Water Research 197 (2021) 117089

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

# Exploration of *E. coli* contamination drivers in private drinking water wells: An application of machine learning to a large, multivariable, geo-spatio-temporal dataset

Katie White[a], Sarah Dickson-Anderson[a,f,*], Anna Majury[b,e], Kevin McDermott[b], Paul Hynds[c], R. Stephen Brown[d], Corinne Schuster-Wallace[a,f]

[a] Department of Civil Engineering, McMaster University, 1280 Main St. W, Hamilton, Ontario, L8S 4L8, Canada
[b] Public Health Ontario, 181 Barrie St, Kingston, Ontario, K7L 3K2, Canada
[c] Environmental Sustainability and Health Institute, Technological University Dublin, Grangegorman Dublin 7, Republic of Ireland
[d] Department of Chemistry and School of Environmental Studies, Queen's University, 99 University Ave, Kingston, Ontario, K7L 3N6, Canada
[e] Department of Biology and Molecular Sciences, Department of Public Health Sciences, School of Environmental Studies, Queen's University, 99 University Ave, Kingston, Ontario, K7L 3N6, Canada
[f] Department of Geography and Planning and Global Institute for Water Security, University of Saskatchewan, 117 Science Place, Saskatoon, Saskatchewan, S7N 5C8, Canada

ARTICLE INFO

*Article history:*
Received 9 November 2020
Revised 22 February 2021
Accepted 23 March 2021
Available online 27 March 2021

*Keywords:*
Private drinking water
Groundwater
*E. coli*
Testing trends
Large dataset
Machine learning

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.



## 1. 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

* Corresponding author.
*E-mail address:* sdickso@mcmaster.ca (S. Dickson-Anderson).

https://doi.org/10.1016/j.watres.2021.117089
0043-1354/

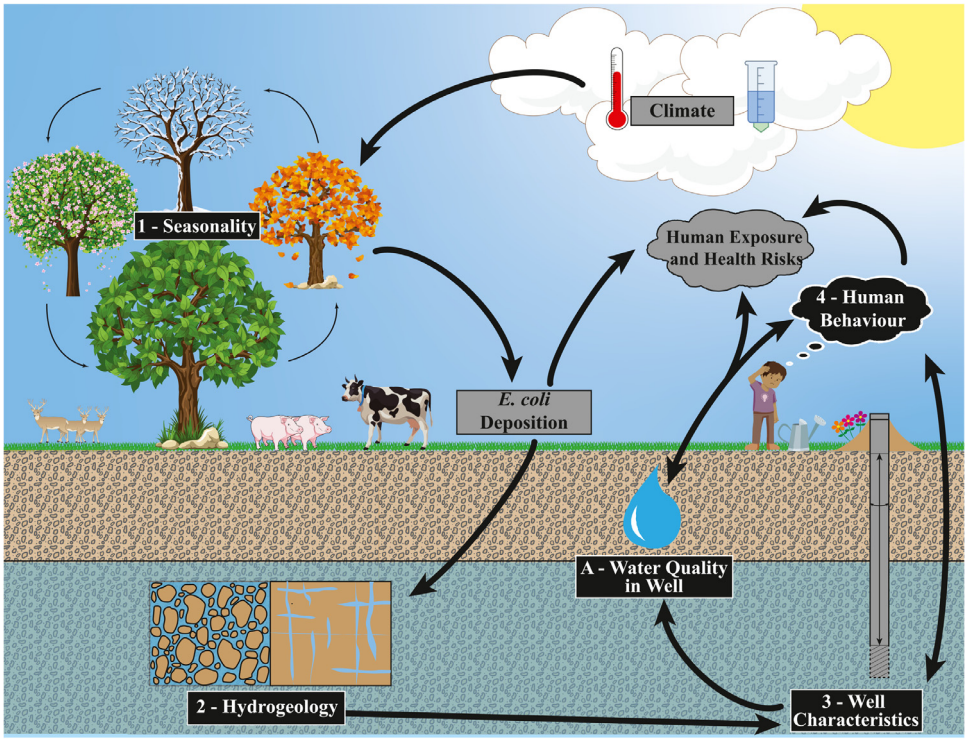


**Fig. 1.** Fate and transport mechanisms driving *E. coli* concentrations in private wells (i.e., contamination risk) considering a coupled-systems approach, adapted from (Di Pelino et al., 2019), where numbered text represent drivers used to develop explanatory models.

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 overexploitation 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 (Fig. 1) (Di Pelino et al., 2019).

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 (Castalletti 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 char-

acteristics (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. Methods

### 2.1. 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 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 S1.1. These sub-classifications are used in lieu of, or alongside, discrete values in some analyses to fit regulatory definitions or account for uncertainty.

