| FFFTTT | | | | | | |
| FFTFFF | | | | | | |
| FFTFFT | | | | | | |
| FFTFTF | | | | | | |
| FFTFTT | | | | | | |
| FFTTF | | | | | | |
| FFTTF | | | | | | |
| FFTTF | | | | | | |
| FFTTF | | | | | | |
| FTFFFF | | | | | | |
| FTFFFT | | | | | | |
| FTFFTF | | | | | | |
| FTFFTT | | | | | | |
| FTFTFF | | | | | | |
| FTFTFT | | | | | | |
| FTFTTF | | | | | | |
| FTTFFF | | | | | | |
| FTTFFT | | | | | | |
| FTTFTF | | | | | | |
| FTTTFF | | | | | | |
| FTTTTF | | | | | | |
| TFFFFFF | | | | | | |
| TFFFFT | | | | | | |
| TFFFTF | | | | | | |
| TFFFTT | | | | | | |
| TFFTFF | | | | | | |
| TFFTTF | | | | | | |
| TFFTTT | | | | | | |
| TFTFFF | | | | | | |
| TFTFFT | | | | | | |
| TFTF | | | | | | |
| TFTTFF | | | | | | |
| TTFFFT | | | | | | |
| TTFFTF | | | | | | |

| | | | | | | |
|--------|--|--|--|--|--|--|
| TTFFFF | | | | | | |
| TTFTFF | | | | | | |
| TTTFFF | | | | | | |

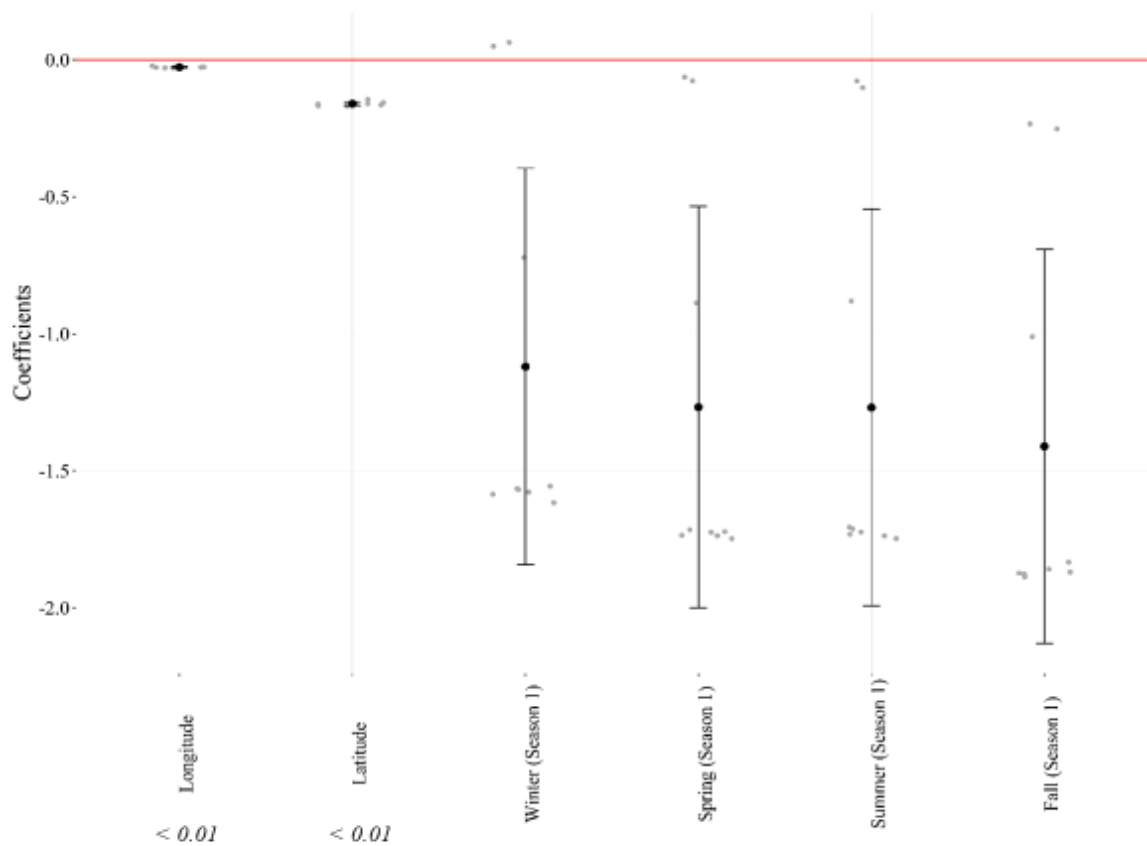


Figure A.2-1: Sensitivity analysis of "most explanatory" seasonality variables considering seasons. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent one standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant.

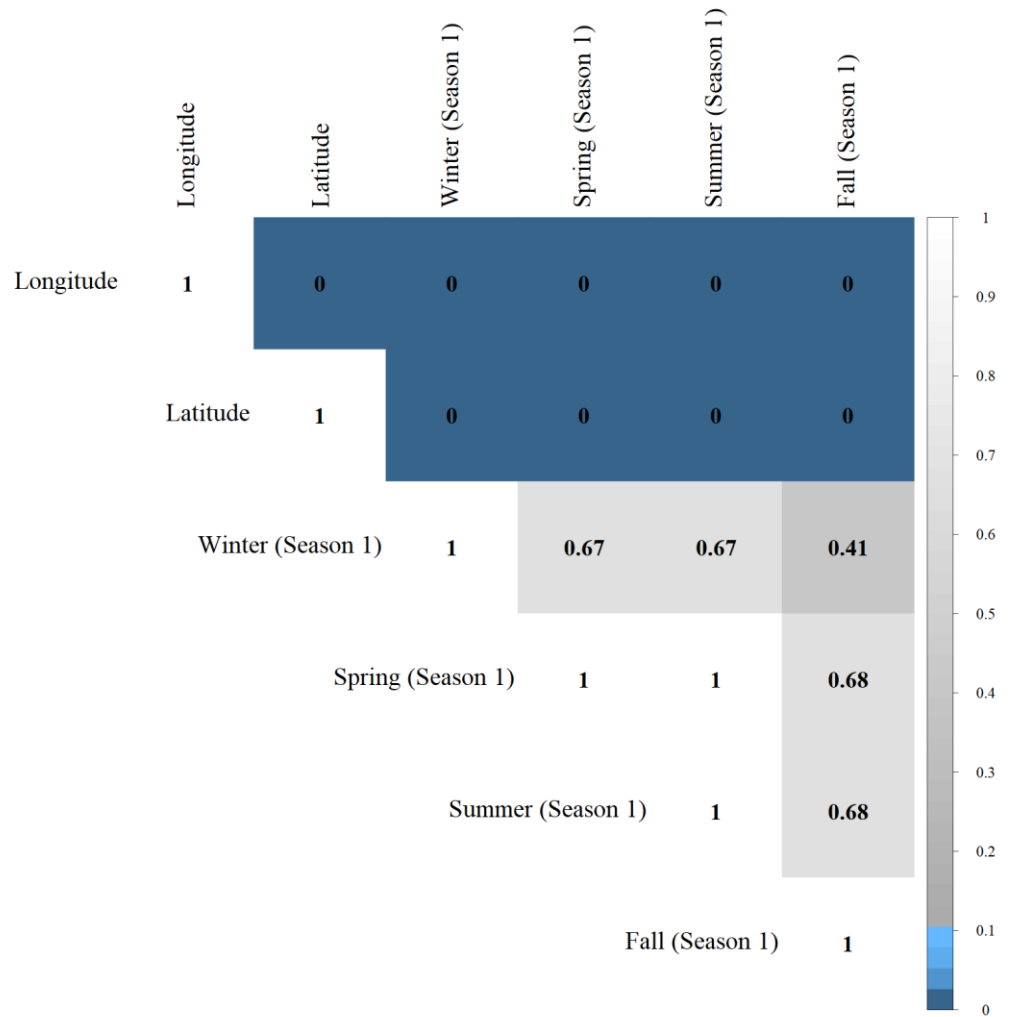


Figure A.2-2: Statistical significance of all seasonality variables in the seasonal assessment model when compared to one another. Statistically significant is defined as $p \leq 0.1$.

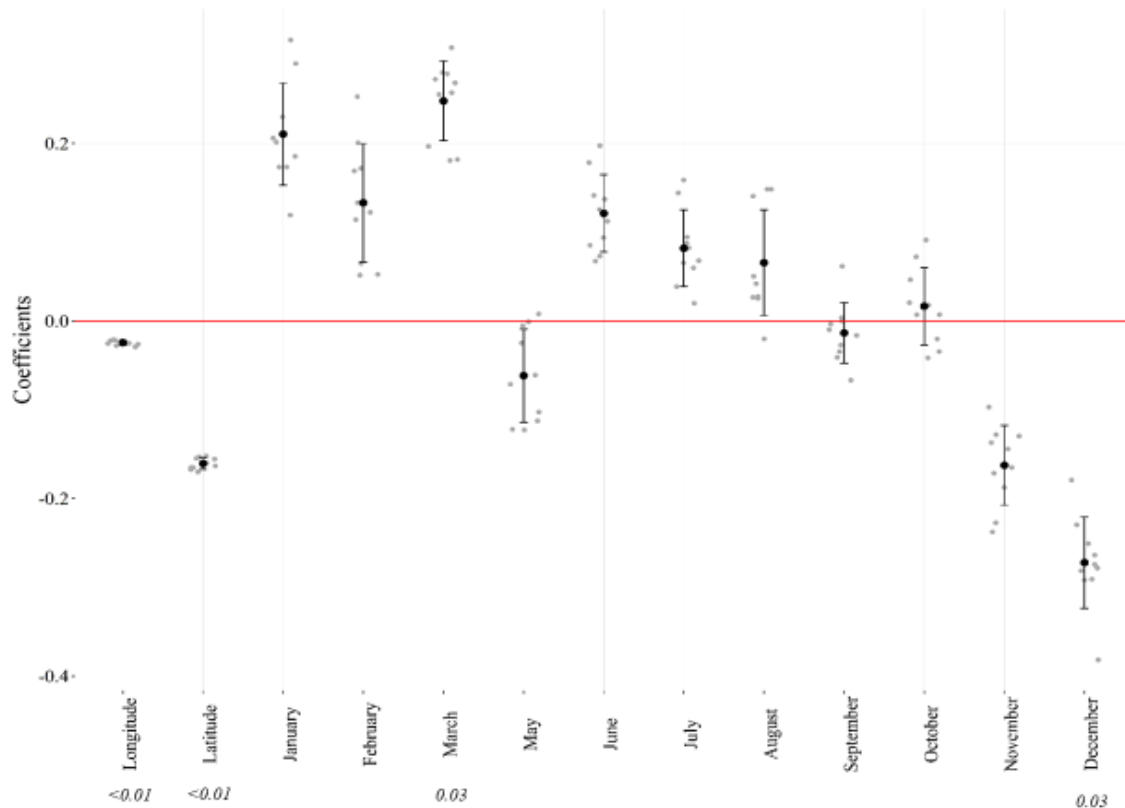


Figure A.2-3: Sensitivity analysis of "most explanatory" seasonality variables considering months. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. Note the absence of April, this was removed as it was deemed insignificant by the models.

