Science of the Total Environment 717 (2020) 134599

Contents lists available at ScienceDirect

Science of the Total Environment

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

Modeling the persistence of viruses in untreated groundwater

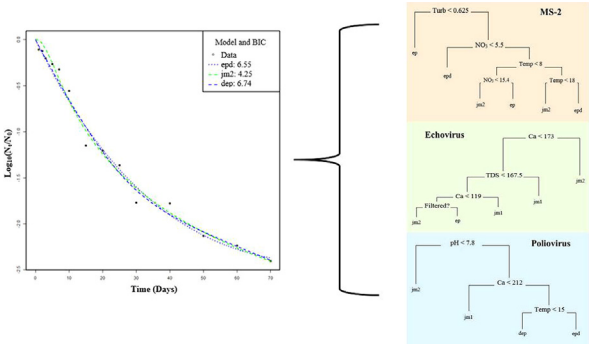
Kara Dean [a], Austin Wissler [a,b], J. Sebastian Hernandez-Suarez [a], A. Pouyan Nejadhashemi [a], Jade Mitchell [a,*]

[a] Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, MI, United States
[b] Water Resources Graduate Program, Oregon State University, Corvallis, OR, United States

HIGHLIGHTS

- Virus persistence in water has inactivation kinetics not fully captured by linear models.
- Three nonlinear models best fit viral persistence data most frequently.
- Classification trees provide a novel framework for persistence model selection.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 July 2019
Received in revised form 18 September 2019
Accepted 20 September 2019
Available online 22 November 2019

Editor: Damia Barcelo

Keywords:
Persistence
Inactivation
Viruses
Indicators
Decay rate

ABSTRACT

Several factors can affect virus behavior and persistence in water sources. Historically linear models have been used to describe persistence over time; however, these models do not consider all of the factors that can affect inactivation kinetics or the observed patterns of decay. Meanwhile, applying the appropriate persistence model is critical for ensuring that decision makers are minimizing human health risk in the event of contamination and exposure to contaminated groundwater. Therefore, to address uncertainty in predictions of decay or virus concentrations over time, this study fit seventeen different linear and nonlinear mathematical models to persistence data from a previously conducted sampling study on drinking water wells throughout the United States. The models were fit using Maximum Likelihood Estimation and the best fitting models were determined by the Bayesian Information Criterion. The purpose of the study was to identify the best model for estimating decay of viruses in groundwater and to determine if model uncertainty contributes to erroneous predictions of viral contamination when only conventional models are considered. For the datasets analyzed in this study, the Juneja and Marks models and the exponential damped model were more representative of the persistence of viruses in groundwater than the traditionally used linear models. The results from this study were then evaluated with classification trees in order to identify more relevant modeling methodology for future research. The classification trees aid in narrowing the scope of appropriate persistence models based on characteristics of the experimental conditions and water sampled.

1. Introduction

From 1971 to 2008, 30% of the waterborne disease outbreaks in the U.S. were from the consumption of untreated groundwater, causing at least 23,478 infections (Murphy et al., 2017;

* Corresponding author at: 524 S. Shaw Lane, 206 Farrall Hall, Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, MI, United States.
E-mail address: jade@msu.edu (J. Mitchell).

https://doi.org/10.1016/j.scitotenv.2019.134599
0048-9697/

Wallender et al., 2014). Of the 42 waterborne disease outbreaks reported in the U.S. between 2013 and 2014, three classified untreated groundwater as the system deficiency and these outbreaks resulted in 70 cases of illness (Benedict et al., 2017). The Ground Water Rule does not require disinfection for all groundwater systems, but rather requires ground water systems at risk of fecal contamination to take corrective action such as providing an alternative source of water, eliminating the contamination, or providing treatment (U.S. Environmental Protection Agency, 2006). Sources of possible contamination include septic tank effluent and leaks, agricultural run-off, and permeating surface and storm water (U.S. Environmental Protection Agency, 2006). In the United States alone, over 13 million households utilize private wells for drinking water (U.S. EPA, 2018). Wells are susceptible to various kinds of contamination and recent studies show there may be a lack of "protective action" undertaken by owners, including proper management, maintenance, testing, and treatment (Murphy et al., 2017; Kreutzwiser et al., 2011). To ensure the safety of well users, it is important to understand the potential pathogenic hazards that may pose risks in contaminated groundwater.