### 2.2. 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 1) were required to enable the assessment of factors driving the presence of *E. coli* in private wells in Ontario, as described in S1.1. 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.

### 2.3. Statistical analyses

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 S1.1). 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

**Table 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 |
| *E. coli* Result | Number of *E. coli* reported in sample by laboratory. Laboratory Reporting Range: 0 – 80 CFU/100 mL | non-detects (ND): 0<br>Category 1: 1-10<br>Category 2: 11-50<br>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: $\leq 5$<br>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 S2.1)<br>Unconsolidated (further categorised as high, medium, low permeability) (See Table S2.2) |
| Pump Test | 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) | Specific Capacity (GPM/m) = Pumping Rate/Drawdown<br>Low (0 - <3.3 GPM/m)<br>Moderate (3.3 – 16.4 GPM/m)<br>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)<br>Moderate 1 ($12.5$ m $\leq x < 18.3$ m)<br>Moderate 2 ($18.3$ m $\leq x < 24.4$ m)<br>Moderate 3 ($24.4$ m $\leq x < 31.1$ m)<br>Moderate 4 ($31.1$ m $\leq x < 41.8$ m)<br>Moderate 5 ($41.8$ m $\leq x < 61$ m)<br>Deep ($\geq 61$ m) |
| Date of Well Construction | Year well construction was completed | |
| Status | Qualitative microbiology comments based on laboratory processing of the sample (See Table S2.3) | |

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 S1.1). 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).

#### 2.3.1. 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 S2.4). 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 S2.4) were selected to develop a series of models to explain *E. coli* concentrations, each exploring different elements of the risk pathway (Fig. 1): seasonal (Driver 1 in Fig. 1), hydrogeological (Driver 2 in Fig. 1), well characteristics (Driver 3 in Fig. 1), and testing practices (Driver 4 in Fig. 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; Rigby and Stasinopoulos, 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.

#### 2.3.2. 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. 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 S1.2). Decay curves were then created utilizing the nonlinear least squares (nls) method (see S1.2) to estimate the parameters ($y_o$, $y_f$, $\alpha$) 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 S1.2).

## 3. Results and discussion

Models for each driver are described and discussed in the following sections. Each model is described in (Table S2.5) and summarised in Fig. 2.

### 3.1. Seasonal drivers (driver 1 in Fig. 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 (Fig. 2; Table S2.5; Figures S2.1-S2.6). 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 $0.16 \pm 0.01$ CFU/100mL (p-value $< 0.01$) (Figure S2.1; Table S2.5). 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 8$^{th}$ and May 3$^{rd}$, respectively, compared to September 16$^{th}$ and June 3$^{rd}$ 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 S2.5). 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 ex-