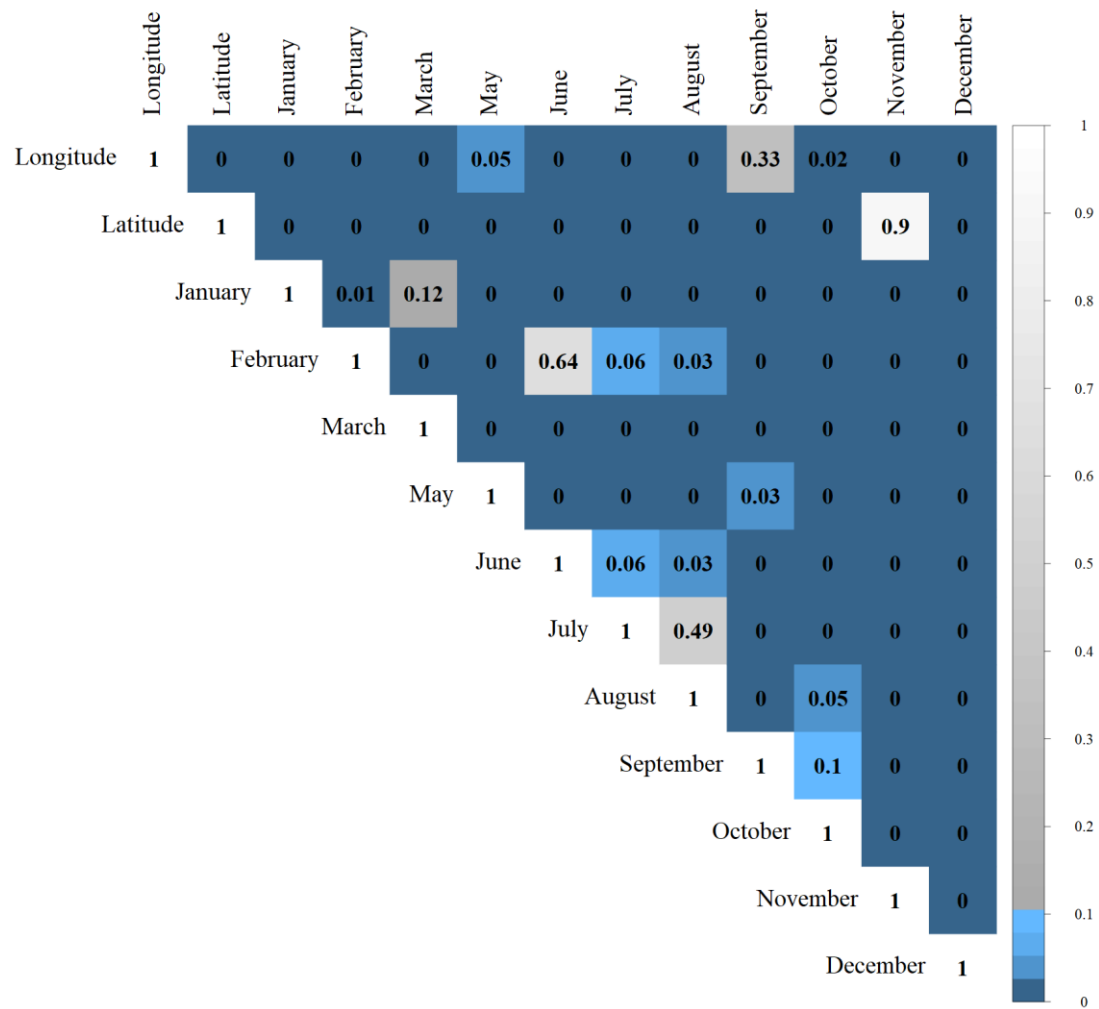


Figure A.2-4: Statistical significance of all seasonality variables in the monthly assessment model when compared to one another. Statistical significance is set to $p \leq 0.1$.

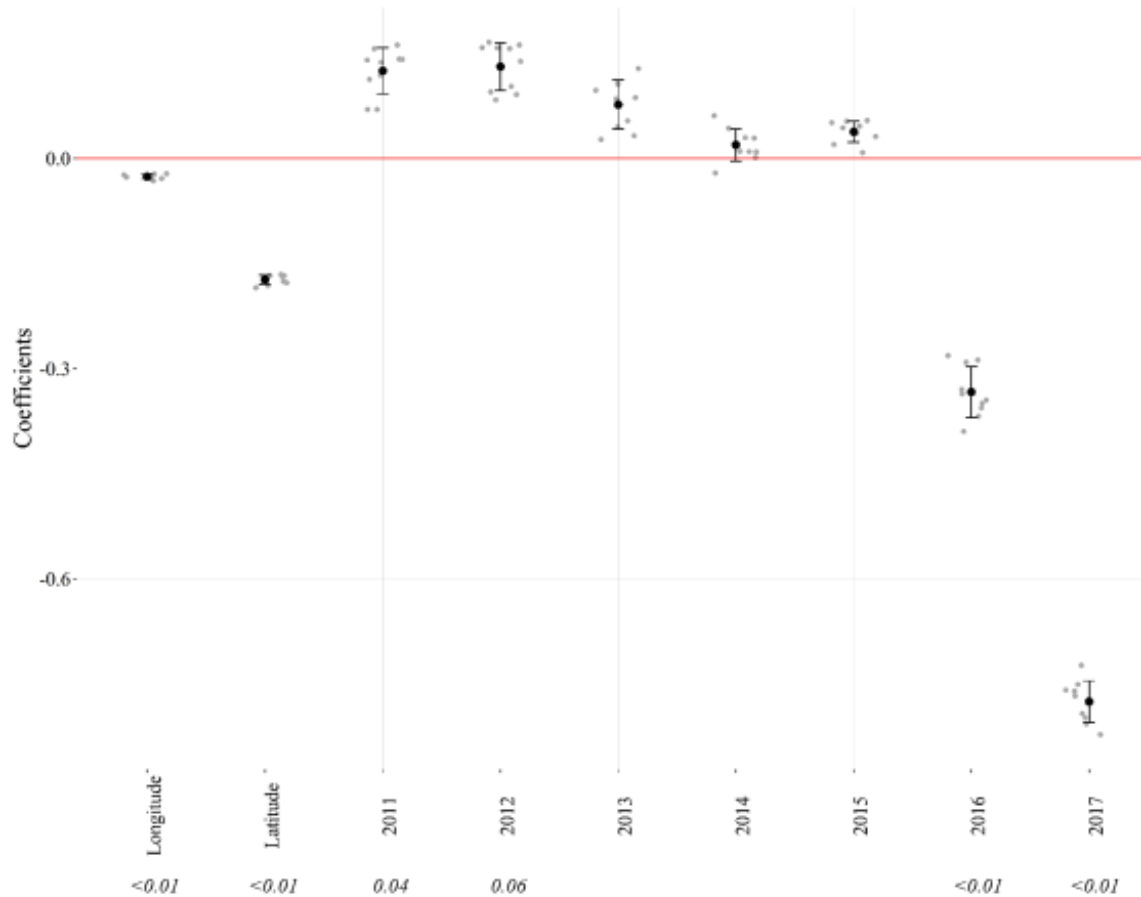


Figure A.2-5: Sensitivity analysis of "most explanatory" seasonality variables considering years. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant. Note the absence of 2010, this was not considered explanatory in the models.

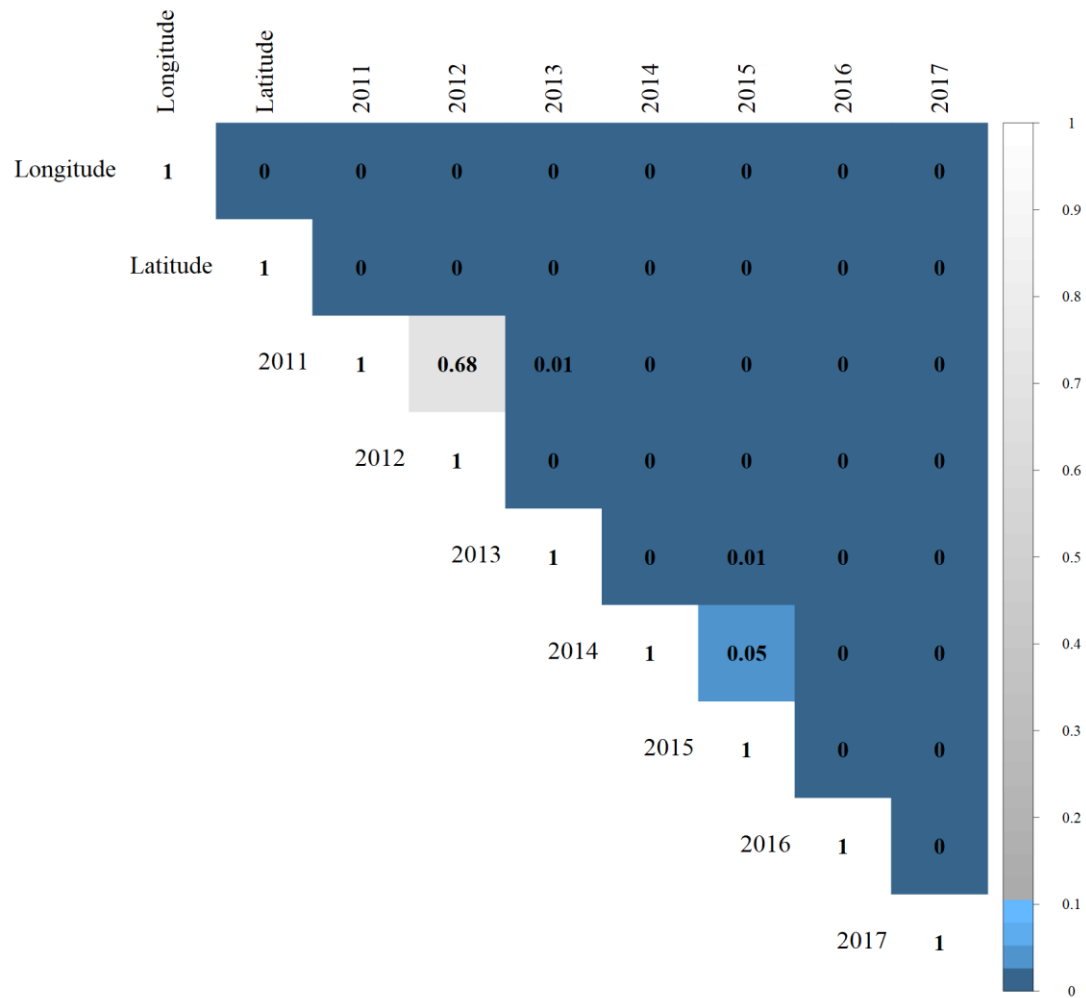


Figure A.2-6: Statistical significance of all seasonality variables found within the yearly assessment model when compared to one another. Statistical significance is set to $p \leq 0.1$.

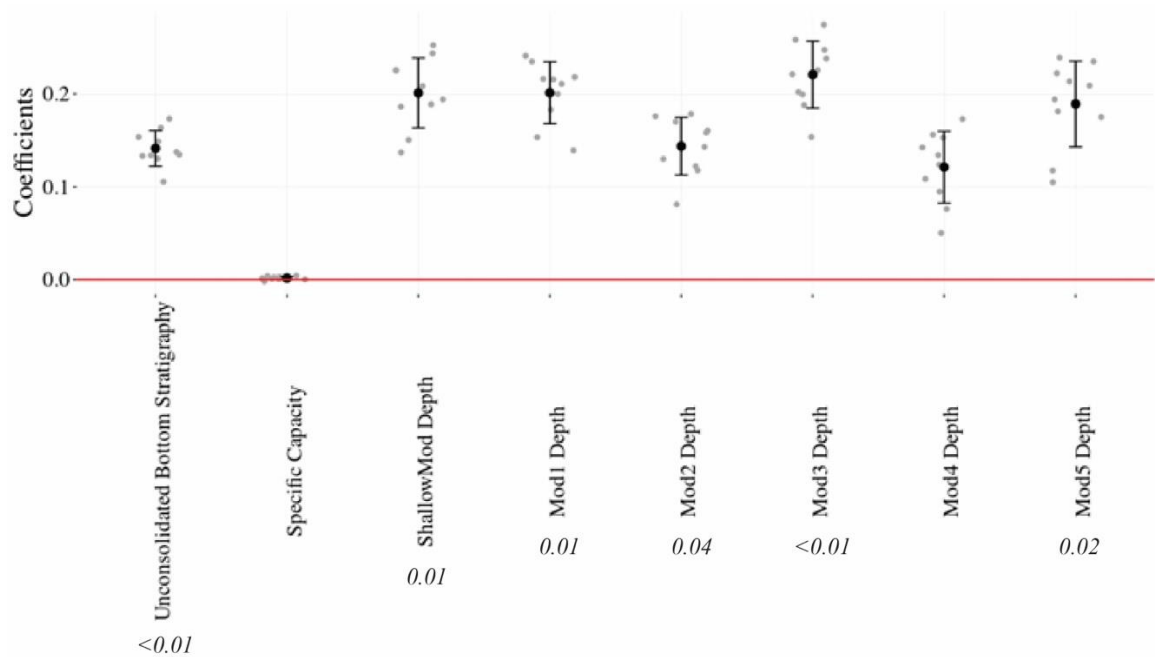


Figure A.2-7: Sensitivity analysis of "most explanatory" hydrogeological variables. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant.