Groundwater sources are at risk of contamination from over 100 viral and several bacterial pathogens that are known to be a concern (Macler and Merkle, 2000). Enteric viruses can be shed in large quantities from infected individuals and their small size facilitates transport through the environment (World Health Organization, 2017; Murphy et al., 2017; Gerba and Rose, 1990). Enteric viruses are one of the most common causes of human infection and as such, there is a growing interest in determining how the persistence of these pathogens in water can be modeled to aid in predicting drinking water safety (World Health Organization, 2017; Ratjar et al., 2008).

Viruses are not specifically measured in many drinking water sources because of resource constraints and complex assays (Leclerc et al., 2000). Instead of monitoring viruses directly, it is common practice to use indicator bacteria (i.e., fecal coliforms, *E. coli*, etc.) as a surrogate for determining contamination of a water source. Indicators and surrogates are useful when under similar conditions they exhibit the same behavior as pathogenic species. If an indicator bacteria does not exhibit decay behavior the same as the viruses of concern, there is potential for erroneous predictions of remaining concentrations after a given time period and misleading assessments of risk. Hence, understanding how common pathogenic agents and their respective indicators persist in aqueous environments may help public health officials prevent and predict waterborne disease outbreaks.

Quantitative microbial risk assessment (QMRA) is a formal process widely used in water quality for evaluating sources of pathogenic hazards and their transport through the environment and across exposure pathways leading to consumption by humans (Haas et al., 2014). QMRA seeks to integrate scientific information into a computational framework describing microbial behaviors and interactions across these pathways that include pathogen fate and therefore, persistence or inactivation within the environment. As with all modeling approaches, it is important to understand uncertainty in the context of (1) data generation (or experimental error) and variability; (2) selection and evaluation of model structures; and (3) parameter estimation. This study deals with the latter two as previous researchers have addressed the former in the context of inactivation (Gronewold et al., 2011).

Microbial inactivation model structures generally consists of four types of survival curves: linear, curves with shoulders, curves with tails, and sigmoidal curves (Xiong et al., 1999). Additionally, empirical models describing multiple population dynamics have emerged from studies under varying experimental factors such as initial concentration, temperature, or relative humidity (Mitchell and Akram, 2017; Enger et al., 2018). As a generalizable form, these models consist of one, two or three parameters. Because environmental and sampling conditions have an impact on decay rates, assessments of decay may consist of studies that utilize calculated decay rates (from an assumed model) and incorporate statistical analyses to explain variability associated with these explanatory variables (Brooks and Field, 2016). In an attempt to understand the persistence of viruses in groundwater sources, the effect of several factors including the climate and soil characteristics must be considered. Temperature can also affect observed persistence behavior, as well as the possible presence or absence of indigenous bacteria (Yates et al., 1985; Yates et al., 1990).