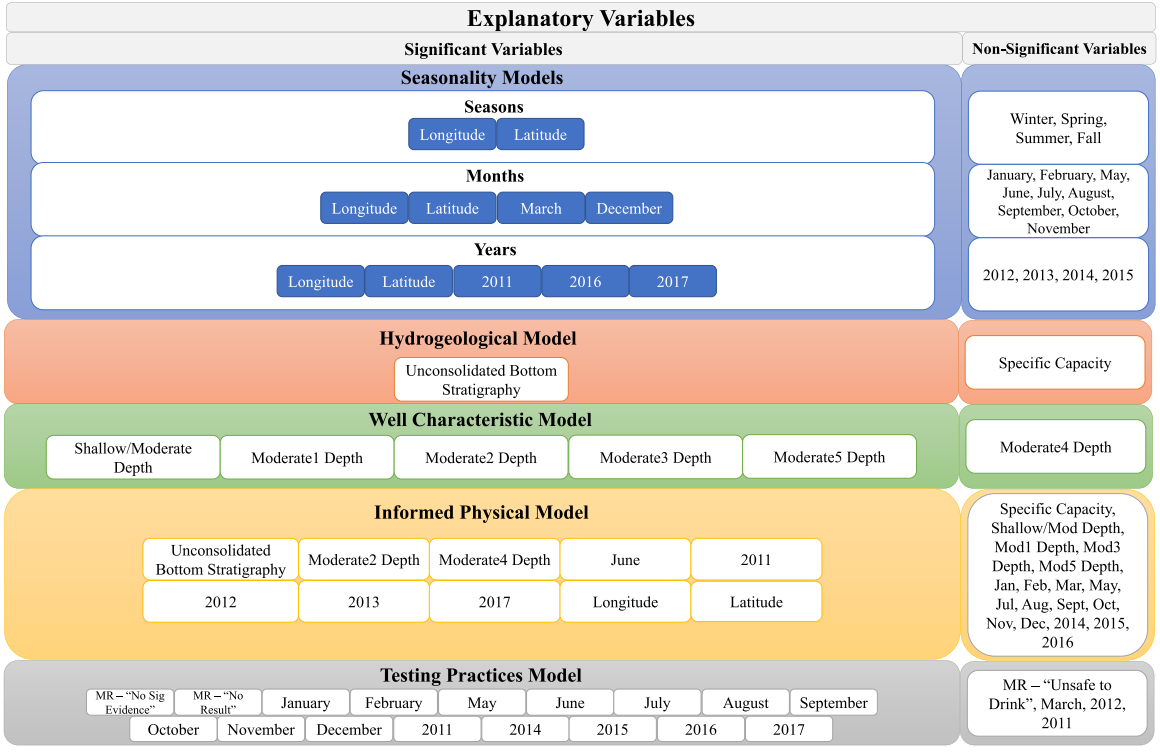


**Fig. 2.** Summary of explanatory variables across "best" models for each driver (Fig. 1).

plain 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) (Fig. 4), although these results do not represent the largest increase in *E. coli* concentrations (Fig. 2; Table S2.5). 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). Additionally, 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-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) (Fig. 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-value = 0.03). December emerges as a month with some of the lowest numbers of adverse *E. coli* sample results (3.1%; n = 799) (Fig. 4) and largest explanatory *E coli* concentration decrease (-0.27 ± 0.05 CFU/100mL, p-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-value < 0.01), with a decrease in average *E. coli* concentration of 0.16 ± 0.01 CFU/100mL per half-degree of latitude (Figure S2.3-S2.4; Table S2.5). 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 (Fig. 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-value < 0.01), except for 2011 to 2012 (Figure S2.5-S2.6). 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-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 S2.5).

### 3.2. Hydrogeological drivers (driver 2 in Fig. 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-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 S2.6).

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 S2.7). 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.

### 3.3. Well characteristics (driver 3 in Fig. 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 im-

proved 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 S2.7-S2.8). 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 $\leq 0.0027$) (Table S2.8). 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.

### 3.4. Informed physical model

The findings from the seasonal, hydrogeological, and well characteristic models were used to create an informed physical model (Fig. 2, Figure S2.9-S2.10). Based on model outputs (Table S2.5) 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.

### 3.5. Testing practices (driver 4 in Fig. 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 S2.5). 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 ($\leq 16$ 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 concerning 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 (Fig. 2; Figure S2.11-S2.12). 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 S2.5). 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 S2.5).

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-value = $< 0.0001$), or "may be unsafe" (13%, n=31,656) (decay rate = 0.40, p-value = $< 0.0001$), or "unsafe to drink" (8%, n=20,341) (decay rate = 0.40, p-value = $< 0.0001$) (Fig. 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-value = $< 0.0001$; "no result", decay rate = 0.16080, p-value = $< 0.0001$; "may be unsafe", decay rate = 0.15107, p-value = $< 0.0001$; "unsafe to drink", decay rate = 0.12197, p-value = $< 0.0001$) (Table S2.9). Routine testing indicates greater awareness of appropriate well stewardship practices (Lavallee et al., 2020).

The majority of samples were submitted for testing in July (Fig. 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