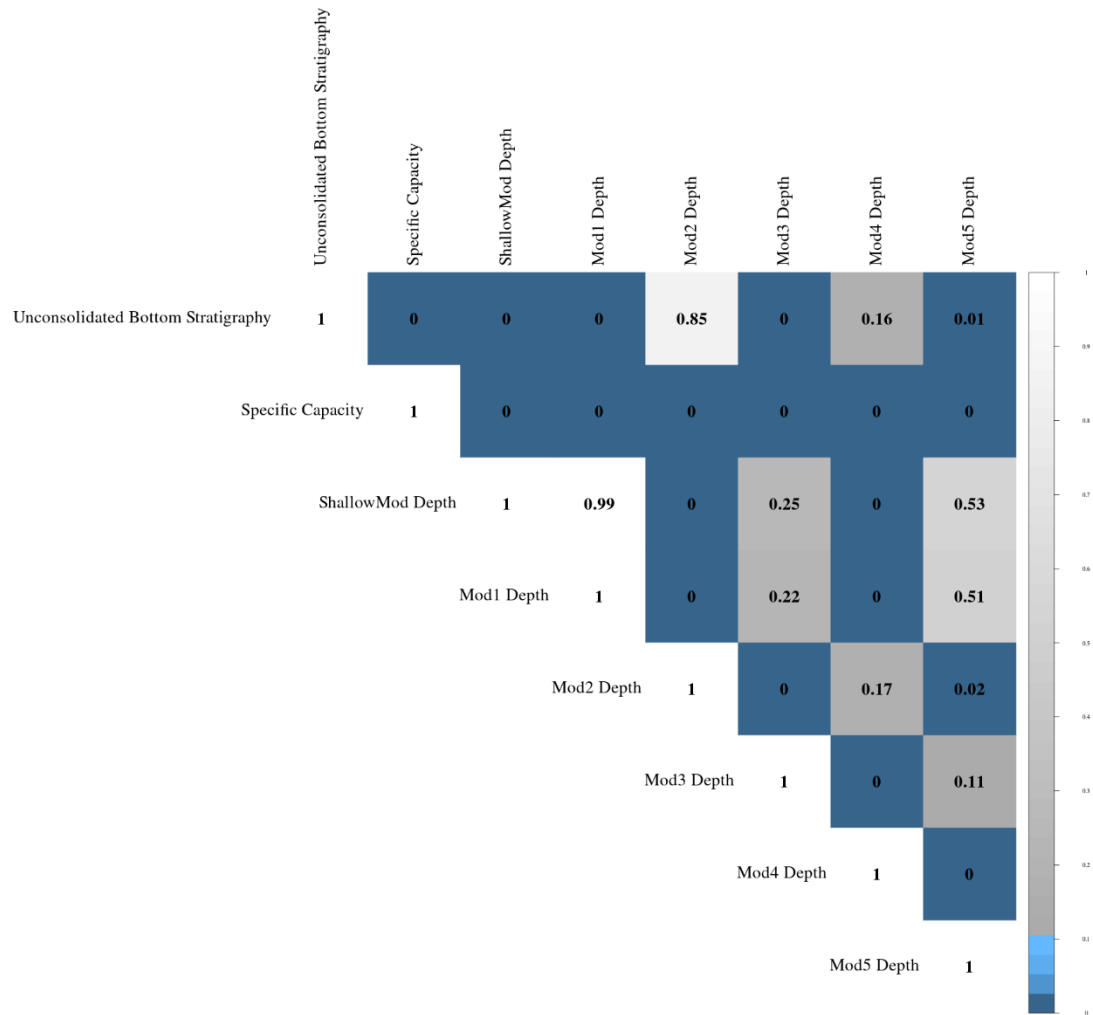


Figure A.2-8: Statistical significance of all hydrogeological variables when compared to one another. Statistical significance is set to $p \leq 0.1$.

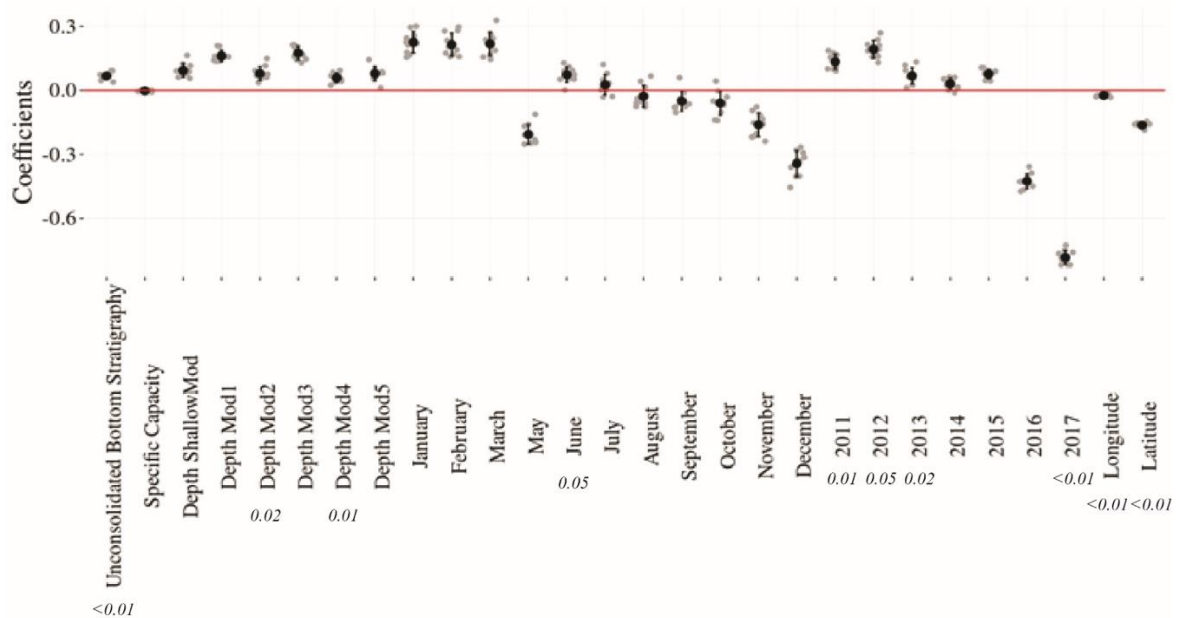


Figure A.2-9: Sensitivity analysis of "most explanatory" informed physical variables. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant.

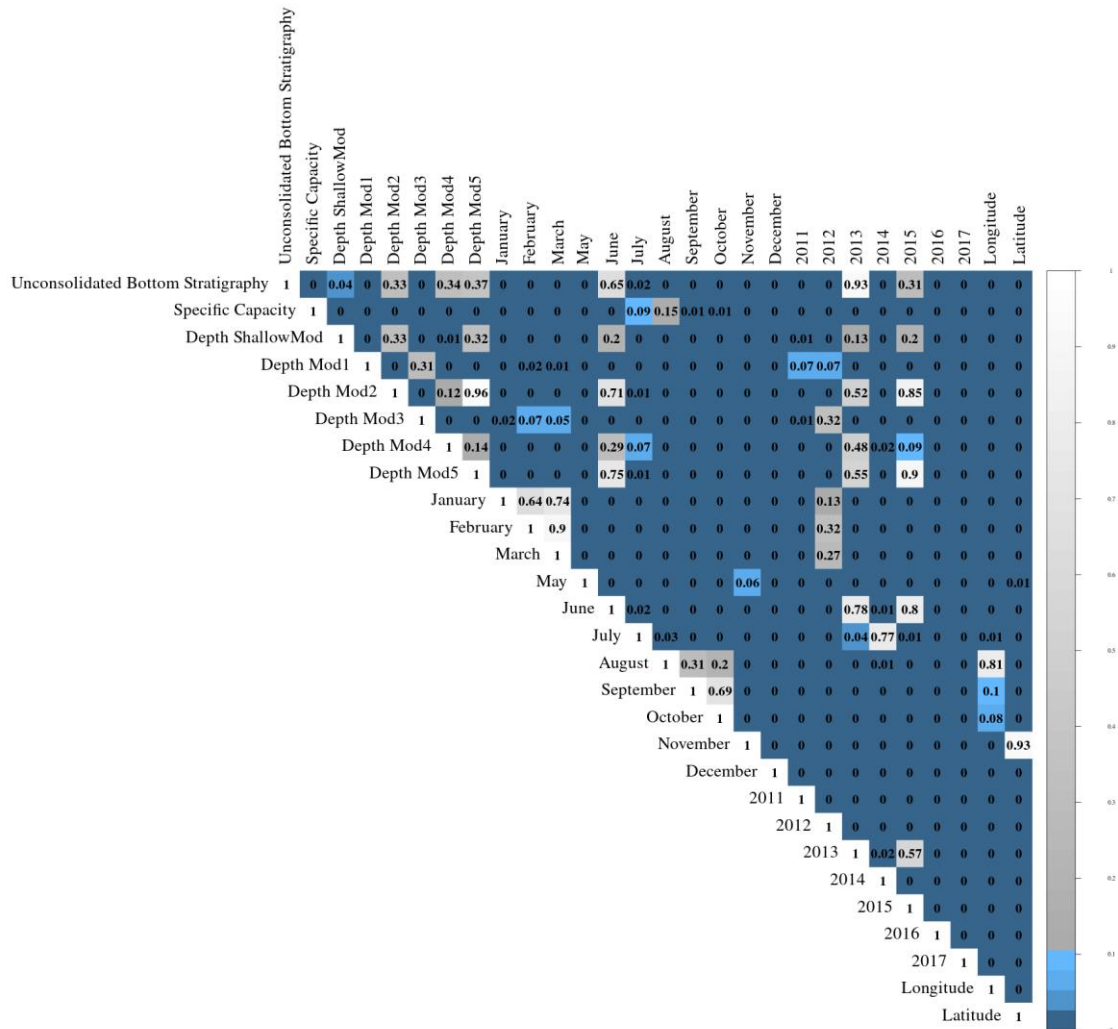


Figure A.2-10: Statistical significance of all informed physical variables when compared to one another. Statistical significance is set to $p \leq 0.1$.

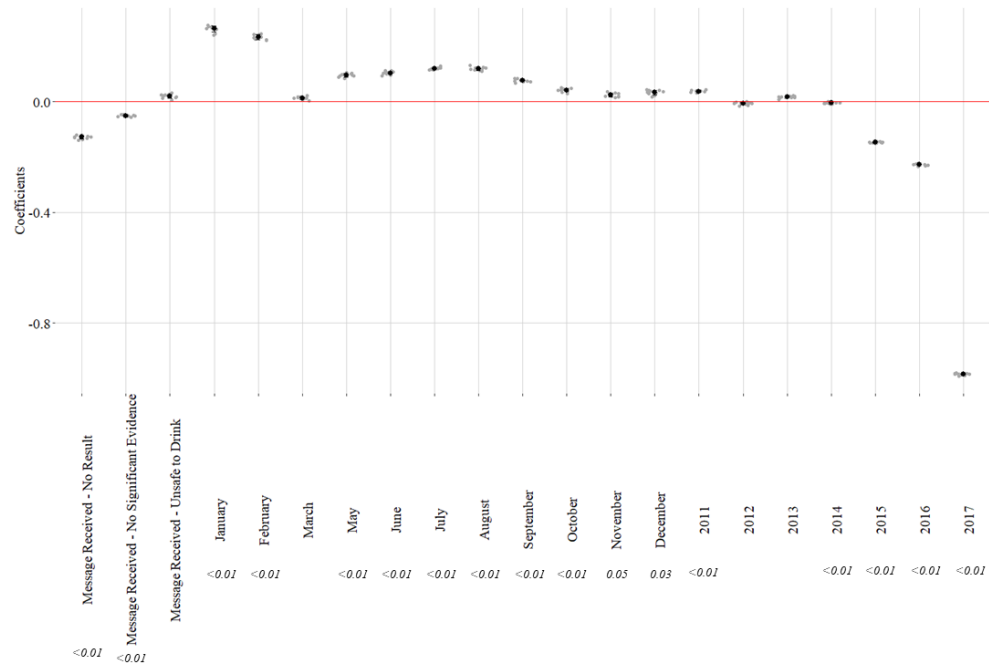


Figure A.2-11: Sensitivity analysis of "most explanatory" well testing variables. Red horizontal line represents a coefficient of zero, black points represent the mean of the variable coefficient, black bars represent the standard deviation, and grey points represent jittered data points. Level of significance listed below variable names where they are significant.