This study specifically addresses the persistence of poliovirus 1, echovirus 1, and the bacteriophage MS-2 from a previously conducted study on well water from multiple locations across the United States (Yates, 1984; Yates et al., 1985). These microorganisms were chosen by Yates (1984) because at the time the *Picornaviridae* family was the most widely studied group of enteric viruses in sewage (Yates, 1984). The enteroviruses of the *Picornaviridae* family have been associated with a range of outcomes including asymptomatic infections, mild respiratory illness, fever, meningitis, or myocarditis (Abbaszadegan and Alum, 2016; Kocwa-Haluch, 2001; Yates, 1984). Enteroviruses are the second most common cause of viral infections, causing an estimated 30–50 million enteroviral infections per year in the U.S., and have been listed by the U.S. Environmental Protection Agency (EPA) on the fourth Contaminant Candidate List (Ratjar et al., 2008, U.S. Environmental Protection Agency, 2016). The bacteriophage MS-2 was chosen because it is used as an indicator or surrogate for viral pollution in water sources (Kott et al., 1974; Nkwe et al., 2015). Comparing the persistence of the bacteriophage to the viruses assesses the validity of using MS-2 as an indicator for virus persistence in groundwater, a practice with questionable reliability (Leclerc et al., 2000).

In this paper, 17 linear and non-linear mathematical models were fit to persistence data of poliovirus 1, echovirus 1, and MS-2 with the purpose of determining whether model uncertainty can contribute to erroneous predictions of viral concentrations when conventional models are used. Historically, log-linear models have been used to describe the persistence of microbial populations, and although sufficient as an estimator, these models do not take into account the presence of sub-populations within a sample that may present different inactivation kinetics (Mitchell and Akram, 2017). Previous studies have cited a reliance on first order decay rates as a limitation, as it only allows for an analysis of data uncertainty and not model uncertainty (Brooks and Fields, 2016). After the best fitting models were determined for each dataset, a decision analytic tool, Classification and Regression Trees (CART), were used to describe which model best fit each dataset according to the experimental conditions and sample characteristics. The intention of this analysis is to make the findings of this study easier to implement in future research studies evaluating decay of viruses in untreated drinking water from groundwater sources.

2. Materials and methods

2.1. Data collection

Persistence data from Yates (1984) and Yates et al. (1985) for MS-2, poliovirus 1, echovirus 1 were data mined. Yates (1984) collected one-liter groundwater samples from wells in six different states in the United States. Physical and chemical analyses included measurements of nitrates, ammonia, pH, magnesium hardness, calcium hardness, total hardness, total dissolved solids and turbidity and were performed after collection. Separate water

samples of 50 mL were inoculated with poliovirus 1, echovirus 1 and MS-2 to achieve final concentrations of 10^4–10^6 plaque forming units (PFU)/mL and then incubated at the desired temperature between 4 °C and 23 °C for the duration of the experiment. Before the viruses or bacteriophage were added and incubated, 33 of the samples were filtered to remove any indigenous bacteria and the remaining 43 samples were left unfiltered. Assays were performed on days 1, 2, 3, 5, 7, 10, 15, 20, 25, and 30 and concentrations were recorded in $\log_{10}$ (PFU/mL). The individual experiments conducted by Yates (1984) were analyzed in this study as separate datasets.

2.2. R persistence modeling procedure

The statistical programming language, R, version 3.4.3 (R Development Core Team, 2017), was used as a modeling tool to fit the 17 linear and nonlinear persistence models shown in Table 1 to each dataset. The code optimized the fit of the models to each of the 76 datasets using Maximum Likelihood Estimation (MLE). Eq. (1) is the likelihood of a model (L_m) where θ represents the parameter values of the model, x represents the data and M is the model used (Mitchell and Akram, 2017). MLE optimizes the likelihood that the observed data were generated from a model with parameter values θ at the global maximum of the likelihood function. If the parameters of a model failed to converge to a single value, the model was excluded from the analysis.

$$L_m = p(x|\theta, M) \tag{1}$$

To compare the fits of different models to each dataset, the Bayesian Information Criterion (BIC) was used, where the lowest absolute BIC value corresponds with the best fitting model from those tested. The BIC is defined by Eq. (2) where k is the total number of parameters, n is the number of data points in the observed data (x), and L_m is the maximum likelihood of the model. If two or more models have BIC values that differ by less than 2, there is not enough evidence to consider them statistically different and they are treated as equal selections in terms of fit (Dziak et al., 2012). BIC was chosen as the criterion for model selection in this study because the datasets have a low number of observations and the BIC provides a greater penalty for models with additional parameters in order to identify the most parsimonious model (Dziak et al., 2012).