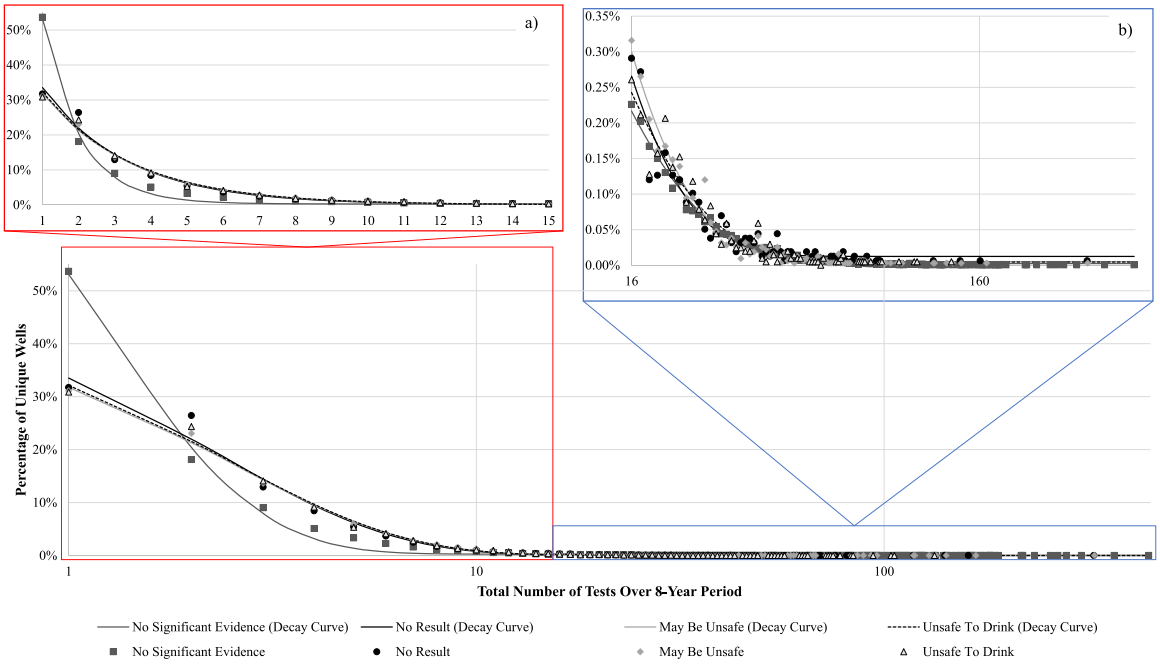


**Fig. 3.** Percentage of individual wells versus number of times tested over the eight-year period given an initial sample that was "no significant evidence" (73% of wells), "no result" (6%), "may be unsafe" (13%), or "unsafe to drink" (8%). Insets show a) under two times per year threshold (i.e., 1-15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

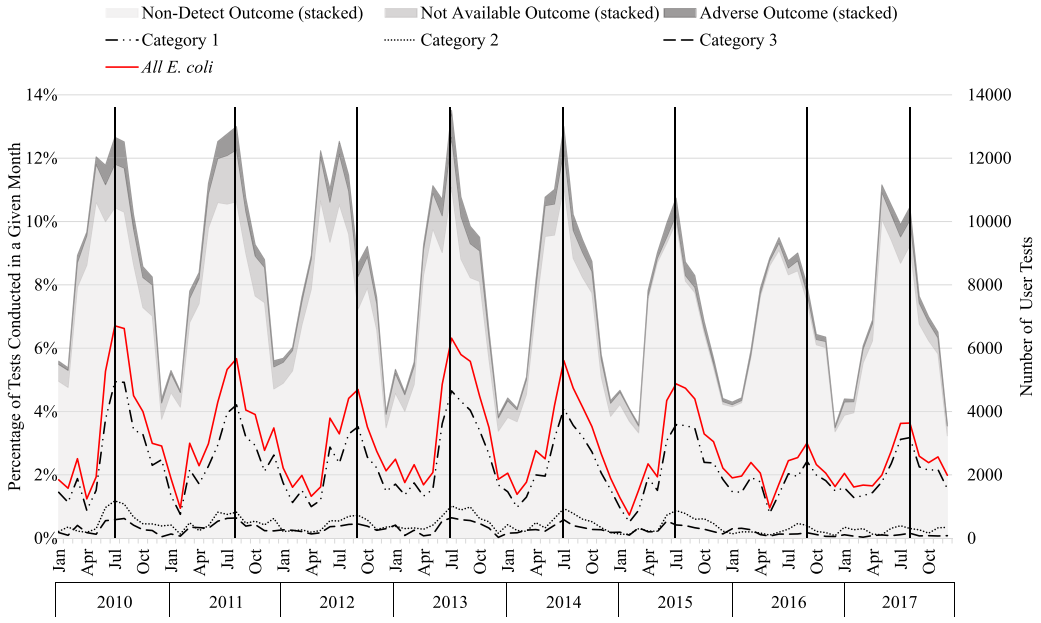


**Fig. 4.** Occurrence of private well testing and adverse results over the study period, where non-detects are defined as 0 CFU/100mL, "All *E. coli*" represents the summation of the three *E. coli* concentration categories. Vertical bars are extensions of peak adverse points each year.

(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 (Fig. 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 S2.11), 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 (Fig. 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.

## 4. Study limitations

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.

## 5. 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 inter-annual 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.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors wish to acknowledge Dr. Ahmed Yosri Ahmed for his review of this manuscript, and the Philomathia Foundation for funding for this work.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117089.

## References

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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. University of Calgray doi: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. doi:10.1016/j.scitotenv.2019.03.045.

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., 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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). Appl. Statist. 54 (part 3), 507–554. doi: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. doi:10.1007/s10533-015-0116-1.

Sainani, K.L., 2014. Explanatory versus predictive modeling. PM R 6, 841–844. doi: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. doi: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. J. Stat. Softw. 23, 1–46. doi:10.18637/jss.v023.i07.

UN Water, 2015. Water for a sustainable world. United Nations World Water Dev. 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. doi:10.1016/j.advwatres.2020.103800.