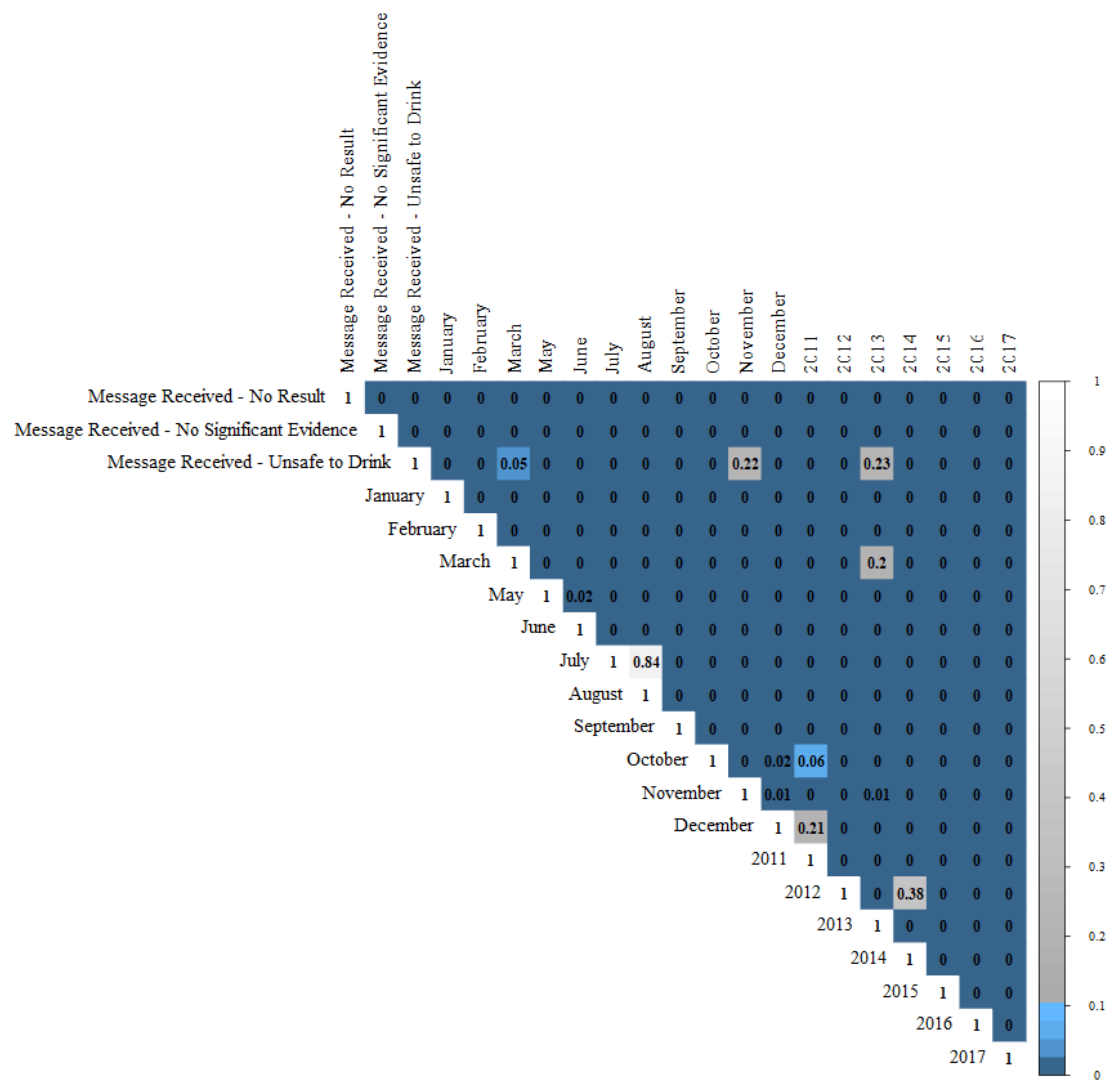


Figure A.2-12: Statistical significance of all testing practice variables when compared to one another. Statistical significance is set to $p \leq 0.1$.

A.2.3 Works Cited

Bain, R., Cronk, R., Hossain, R., Bonjour, S., Onda, K., Wright, J., Yang, H., Slaymaker, T., Hunter, P., Prüss-Ustün, A., Bartram, J., 2014. Global assessment of exposure to faecal contamination through drinking water based on a systematic review. *Trop. Med. Int. Heal.* 19, 917–927. <https://doi.org/10.1111/tmi.12334>

Freeze, R.A., Cherry, J.A., 1978. *Groundwater*.




Ontario Agency for Health Protection and Promotion, 2019. *Public Health Inspector's Guide to Environmental Microbiology Laboratory Testing*.




Public Health Ontario, 2020. Well Water Testing (Private Drinking Water) | Public Health Ontario [WWW Document]. URL <https://www.publichealthontario.ca/en/laboratory-services/well-water-testing?tab=4> (accessed 11.2.20).

A.3 CHAPTER 3 APPENDICES

A.3.1 Tables and Figures

Table A.3-1: Tables adapted from Caldeira et al. (2019) and Mastrandrea et al. (2010). A – agreement is the qualitative measure of the level of concurrence in the literature, B – evidence is the amount of work supporting a finding, C and D– level of confidence is determined by a summary of the evidence and agreement levels, where darker blocks (HA&RE) represent a higher confidence than lighter blocks (LA&LE) (Caldeira et al., 2019; Mastrandrea et al., 2010).

| A - Agreement | |
|---------------|---|
| High |  |
| Medium |  |
| Low |  |

| B - Evidence | | Number of Papers/Reports |
|--------------|---|--------------------------|
| Robust |  | 4+ |
| Medium |  | 2-3 |
| Limited |  | 1 |

| C – Level of Confidence | | |
|--------------------------------------|-------------------------------------|-------------------------------------|
| High Agreement Limited Evidence | High Agreement Medium Evidence | High Agreement Robust Evidence |
| Medium Agreement Limited Evidence | Medium Agreement Medium Evidence | Medium Agreement Robust Evidence |




| D - Categories | Confidence Symbols | Likelihood and Agreement Combination |
|----------------|--------------------|--------------------------------------|
| Very High | ★★★★ | HA&RE |
| High | ★★★ | HA&ME; MA&RE |

| | | |
|--------------------------------------|-------------------------------------|----------------------------------|
| Low Agreement Limited Evidence | Low Agreement Medium Evidence | Low Agreement Robust Evidence |
|--------------------------------------|-------------------------------------|----------------------------------|

| | | |
|----------|----|------------------------|
| Medium | ★★ | ME&MA; HA&LE; LA&RE |
| Low | ★ | MA&LE; LA&ME |
| Very Low | ★ | LA&LE |

Table A.3-2: Summary table comparing ordered LULCs based on impact to *E. coli* contamination potential utilizing regression analyses and corresponding literature. Icons in “Key Process Notes” column are defined in Table A.3-1 to depict likelihood, evidence, and level of confidence of findings based on regression analyses and literature.

| Defined for Current Study (Based on Regression Analyses) | | | Literature Summaries | | | |
|---|---|--|---|---|---|--|
| Land Cover Category (Impact on <i>E. coli</i> Contamination Potential) | Category Description | Key Process Notes | Category | Category Description and Study Type | Process Description | Impact on <i>E. coli</i> Contamination Potential |
| Pastoral/ Agricultural (Very High) | Continuous row crops and mixed crops are rotated with perennial crops. Also includes hay/pasture, orchards, vineyards, nurseries, rural | <ul style="list-style-type: none"> High <i>E. coli</i> loading from domestic animals and land-based application of manure High potential for connectivity to and transport | Agriculture (Paule-Mercado et al., 2016) | Cultivation, no livestock farms exist, low impervious percentage, dry and paddy fields Field – Stormwater Runoff | Domestic and wild animal faeces, sewer and septic cross connection leakages and overflow. Rice paddies receive intensive application of | Moderate |




| | | | | | | |
|--|--|---|------------------------------------|--|---|-----------|
| | properties not in production, urban brown fields, and clearings within forests. | through groundwater as often well irrigated and, in some cases, actively tilled | | | fertilizers, fields then flooded to facilitate increase in nutrient uptake by rice plants; large precipitation event can cause overland flow. | |
| | <div>    </div> | | Cultivated (Namugize et al., 2018) | NA Note: study conducted in lower-income area of South Africa | Application of fertilizers | Moderate |
| | | | Agriculture (Jabbar et al., 2019) | Field - River Pasture/Hay (for livestock grazing or production of seed or hay crops) and cultivated crops (land being actively tilled, orchards, vineyards, annual crops) | Excess nutrients through fertilizers and manure, runoff from pesticide and herbicides, and increased | Very High |




| | | | | |
|--|--|---|---|--|
| | | Field - River | turbidity level due to sedimentation and soil erosion. Note: this mainly impacts N and P. | |
| | | <p>Cropland/ Pasture (Dusek et al., 2018)</p> <p>Pasture/Hay (for livestock grazing or production of seed or hay crops) and cultivated crops (land being actively tilled, orchards, vineyards, annual crops)</p> <p>Field - Soils</p> | <p>Pasture - Migration of <i>E. coli</i> into the soil at higher frequency due to more frequent fecal deposition, lower selection pressure leading to enhanced survival (or growth). Cropland – manured croplands, <i>E. coli</i> more prevalent in</p> | <p>Very High (Pasture)</p> <p>Low (Cropland)</p> |

| | | | |
|--|--|---|---|
| | | | saturated soils, high levels of diurnal and anthropogenic variation under soil conditions may lead to lower prevalence. |
| | Pasture (Elliott et al., 2016) | Arable crops, vegetables, potatoes, kiwifruit, pip fruit, viticulture, vineyards, hill and intensive sheep/beef pasture | NA Moderate |
| | Agricultural Land (Petersen and Hubbart, 2020) | Field – River/Model Definition Actively grazed pasture, dairy cattle grazing pastures, holding pens, and livestock manure stacks | NA High |
| | | Literature Review | |

| | | | | |
|--|--|--|--|----------|
| | | <p>Agriculture (Gregory et al., 2019)</p> <p>Ungrazed rangeland – annual mowing or haying Managed hay pasture – seasonal haying, ungrazed currently Cultivated cropland – conventional tillage, commercial fertilizer, ungrazed</p> <p>Field - Soils</p> | <p>Cattle from nearby pastures and pets from neighboring properties unexpectedly found in watersheds. Avian wildlife found more in cultivated cropland and hay pasture (vs. native prairie).</p> | Moderate |
| | | <p>Agriculture (Hua, 2017)</p> <p>Agriculture activities (not clearly defined, assumed to be mainly cropland, potentially cattle/poultry)</p> <p>Field - River</p> | <p><i>E. coli</i> strongly connected to raw and municipal sewage from domestic and poultry farms. Farms with high fertilizer usage causes <i>E. coli</i></p> | Moderate |




| | | | | | | |
|--------------------------------------|--|--|-----------------------------------|---|---------------|--|
| | | | | | | presence; indirectly, agriculture activities could disrupt the soil structure and cause <i>E. coli</i> presence in rivers. |
| | | | | | | High wastewater from non-point sources, i.e., tile drains and ditches. |
| | | | Pastoral (Larned et al., 2004) | e.g., Dairy farming, row cropping (i.e., orchards, vineyards, short rotation) | Field - River | High |
| Open/ Bareland (High) | Unconsolidated materials subject to active processes (i.e., wave energy, erosion, etc.). Containing <25% tree or shrub coverage. | <ul style="list-style-type: none"> • <i>E. coli</i> loading from wildlife • Potential for higher groundwater infiltration due to “unconsolidated materials”, i.e., minimal natural attenuation | Barren Land (Jabbar et al., 2019) | Areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits and other accumulations of earthen material. Generally, vegetation <15% total cover | NA | Low/Moderate |







| | | | | | |
|--------------|--|---|--------------------------------|---|----------|
| | | <ul style="list-style-type: none"> High runoff potential    | Field - River | | |
| | | | Open Space (Hua, 2017) | <p>Including all land areas that exposed soil and barren areas are influenced by humans. Described as transition area for built-up area converted from agriculture, as well as forest into agricultural activities</p> <p>Risk due to hydrologic modifications such as dredging, water diversion, and channelization.</p> | Moderate |
| Urban (High) | <p>Highways, roads, linear frequencies of structures more than 10 per 500m or 4 per 1-hectare box. Green space or permeable surfaces can be present.</p> | <ul style="list-style-type: none"> Potential <i>E. coli</i> loading from overflowing sewers, wastewater treatment facilities, leaking pipes or septic tanks, and domestic animals | Field - River | | |
| | | | Urban (Larned et al., 2004) | <p>Including manufacturing, urban housing, industrial, roads</p> <p>Field - River</p> <p>Characterized by high surface runoff and periodically high stormwater discharge. High wastewater input from non-point sources: storm drains</p> | High |


| | | | | |
|--|--|---|---|-----------|
| | <ul style="list-style-type: none"> • Lower potential for connectivity to groundwater due to high levels of impermeable surfaces • High surface runoff <div>    </div> | | and leaking sewers; lower interception and lower infiltration rates. Highlighted positive trends in conductivity. | |
| | <div> Urban/built-up (Namugize et al., 2018) </div> | <div> NA Note: study conducted in lower-income area of South Africa Field - River </div> | Increased <i>E. coli</i> linked to rapid urbanization, associated with expansion of informal settlements, stray livestock, poor sanitary systems. | High |
| | <div> Ground (Paule-Mercado et al., 2016) </div> | <div> Parking lot, residential, road, and commercial. High impervious percentage </div> | Sources of fecal contamination include domestic animal feces, sewer and | Very High |