$$\mathrm{BIC} = k \ln(n) - 2\ln(L_m) \tag{2}$$

A visual inspection of the output graphics for each dataset along with the BIC values for the models fit, made it possible to determine approximately three to five best fitting models for each dataset. However, BIC comparisons do not indicate an overall goodness of fit for any of the models tested, rather a relative ranking of the best fitting models. In order to analyze the goodness of fit for the models with the lowest BICs, the likelihood ratio test was performed. The statistical deviance of the model was compared to the χ^2 test statistic at the degrees of freedom for a significance level of 0.05 given the number of observed data points and the number of parameters of each model.

2.3. T90 calculations

A T90 value is the amount of time it takes for a one-log or 90% reduction in the concentration of a pathogen and it is generally relevant in the linear region of the decay curve (Mitchell and Akram, 2017). T90 values were calculated from the best fitting model of each dataset, unless modeling resulted in more than one best fitting model according to the BIC values. In this case, model averaging was completed to find the most representative T90 value. The model averaging was done using the following methodology outlined in Haas et al. (2014) that assigns a weighting value, w_i, to each of the best fitting models. The weighting value is calculated using Eq. (3). Multiplying each T90 value by the model's weighting value and summing them gives a final model averaged T90 value.

$$w_i = \frac{\exp(-\frac{1}{2}BIC_i)}{\sum_{j-1}^{i}\exp(-\frac{1}{2}BIC_j)} \tag{3}$$

The resulting T90 values for all datasets were used as a metric to compare persistence across each pathogen and each combination of environmental conditions (location, filtered vs. nonfiltered,

Table 1
Models Fit to Persistence Data.

Model	Notation	Equation[a]	Reference
Exponential	ep	$\frac{N_t}{N0} = \exp(-k_1 t)$	(Chick, 1908)
Logistic	lg1	$\frac{N_t}{N0} = \frac{2}{1+\exp(-k_1 t)}$	(Kamau et al., 1990)
Fermi	lg2	$\frac{N_t}{N0} = \frac{1}{1+\exp(k_1(t-k_2))}$	(Peleg, 1995)
Exponential damped	epd	$\frac{N_t}{N0} = \exp(-k_1 t \times \exp(-k_2 t))$	(Cavalli-Sforza et al., 1983)
Juneja and Marks 1	jm1	$\frac{N_t}{N0} = 1-(1-\exp(-k_1 t))^{k_2}$	(Juneja et al., 2001)
Juneja and Marks 2	jm2	$\frac{N_t}{N0} = \frac{1}{1+\exp(k_1+k_2\ln(t))}$	(Huang et al., 2006)
Gompertz 2	gz	$\frac{N_t}{N0} = \exp\left[-\frac{k_1}{k_2}(\exp(k_2 t)-1)\right]$	(Wu et al., 2004)
Weibull	wb	$\frac{N_t}{N0} = 10^{-((t/k_1)^{k_2})}$	(Peleg, 2003)
Lognormal[b]	ln	$\frac{N_t}{N0} = 1-0.5(1+\mathrm{erf}(\frac{\ln(t)-\mu}{\sqrt{2}\sigma}))$	(Aragao et al., 2007; R Core Team, 2018)
Gamma[c]	gam	$\frac{N_t}{N0} = \frac{1}{\Gamma z}\int_0^t t^{(\alpha-1)}\exp(-t)dt$	(van Gerwen and Zwietering, 1998; R Core Team, 2018)
Broken-line	bi	$\frac{N_t}{N0} = \exp(-k_1 t), t < k_3$	(Muggeo, 2003)
Broken-line 2	bi2	$\frac{N_t}{N0} = \exp(-k_1 t + k_2(t-k_3)), t \geq k_3$	(Muggeo, 2003)
Double exponential	dep	$\frac{N_t}{N0} = k_3\exp(-k_1 t)+(1-k_3)\exp(-k_2 t)$	(Abraham et al., 1990)
Gompertz 3	gz3	$\frac{N_t}{N0} = 10^{k_1\exp\left[-\exp\left(\frac{-k_2\exp(1)(k_3-t)}{k_1}+1\right)\right]}$	(Gil et al., 2011)
Gompertz-Makeham	gzm	$\frac{N_t}{N0} = \exp\left(-k_3 t - \frac{k_1}{k_2}(\exp(k_2 t)-1)\right)$	(Jodrá, 2009)
Sigmoid type A	sA	$\frac{N_t}{N0} = 10^{((-k_1 t)/\{(1+k_2 t)(k_3-t)\})}$	(Peleg, 2006)
Sigmoid type B	sB	$\frac{N_t}{N0} = 10^{-(k_1\times t^{k_3})/(k_2+t^{k_3})}$	(Peleg, 2006)