| | | | | |
|--|--|--|---|----------|
| | | Field – Surface Water Runoff | septic cross connection leakages and overflow. Impervious Cover may be prone to fecal indicator bacteria from runoff. | |
| | Developed (Petersen and Hubbart, 2020) | Urban (commercial area) and residential Literature Review | Potential artificial sources of <i>E. coli</i> (i.e., water infrastructure), increased runoff. | Moderate |
| | Urban (Jabbar et al., 2019) | Containing anything from open space (<20% impervious; large-lot homes, golf courses), to single-family homes (20-49% impervious), to smaller-lot homes (50-79% impervious), to apartment | Significant load of contaminants from the point and non-point sources. Originates from waste produced by municipal wastewater | High |

| | | | | |
|--|--|--|---|------|
| | | complexes and commercial/industrial (80-100% impervious) | treatment plants and undefined anthropogenic sources. | |
| | | Field - River | | |
| | Artificial Surfaces (Urban) (Elliott et al., 2016) | Field – River/Model Definition | NA | High |
| | Urban (Paule-Mercado et al., 2016) | Residential and complexes, “ground land use” Field – Surface Water Runoff | Domestic animal faeces, sewer and septic cross connection leakages and overflow (from sewers, septic sewers, and WW treatment plants) | High |
| | Built-up Areas (Hua, 2017) | Including all residential, commercial, industrial, and transportation | Main causes are residential activities, industrial activities, and sewage | High |

| | | | | | | |
|-----------------------------|---|---|--|---|--|------|
| | | | Field - River | treatment plants. | | |
| Mines (High) | Open-pit aggregate extraction site. Includes associated infrastructure such as roads, buildings, weigh scales, and ponds. | <ul style="list-style-type: none"> Heavy metal presence correlated with higher odds of <i>E. coli</i> detection Preferential flow paths may be present    | Mines/Quarries (Namugize et al., 2018) | NA Note: study conducted in lower-income area of South Africa | NA | High |
| | | | Mines (Somaratne and Hallas, 2015) | Field - River Abandoned copper mine in an unconfined fractured rock aquifer | Preferential flow pathways present | High |
| | | | Mines (Armah, 2014) | Field - Wells Gold mining community Field - Groundwater | Presence of heavy metals correlated with higher odds of detecting <i>E. coli</i> | High |
| Scrubland (Moderate) | Often tree cover <25%, shallow substrates, low shrubs (<2m). Can have rapidly | <ul style="list-style-type: none"> <i>E. coli</i> loadings likely from wildlife Higher connectivity to and transport | Shrubland (Jabbar et al., 2019) | Dominated by shrubs, less than 5 m tall, shrub canopy greater than 20%. Includes grasses, sedges, herbs, young trees, | NA | Low |




| | | | | | | |
|-------------------------------|--|--|--|---|----|-----|
| | draining soils, may contain exposed bedrock. | within groundwater due to “rapidly draining soils”    | or trees stunted from environmental conditions. Field - River | | | |
| Disturbance (Moderate) | Includes peat removed for consumptive (e.g., trees and peat) and non-consumptive (e.g., conservation, erosion protection) uses. Includes forest clear cut <10 years old (low trees, dead trees, mosses, herbaceous). | <ul style="list-style-type: none"> Disturbed soil structure may lead to high connectivity to and transport within groundwater    | Degraded (Namugize et al., 2018) | NA Note: study conducted in lower-income area of South Africa Field - River | NA | Low |
| Bedrock (Moderate) | Exposed bedrock, often | <ul style="list-style-type: none"> Bedrock could be fractured | NA | NA | NA | NA |




| | | | | | | |
|--------------------------|---|--|-------------------------------------|--|---|----------|
| | times <25% vegetation. | leading to high transport and infiltration rates to groundwater, or not fractured leading to high runoff (Arnaud et al., 2015) | | | | |
| | |  | | | | |
| Forest (Moderate) | >60% tree cover (coniferous, deciduous, mixed). | <ul style="list-style-type: none"> • <i>E. coli</i> loading occurring from wildlife • Higher connectivity to and transport through groundwater (preferential flow paths through roots, animal burrows, and insect activity; low runoff | Forest (Jabbar et al., 2019) | Trees greater than 5m tall and >20% vegetation cover | NA | Moderate |
| | | | Native Forest (Larned et al., 2004) | Tall forest dominated by indigenous conifer, broadleaved or beech species Field - River | Mentions positive trends in conductivity. Climate factors such as air temperature and precipitation affect conductivity | Moderate |




| | | | | | |
|--|--|---------------------------------------|--|--|----------|
| | leading to higher infiltration), though natural attenuation likely present • More consistent and preferred soil conditions for <i>E. coli</i> growth (high level of organic matter buffering pH change and nutrient depletion, and plant cover and shading buffering temperature and soil moisture) | | | via rock weathering and atmospheric input. | |
| | | Forested (Petersen and Hubbart, 2020) | NA Literature Review | Lower <i>E. coli</i> potentially due to lower artificial sources (i.e., livestock manure). Lower <i>E. coli</i> attributed to decreased endotherm population density, lack of artificial sources of <i>E. coli</i> , and decreased run-off. | Low |
| | | Forest (Paule-Mercado et al., 2016) | Deciduous, mixed coniferous, and broadleaf forests Field – Surface Water Runoff | Wild animal faeces | Moderate |

| | | | | | |
|--|------|---|---|--|----------|
| | ★★★★ | Forest (Dusek et al., 2018) | Mainly deciduous Field - Soils | Wild animal faeces leading to elevated fecal deposition, combined with more stable soil conditions (high level of organic matter buffering pH change and nutrient depletion, and plant cover and shading buffering temperature and soil moisture). | High |
| | | Forest Plantations (Namugize et al., 2018) | NA Note: study conducted in lower- income area of South Africa Field - River | Positively correlated with <i>E. coli</i> presence. | Moderate |

| | | | | | | |
|---------------------------|--|---|--|---|---|----------|
| | | | Forest (Hua, 2017) | NA Field - River | Generalized with urban area as “vegetation”. | Moderate |
| Wetlands (Low) | Water table seasonally or permanently at, near, or above substrate surface (can be gently flowing). Vegetation can include woody plants, trees (often <25%), shrubs (often hydrophytic, often >25%). | <ul style="list-style-type: none"> Occasionally referred to as “natural purifiers of water”, can remove pathogens, soils can have higher sorption rates; roots of some vegetation can reduce the populations of pathogenic bacteria (Dordio et al., 2008) <i>E. coli</i> loading may occur due to wildlife activity Roots of vegetation can act as | Herbaceous Wetland (Dusek et al., 2018) | Perennial herbaceous vegetation accounts for greater than 80% of vegetative cover and the soil or substrate is periodically saturated with or covered with water. | NA | Moderate |
| | | | Wooded Wetland (Dusek et al., 2018) | Field - Soils Forest or shrubland vegetation accounts for greater than 20% of vegetative cover and soil or substrate is periodically saturated or covered with water | NA | Moderate |
| | | | Wetlands (Namugize et al., 2018) | Field - Soils NA Note: study conducted in lower- | NA | Low |
| | | | | | | |

| | | | | | | |
|------------------------|--|---|----------------------------------|---|---|----------|
| | | <p>preferential flow paths, increasing transport (Rivera, 2014)</p>    | income area of South Africa | | | |
| | | | Field - River | | | |
| Grassland (Low) | Ground layer dominated by prairie graminoids, tree cover < 60%, shrub cover < 25%. | <ul style="list-style-type: none"> Some loading possible with wildlife Offers high efficiency attenuation for waterborne <i>E. coli</i> when runoff rates are low | Wetland (Jabbar et al., 2019) | Periodically saturated or covered with water, forest or shrubland >20% coverage OR herbaceous vegetation >80% | NA | Moderate |
| | | | Field - River | | | |
| | | | Grassland (Dusek et al., 2018) | <p>Dominated by graminoid or herbaceous vegetation, generally >80% of total vegetation. No tilling but may be utilized for grazing.</p> <p>Field - Soils</p> | When soils are not saturated with water, cooler temperatures were associated with increased prevalence. May contain wildlife activity (i.e., deer). | Low |
| | | | Grassland (Elliott et al., 2016) | Defined as “non-pasture” | NA | Low |

|    | Field – River/Model Definition | | | |
|---|--|--|---|----------|
| | Grassland (Paule- Mercado et al., 2016) | NA Field – Surface Water Runoff | Wild animal feces | Low |
| | Grassland (Jabbar et al., 2019) | Dominated by graminoid or herbaceous vegetation, generally >80% of total vegetation. No tilling but may be utilized for grazing. | NA | Moderate |
| | Grassland (Tate et al., 2006) | Field - River Grasslands composed of annual grasses and forbs such as annual ryegrass, wild oats, soft chess, and redstem filaree Bench-scale – Outdoor Study | Seasonally used for cattle grazing, <i>E. coli</i> loading resulting. Grasslands offer a significant capacity to attenuate waterborne <i>E. coli</i> , though | Low |

| | | | | | | |
|--|--|--|-----------------------------|--|----|---|
| | | | | | | this efficiency decreases with increasing runoff rates. |
| Water (Low) | Permanent water > 2m deep, vegetation coverage <25%. | <ul style="list-style-type: none">No notable <i>E. coli</i> loading, only a factor if other LULC (i.e., agriculture/pastoral) produces runoff resulting in a non-point source loading into water body <div></div> | Water (Jabbar et al., 2019) | Open water, typically <25% vegetation cover Field - River | NA | Low |
| <div><ul style="list-style-type: none">- Field – This study was conducted utilizing field acquired data<ul style="list-style-type: none">o Groundwater – <i>E. coli</i> data acquired was from groundwater sourceso River – <i>E. coli</i> data acquired was from river sourceso Soils - <i>E. coli</i> data acquired was from soil sourceso Surface Water Runoff - <i>E. coli</i> data acquired was from surface water runoff sources- Bench-scale – This study used a bench-scale, more-controlled environment to acquire data</div> | | | | | | |

-
- Model Definition – This study is a proof of concept for a pre-existing model
 - Literature Review – This study summarizes pre-existing literature
-

Table A.3-3: Summary of intercepts and coefficients of regression models exploring LULC (independent variable) and *E. coli* contamination potential (dependent variable).

| | All LULC categories (with all <i>E. coli</i> data) | All LULC categories (with <i>E. coli</i> presence subset) | LULC subset with <30,000 Obs.* in category (with all <i>E. coli</i> data) | LULC subset with <30,000 Obs.* in category (with <i>E. coli</i> presence subset) | LULC subset with >135, 000 Obs.* in category (with all <i>E. coli</i> data) | LULC subset with >135, 000 Obs.* in category (with <i>E. coli</i> presence subset) |
|---|---|--|---|--|---|--|
| Intercept | $3.70 \pm 1.82\text{e-}1$ | $3.72 \pm 1.82\text{e-}1$ | $3.92 \pm 1.07\text{e-}2$ | $3.91 \pm 5.97\text{e-}03$ | $3.51 \pm 2.57\text{e-}2$ | $3.52 \pm 1.89\text{e-}2$ |
| Aggregate Mine | $7.42\text{e-}1 \pm 4.81\text{e-}1$ | $5.85\text{e-}1 \pm 2.45\text{e-}1$ | $5.29\text{e-}1 \pm 5.02\text{e-}1$ | $3.92\text{e-}1 \pm 1.63\text{e-}1$ | X | X |
| Bedrock | $-3.07\text{e-}2 \pm 2.26\text{e-}1$ | $-5.13\text{e-}2 \pm 2.45\text{e-}1$ | $-2.43\text{e-}1 \pm 7.99\text{e-}2$ | $-2.44\text{e-}1 \pm 1.21\text{e-}1$ | X | X |
| Disturbance | $4.22\text{e-}3 \pm 2.55\text{e-}1$ | $-9.63\text{e-}03 \pm 2.25\text{e-}1$ | $-2.08\text{e-}1 \pm 1.49\text{e-}1$ | $-2.02\text{e-}1 \pm 2.06\text{e-}1$ | X | X |
| Forest | $-6.23\text{e-}2 \pm 1.86\text{e-}1$ | $-7.69\text{e-}2 \pm 1.71\text{e-}1$ | X | X | $1.27\text{e-}1 \pm 2.82\text{e-}2$ | $1.25\text{e-}1 \pm 3.49\text{e-}2$ |
| Grassland | $-3.70 \pm 1.82\text{e-}1$ | NA | $-3.92 \pm 1.07\text{e-}2$ | NA | X | X |
| Open/Bareland | $1.03 \pm 7.06\text{e-}1$ | $1.28 \pm 4.83\text{e-}1$ | $8.22\text{e-}1 \pm 6.65\text{e-}1$ | $1.08 \pm 3.89\text{e-}1$ | X | X |
| Pastoral/Agricultural | $3.15\text{e-}1 \pm 1.91\text{e-}1$ | $2.97\text{e-}1 \pm 1.81\text{e-}1$ | X | X | $5.05\text{e-}1 \pm 2.48\text{e-}2$ | $4.99\text{e-}1 \pm 2.31\text{e-}2$ |
| Scrubland | $-7.07\text{e-}2 \pm 2.10\text{e-}1$ | $-1.01\text{e-}1 \pm 1.86\text{e-}1$ | $-2.83\text{e-}1 \pm 4.52\text{e-}2$ | $-2.94\text{e-}1 \pm 4.86\text{e-}2$ | X | X |
| Urban | $2.19\text{e-}1 \pm 1.80\text{e-}1$ | $1.94\text{e-}1 \pm 1.83\text{e-}1$ | X | X | $4.08\text{e-}1 \pm 2.63\text{e-}2$ | $3.96\text{e-}1 \pm 2.04\text{e-}2$ |
| Water | $-4.70\text{e-}1 \pm 1.84\text{e-}1$ | $-4.79\text{e-}1 \pm 1.92\text{e-}1$ | $-6.83\text{e-}1 \pm 5.48\text{e-}2$ | $-6.72\text{e-}1 \pm 5.65\text{e-}2$ | X | X |
| <ul style="list-style-type: none"> • NA represents a LULC category that was to be included in the regression analysis but did not contain any water sample observations • X represents a LULC category that was not considered in an analysis <p>*Due to large discrepancy in number of water sample observations in LULC categories, analyses were also undertaken excluding LULC categories with low and high water sample observations, respectively</p> | | | | | | |

A.4 CHAPTER 4 APPENDICES

A.4.1 Figures and Tables

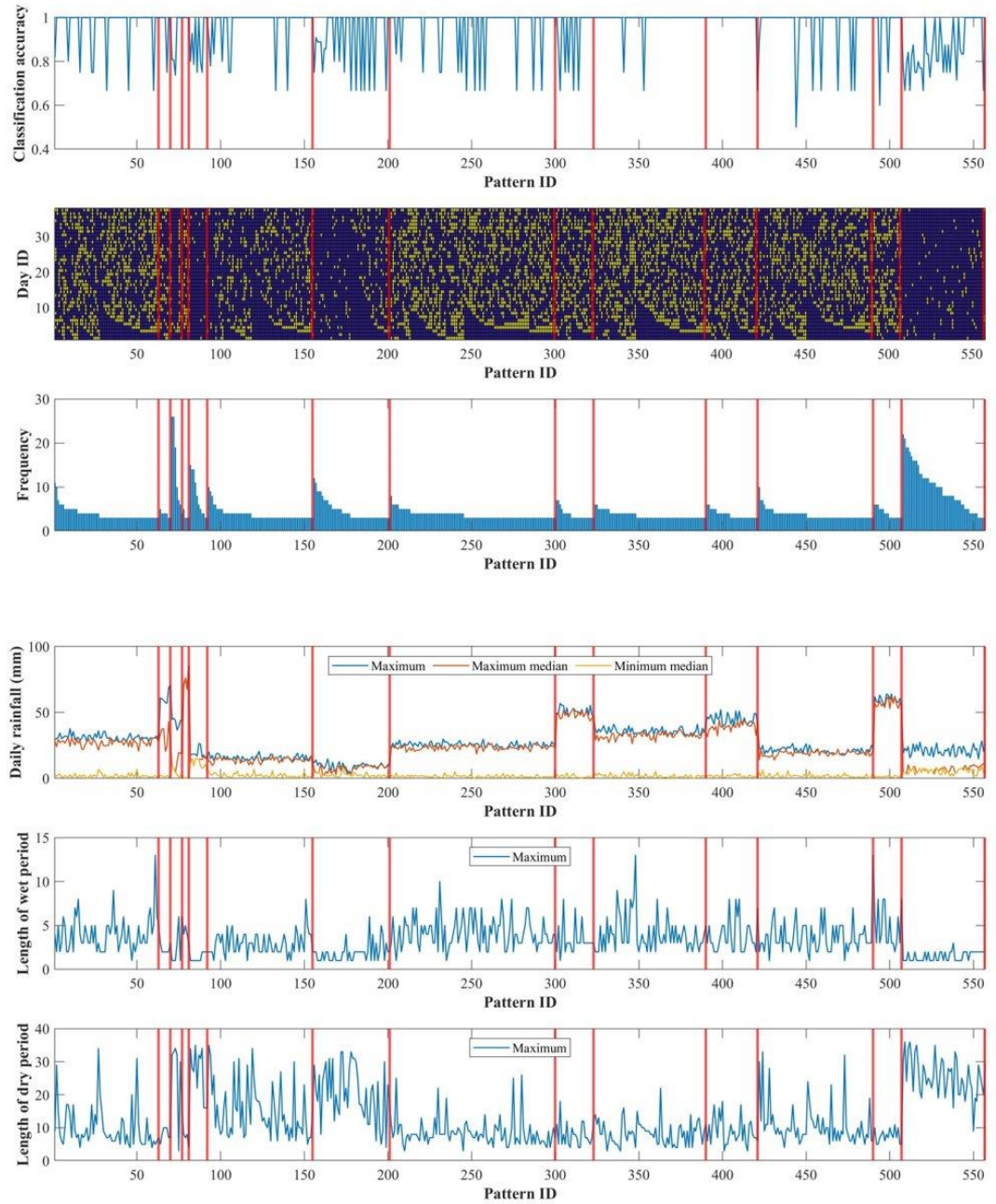


Figure A.4-1: Summary of original 14 clusters based on spectral analysis.

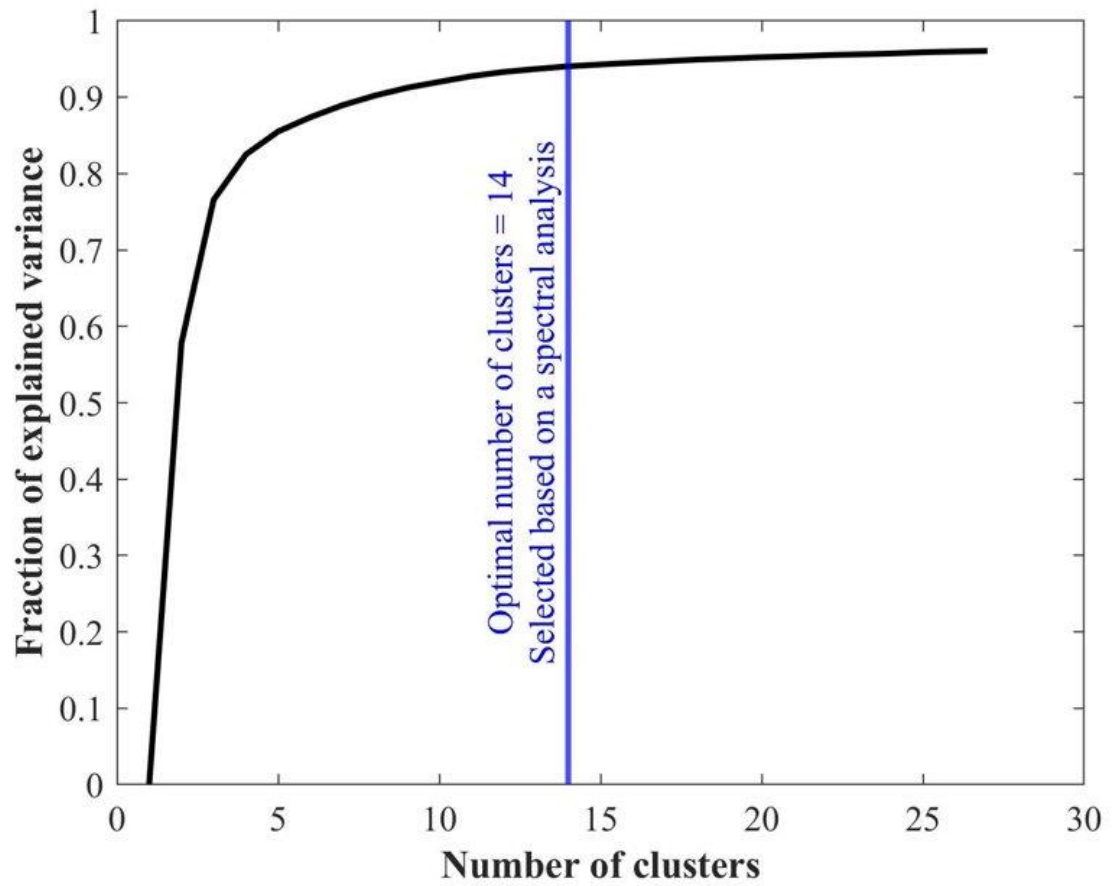


Figure A.4-2: Fraction of explained variance of the model based on the number of clusters allowed.

Table A.4-1: Summary of clusters based on adverse and non-adverse *E. coli* observations from n to n-5 days.

| | Adverse <i>E. coli</i> Observations | | | | | | Non-Adverse <i>E. coli</i> Observations | | | | | | |
|---|-------------------------------------|-----------|-----------|-----------|-----------|-----------|---|-----------|-----------|-----------|-----------|-----------|-----------|
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | All Dry Days | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
| Number of unique patterns | 9 | 2 | 14 | 17 | 12 | 9 | 1 | 14 | 12 | 20 | 5 | 6 | 6 |
| Number of observations | 5388 | 480 | 3630 | 7239 | 3351 | 1310 | 158442 | 89016 | 46540 | 280504 | 26145 | 41732 | 28801 |
| Max number of wet/dry days | 2/5 | 5/1 | 6/3 | 3/5 | 4/3 | 4/4 | | 5/3 | 4/3 | 3/5 | 5/3 | 6/3 | 3/4 |
| Mean number of wet days | 1.5±0.5 | 5±0 | 4.5±1.5 | 2±1 | 3.5±0.5 | 3±1 | 0/6 | 4±1 | 3.5±0.5 | 2±1 | 4±1 | 4.5±1.5 | 2.5±0.5 |
| Mean number of dry days | 4.5±0.5 | 1±0 | 1.5±1.5 | 4±1 | 2.5±0.5 | 3±1 | | 2±1 | 2.5±0.5 | 4±1 | 2±1 | 1.5±1.5 | 3.5±0.5 |
| Max rainfall range (i.e., the maximum of each pattern's range within each cluster) | 76.63 | 67.61 | 62.22 | 65.48 | 85.28 | 44.01 | | 76.99 | 65.71 | 85.21 | 112.76 | 87.70 | 63.87 |
| | | | | | | | -- | | | | | | |
| Minimum | 0.04 | 2.05 | 1.06 | 0.01 | 1.14 | 0.66 | | 0.35 | 0.51 | 0.01 | 0.51 | 0.35 | 0.23 |
| Mean | 11.9±9.8 | 19.0±10.4 | 16.3±9.7 | 11.2±9.4 | 16.3±10.6 | 13.1±8.2 | | 15.1±9.5 | 13.9±8.9 | 9.3±8.3 | 16.8±11.0 | 14.8±9.9 | 10.5±7.9 |
| Max/Min number of consecutive wet days | 2/1 | 5/3 | 6/2 | 3/1 | 4/2 | 3/1 | | 5/2 | 4/2 | 2/1 | 4/2 | 6/2 | 1/1 |
| | | | | | | | 0 | | | | | | |
| Mean | 1.5±0.5 | 4±1 | 4±2 | 2±1 | 3±1 | 2±1 | | 3.5±1.5 | 3±1 | 1.5±0.5 | 3±1 | 4±2 | 1±0 |

| | | | | | | | | | | | | | |
|---|---------|-------|---------|-------|-------|-------|-------|-------|-------|---------|---------|---------|-------|
| Max/Min number of consecutive dry days | 5/2 | 1/1 | 3/0 | 5/1 | 3/1 | 3/1 | 6 | 3/1 | 3/1 | 5/2 | 2/1 | 3/0 | 3/1 |
| Mean | 3.5±1.5 | 1±0 | 1.5±1.5 | 3±2 | 2±1 | 2±1 | | 2±1 | 2±1 | 3.5±1.5 | 1.5±0.5 | 1.5±1.5 | 2±1 |
| LULC ratio: | | | | | | | | | | | | | |
| Very High | 0.416 | 0.385 | 0.429 | 0.416 | 0.417 | 0.413 | 0.401 | 0.386 | 0.380 | 0.385 | 0.392 | 0.389 | 0.380 |
| High | 0.285 | 0.277 | 0.275 | 0.280 | 0.280 | 0.292 | 0.313 | 0.310 | 0.311 | 0.315 | 0.312 | 0.314 | 0.315 |
| Moderate | 0.226 | 0.248 | 0.230 | 0.233 | 0.223 | 0.217 | 0.224 | 0.236 | 0.238 | 0.231 | 0.229 | 0.228 | 0.233 |
| Low | 0.073 | 0.090 | 0.066 | 0.071 | 0.079 | 0.078 | 0.062 | 0.068 | 0.070 | 0.068 | 0.066 | 0.069 | 0.072 |
| Stratigraphy ratio: | | | | | | | | | | | | | |
| Bedrock | 0.715 | 0.765 | 0.723 | 0.711 | 0.703 | 0.718 | 0.628 | 0.645 | 0.648 | 0.642 | 0.633 | 0.654 | 0.647 |
| Overburden | 0.285 | 0.235 | 0.277 | 0.289 | 0.297 | 0.282 | 0.372 | 0.355 | 0.352 | 0.358 | 0.367 | 0.346 | 0.353 |
| <i>E. coli</i> ratio: | | | | | | | | | | | | | |
| High | 0.078 | 0.100 | 0.100 | 0.092 | 0.096 | 0.095 | -- | -- | -- | -- | -- | -- | -- |
| Moderate | 0.154 | 0.156 | 0.151 | 0.148 | 0.138 | 0.153 | | | | | | | |
| Low | 0.768 | 0.744 | 0.749 | 0.760 | 0.766 | 0.752 | | | | | | | |

Table A.4-2: Summary of clusters based on adverse and non-adverse *E. coli* observations from n to n-16 days.

| | Adverse <i>E. coli</i> Observations | | | | | | Non-Adverse <i>E. coli</i> Observations | | | | | | |
|--|-------------------------------------|-----------|-----------|-----------|-----------|-----------|---|-----------|-----------|-----------|-----------|-----------|-----------|
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | All Dry Days | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
| Number of unique patterns | 335 | 98 | 347 | 168 | 30 | 410 | 1 | 701 | 364 | 1278 | 179 | 1136 | 806 |
| Number of observations | 2695 | 876 | 1937 | 1229 | 197 | 2731 | 55718 | 34759 | 52491 | 111038 | 15091 | 71054 | 86467 |
| Max number of wet/dry days | 13/16 | 14/16 | 14/16 | 13/16 | 9/15 | 13/16 | | 14/15 | 15/16 | 15/16 | 15/15 | 14/15 | 14/16 |
| Mean number of wet days | 7±6 | 7.5±6.5 | 7.5±6.5 | 7±6 | 5.5±3.5 | 7±6 | 0/17 | 8±6 | 8±7 | 8±7 | 8.5±6.5 | 8±6 | 7.5±6.5 |
| Mean number of dry days | 10±6 | 9.5±6.5 | 9.5±6.5 | 10±6 | 11.5±3.5 | 10±6 | | 9±6 | 9±7 | 9±7 | 8.5±6.5 | 9±6 | 9.5±6.5 |
| Max rainfall range (i.e., the maximum of each pattern's range within each cluster) | 37.69 | 61.93 | 20.57 | 48.67 | 85.28 | 29.04 | -- | 22.69 | 62.94 | 38.99 | 104.11 | 30.74 | 49.54 |
| Minimum | 0.1 | 0.07 | 0.4 | 0.73 | 2.4 | 0.08 | | 0.5 | 0.01 | 0.01 | 0.64 | 0.13 | 0.02 |
| Mean | 18.8±10.1 | 26.8±18.5 | 11.6±4.7 | 24.3±13.2 | 37.0±22.7 | 16.5±7.5 | | 11.5±4.5 | 16.7±13.5 | 15.6±9.2 | 25.8±17.3 | 14.8±6.4 | 15.7±11.0 |
| Max/Min number of consecutive wet days | 12/1 | 8/1 | 11/1 | 11/1 | 8/1 | 10/1 | 0 | 13/1 | 11/1 | 13/1 | 12/1 | 11/1 | 11/1 |
| Mean | 6.5±5.5 | 4.5±3.5 | 6±5 | 6±5 | 4.5±3.5 | 5.5±4.5 | | 7±6 | 6±5 | 7±6 | 6.5±5.5 | 6±5 | 6±5 |
| Max/Min number of | 15/2 | 13/1 | 13/2 | 14/1 | 14/3 | 16/1 | 17 | 13/1 | 15/1 | 16/1 | 12/1 | 13/1 | 16/2 |

| | | | | | | | | | | | | | |
|---------------------------------|---------|-------|---------|---------|---------|---------|-------|-------|-------|---------|---------|-------|-------|
| consecutive dry days | | | | | | | | | | | | | |
| Mean | 8.5±6.5 | 7±6 | 7.5±5.5 | 7.5±6.5 | 8.5±5.5 | 8.5±7.5 | | 7±6 | 8±7 | 8.5±7.5 | 6.5±5.5 | 7±6 | 9±7 |
| LULC ratio: | | | | | | | | | | | | | |
| Very High | 0.421 | 0.430 | 0.423 | 0.437 | 0.345 | 0.434 | 0.413 | 0.384 | 0.399 | 0.383 | 0.385 | 0.374 | 0.396 |
| High | 0.291 | 0.280 | 0.279 | 0.302 | 0.269 | 0.283 | 0.302 | 0.329 | 0.321 | 0.323 | 0.311 | 0.324 | 0.323 |
| Moderate | 0.218 | 0.220 | 0.226 | 0.190 | 0.269 | 0.208 | 0.228 | 0.216 | 0.216 | 0.224 | 0.230 | 0.229 | 0.215 |
| Low | 0.070 | 0.070 | 0.073 | 0.072 | 0.117 | 0.075 | 0.058 | 0.071 | 0.064 | 0.070 | 0.074 | 0.073 | 0.066 |
| Stratigraphy ratio: | | | | | | | | | | | | | |
| Bedrock | 0.713 | 0.693 | 0.707 | 0.710 | 0.741 | 0.715 | 0.626 | 0.639 | 0.621 | 0.630 | 0.651 | 0.638 | 0.630 |
| Overburden | 0.287 | 0.307 | 0.293 | 0.290 | 0.259 | 0.285 | 0.374 | 0.361 | 0.379 | 0.370 | 0.349 | 0.362 | 0.370 |
| E. coli ratio: | | | | | | | | | | | | | |
| High | 0.093 | 0.103 | 0.090 | 0.103 | 0.091 | 0.090 | -- | -- | -- | -- | -- | -- | -- |
| Moderate | 0.146 | 0.150 | 0.155 | 0.148 | 0.183 | 0.149 | | | | | | | |
| Low | 0.761 | 0.747 | 0.755 | 0.749 | 0.726 | 0.761 | | | | | | | |

A.5 CHAPTER 5 APPENDICES

A.5.1 Figures and Tables

Closest Driving Time vs. First Message Received

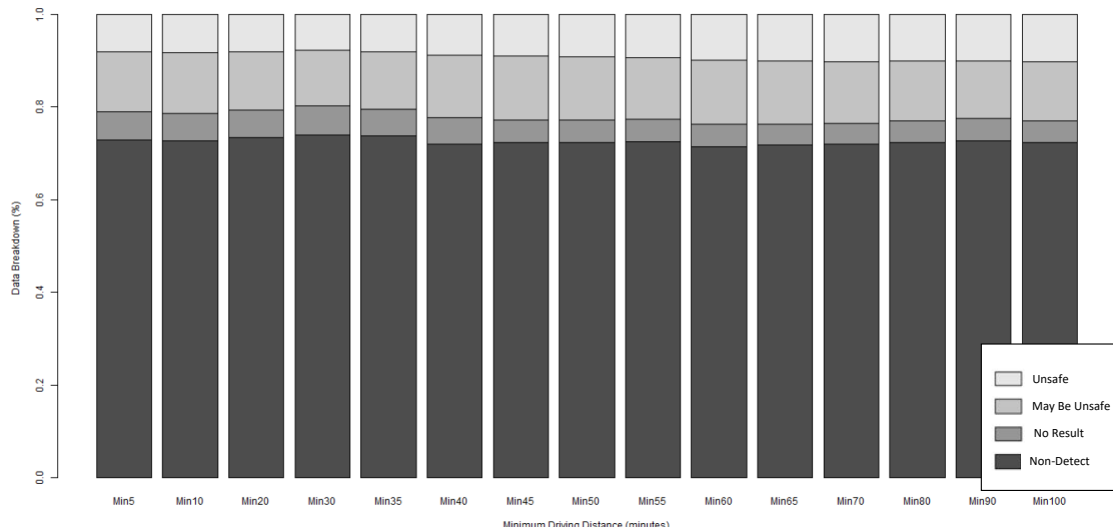


Figure A.5-1: Closest driving time bins of closest drop-off location versus first test message received.

Closest Driving Time vs. Year of Test

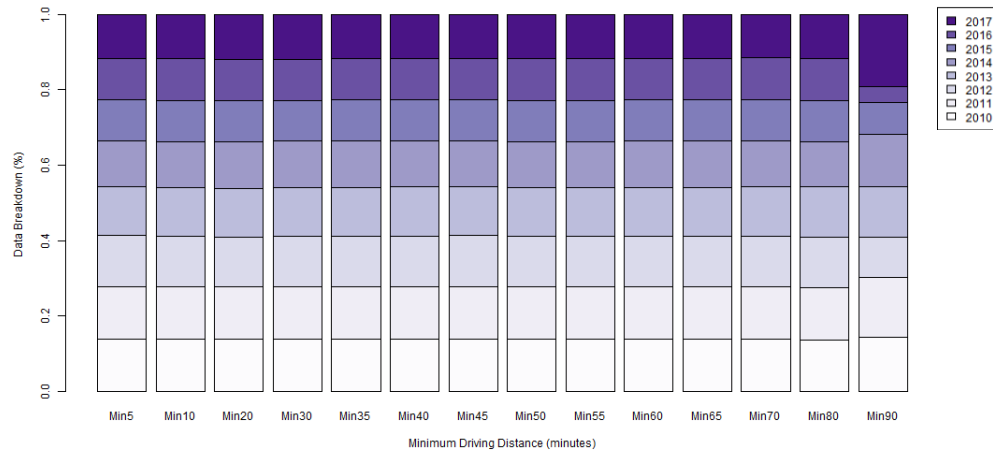


Figure A.5-2: Closest driving time bins of closest drop-off location versus year of test.

Closest Driving Time vs. Month of Test

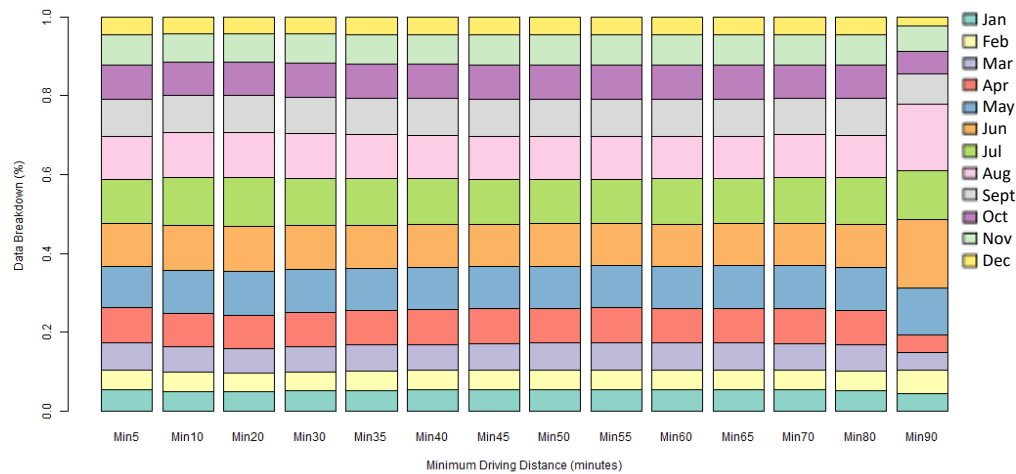


Figure A.5-3: Closest driving time bins of closest drop-off location versus month of test.

Closest Driving Time vs. Week Day of Test

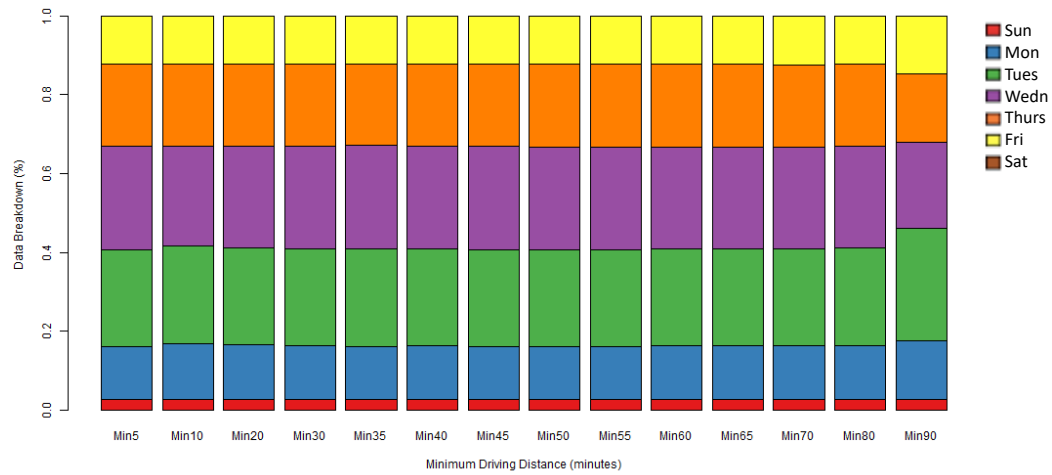


Figure A.5-4: Closest driving time bins of closest drop-off location versus weekday of test.

Table A.5-1: Summary of SWE inclusion criteria based on Ontario Plant Hardiness Zones.

| Hardiness Zone | Summer | Autumn | Winter | Spring |
|-----------------------|--------------------|--------------------|-------------------|-------------------|
| 0b | (Jun 15 – Jul 30) | [Jul 30 – Dec 10) | [Dec 10 – Jun 13] | (Jun 13 – Jun 15] |
| 1a | (Jun 19 – Aug 5) | [Aug 5 – Dec 10) | [Dec 10 – Jun 6] | (Jun 6 – Jun 19] |
| 1b | (Jun 20 – Aug 21) | [Aug 21 – Dec 12) | [Dec 12 – Jun 18] | (Jun 18 – Jun 20] |
| 2a | (Jun 7 – Sept 9) | [Sept 9 – Dec 13) | [Dec 13 – Jun 4] | (Jun 4 – Jun 7] |
| 2b | (Jun 7 – Sept 9) | [Sept 9 – Dec 13) | [Dec 13 – Jun 3] | (Jun 3 – Jun 7] |
| 3a | (Jun 3 – Sept 16) | [Sept 16 – Dec 14) | [Dec 14 – Jun 2] | (Jun 2 – Jun 3] |
| 3b | (May 27 – Sept 17) | [Sept 17 – Dec 15) | [Dec 15 – May 24] | (May 24 – May 27] |
| 4a | (May 24 – Sept 22) | [Sept 22 – Dec 18) | [Dec 18 – May 21] | (May 21 – May 24] |
| 4b | (May 24 – Sept 22) | [Sept 22 – Dec 21) | [Dec 21 – May 15] | (May 15 – May 24] |
| 5a | (May 17 – Sept 26) | [Sept 26 – Dec 27) | [Dec 27 – May 5] | (May 5 – May 17] |
| 5b | (May 11 – Oct 1) | [Oct 1 – Dec 29) | [Dec 29 – Apr 27] | (Apr 27 – May 11] |
| 6a | (May 3 – Oct 8) | [Oct 8 – Dec 31) | [Dec 31 – Apr 22] | (Apr 22 – May 3] |
| 6b | (Apr 30 – Oct 13) | [Oct 13 – Jan 3) | [Jan 3 – Apr 7] | (Apr 7 – Apr 30] |
| 7a | (Apr 25 – Oct 20) | [Oct 20 – Jan 6) | [Jan 6 – Apr 1] | (Apr 1 – Apr 25] |

Table A.5-2: Summary of impact to *E. coli* in “summer” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable.

| Variables | | LASSO – Cat | LASSO – Disc/Cont | MARS – Cat | MARS – Disc/Cont | GAMLSS - Cat | GAMLSS – Disc/Cont |
|-----------|----------|--|--|-------------------------------|--|----------------------|----------------------------|
| DriveTime | | Decreases | Decreases | | | | Decreases concentration |
| SpecCap | | Decrease presence | Decrease presence | | | | |
| LULC | | Decreases | Decrease, magnitude 2 greater for concentration | Hinge @ 2, decreases after | Hinge @ 2, decrease after | Decreases | Decreases |
| WellDept | | Decreases | Decreases | | | Decrease presence | Decreases |
| GiBin | Hotspot | Increase presence, one magnitude more than coldspot Increase concentration | Increases | | | | |
| | Coldspot | Increase presence | | | | | |
| TestFreq | | Increase presence Decrease concentration | Increase presence Decrease concentration | | Hinge @ Freq~10 & Den~15.4, increase concentration | | |

| | | | | | | | |
|--|----------------------------|---|---|----------------------------------|---|-------------------------|-----------|
| Rainfall Wet/Dry Cycles | | Increase presence Decrease concentration | Increase presence Decrease concentration | | | | |
| Rainfall Maximum Consecutive Wet Days | | Decrease concentration | Decrease concentration | | | | |
| Well Density | Category 1 | Increases presence | Decrease | | Hinge @ Cat3/16.1, decreases to hinge then plateaus Hinge @ Freq~10 & Den~16.6, increase concentration | Second highest increase | Decreases |
| | Category 2 | | | | | Third highest increase | |
| | Category 3 | Increases concentration | | | | Highest increase | |
| | Category 4 | Decreases | | If True, concentration decreases | | | |
| Bottom Casing Material | Cracking | Increases | Decreases | | | | |
| | OpenHole | Increases presence, two mag less than Crack | | | | Increase concentration | |
| | Potential Corrosion | Decrease presence | | | | | |
| | Stainless | | | | | Decrease presence | |

| | | | | | | | |
|-----------------------------|------------|---|---|--|--|-------------------|-----------|
| Bottom of Well Stratigraphy | Bedrock | Increase presence Decrease concentration | Increase presence Decrease concentration | | | Decrease presence | Decreases |
| | Overburden | | | | | | |

Table A.5-3: Summary of impact to *E. coli* in “shoulder” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Cont” represents the use of continuous variable.

| Variables | | LASSO – Cat | LASSO – Cont | MARS – Cat | MARS – Cont | GAMLSS - Cat | GAMLSS – Cont |
|-----------|----------|--|-------------------|---|---|-----------------|-------------------------|
| DriveTime | | Increases presence | Increase presence | | | | |
| SpecCap | | Decreases | Decreases | | | | Decreases concentration |
| LULC | | Decreases | Decreases | Hinge @LULC=2, 1&2 the same, decreases after Hinge @ LULC=2 & Overburden=T, concentration much higher for LULC=1 | Hinge @LULC=2, slightly increases from 1 to 2, then decreases | Decreases | Decreases |
| WellDept | | Decreases | Decreases | | | Decreases | Decreases |
| GiBin | Hotspot | Decreases (1 magnitude more for concentration) | | | | | |
| | Coldspot | Increase presence | | | | | |

| | | | | | | | |
|---|----------------------------|---|---|---|--|-----------|-----------|
| TestFreq | | Increase presence | Increase presence | | | | |
| Total Water (Rainfall + Snow Melt) | | Increases | Increases | | | Increases | Increases |
| Rainfall Wet/Dry Cycles | | Decrease presence | Decreases | | | | |
| Maximum Consec Melt Day Sum | | Decreases | Decreases | | | Decreases | Decreases |
| Bottom Casing Material | Cracking | Decreases | Decrease presence Increase concentration | | | | |
| | OpenHole | Increases presence | | | | | |
| | Potential Corrosion | Decrease presence | | | | | |
| | Stainless | Decrease presence | | | | | |
| Bottom of Well Stratigraphy | Bedrock | Increase presence Decrease concentration | Increase presence | | | Decreases | Decreases |
| | Overburden | Decrease presence | Decrease concentration | Hinge @ LULC=2 & Overburden=T, concentration much higher for LULC=1 | | | |

Table A.5-4: Summary of impact to *E. coli* in “winter” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable.

| Variables | | LASSO – Cat | LASSO – Disc/Cont | MARS – Cat | MARS – Disc/Cont | GAMLSS - Cat | GAMLSS – Disc/Cont |
|-----------|----------|-----------------------|----------------------|--|--|----------------------|--------------------------|
| DriveTime | | Decreases | Decreases | | | Decrease presence | Decreases |
| SpecCap | | Decreases presence | Decrease presence | | | | |
| LULC | | Decreases | Decreases | Hinge @ LULC = 2 & Melt Days = 1; decreases as LULC >=2 and Melt Days>1 | Hinge @ LULC = 2 & Melt Days = 1; decreases as LULC >=2 and Melt Days>1 | | |
| WellDept | | Decrease presence | Decreases | | | | |
| GiBin | Hotspot | Decrease presence | Decrease presence | | | | |
| | Coldspot | Increase presence | | | | | |

| | | | | | | | |
|------------------------------|-------------------|---|---|---|---|-------------------|-----------|
| TestFreq | | | Decrease concentration | | | Decrease presence | Decreases |
| Number of Melt Days | | Increase presence Decrease concentration | Increase presence Decrease concentration | Hinge @ LULC = 2 & Melt Days = 1; decreases as LULC >=2 and Melt Days>1 | Hinge @ LULC = 2 & Melt Days = 1; decreases as LULC >=2 and Melt Days>1 | | |
| Maximum Melt Day (mm) | | Increase presence | Increase presence | | | | |
| Well Density | Category 1 | Increases | Decreases | | | | Decreases |
| | Category 2 | Increase presence (less than Cat1) | | | | | |
| | Category 3 | | | | | | |
| | Category 4 | Decreases | | | | | |
| Flow Accumulation | Category 1 | | Decreases | | | | |
| | Category 2 | Increase presence | | | | | |

| | | | | | | | |
|------------------------------------|-------------------|---|---|--|--|--|--|
| | Category 3 | Decreases | | | | | |
| | Category 4 | Decrease presence | | | | | |
| Bottom of Well Stratigraphy | Bedrock | Increase presence Decrease concentration | Increase presence Decrease concentration | | | | |
| | Overburden | | | | | | |

Table A.5-5: Summary of impact to *E. coli* in “all seasons” models if one variable was increased at a time, only model deemed significant variables record an impact. “Cat” represents the use of categorical variables, “Disc/Cont” represents the use of discrete and/or continuous variable.

| Variables | | LASSO – Cat | LASSO – Disc/Cont | MARS – Cat | MARS – Disc/Cont | GAMLSS - Cat | GAMLSS – Disc/Cont |
|-----------|----------|--|-----------------------|----------------------------------|-------------------------------|-----------------------|--------------------------|
| DriveTime | | Decreases | Decreases | | | Decreases presence | |
| SpecCap | | Decreases | Decreases | | | | |
| LULC | | Decreases | Decreases | Hinge @ 2, decreases after | Hinge @ 2, decreases after | Decreases | Decreases |
| WellDept | | Decreases | Decreases | | | Decreases | Decreases |
| GiBin | Hotspot | Increases presence | Increases | | | | |
| | Coldspot | Increases presence (slightly more) | | | | | |
| TestFreq | | Increases presence | Increases presence | | | Decreases | Decreases |

| | | | | | | | |
|-------------------------------|-------------------|---|---|------------------------------|---|------------------------|-----------|
| | | Decreases concentration | Decreases concentration | | | | |
| Total Water | | Increase presence | Increase presence Decrease concentration | Hinge @ 282, increases after | Hinge @ 297, increases after | | |
| Well Density | Category 1 | Increases | Decreases | | Hinge @ 41/Cat 3, decreases to 41 then plateaus | Increases | Decreases |
| | Category 2 | Increase presence Decrease concentration | | | | Increases | |
| | Category 3 | Increases | | | | Increases | |
| | Category 4 | Decreases | | Decrease if True | | | |
| Bottom Casing Material | Cracking | Increase presence (1 mag more than rest) | Decrease presence | | | | |
| | OpenHole | Increase presence | | | | Increase concentration | |

| | | | | | | | |
|------------------------------------|----------------------------|---|---|--|--|-------------------|-----------|
| | Potential Corrosion | Decreases presence (-0.015) | | | | | |
| | Stainless | Decreases presence (-0.0185, 2 iter) | | | | | |
| Bottom of Well Stratigraphy | Bedrock | Increase presence Decrease concentration | Increase presence Decrease concentration | | | Decrease presence | Decreases |
| | Overburden | Decrease presence Increase concentration | | | | | |