[a] N_t denotes the number of organisms remaining at time t; N0 denotes the initial number of organisms; k_1, k_2, and k_3 are model parameters. Model parameters, k_i, are constants in each model that have different units depending on the model.
[b] The lognormal equation was calculated in R using the *plnorm()* function.
[c] For the gamma model, the approximation shown was given in the RDocumentation for the gamma distribution (R Core Team, 2018) and was calculated in R using the *pgamma()* function.

temperature), as the T90 is often presented as a metric in microbiological studies. Additionally, the calculation of the T90 value provides a consistent metric for comparison of T90 values or decay rates previously published for pathogens when the log-linear model was used to the alternative best-fitting models determined in this study.

A three-way ANOVA was completed using the aov function from the stats package in R to determine the effect of the three independent variables (temperature, location, and filter status) on the dependent variable (T90) at a significance level of 0.05 (R Core Team, 2018). There were enough MS-2 datasets to also analyze the effects of any possible interactions between the three independent variables. For this analysis, the independent variable "location" is used to represent all of the sampling sites for monitoring water quality data. This variable is further divided into the individual water quality variables in the following classification tree analysis.

2.4. Classification trees

Once the set of best models for each dataset were obtained, a classification tree was generated for each of the two viruses and the bacteriophage using the type of best fit model as the response variable. Factors such as temperature, sample filtration, pH, turbidity, total dissolved solids, and concentrations for ammonia, nitrates, and hardness (total, calcium, and magnesium) were considered the explanatory variables. This approach attempted to identify the most suitable persistence model given attributes describing each sample's general water quality and the laboratory test conditions.

The tree-based classification method utilized the Classification and Regression Trees (CART) algorithm (Breiman et al., 1984). CART algorithms divide the predictor space into a set of M rectangular regions and then fits a simple model in each region (Hastie et al., 2009). When working with classification trees, the response variable must be categorical. Trees are usually grown to a maximum by repeatedly performing binary partitioning, resulting in a dendrogram with M terminal nodes, one per region. CART algorithms work top-down by choosing an optimal combination of the splitting variable and split location at each step that best matches the distribution of classes at each node. Then, a cross-validation approach is employed to prune the tree model by collapsing its internal nodes while preventing overfitting. The latter process is commonly referred to as cost-complexity pruning (Ripley, 1996). Growing and pruning classification trees require the definition of a node impurity measure based on the proportion of observations, p_{mk}, with class k in node m:

$$p_{mk} = \frac{1}{N_m} \sum_{xi \in R_m} I(y_i = k) \tag{4}$$

where, N_m is the number of observations in node m, R_m is the region corresponding to node m, x_i are the values for the explanatory variables from the predictor space, y_i are the response variable observations related to x_i, and I is the indicator function. The majority class $k(m)$ is the one with the largest p_{mk}. Three common node impurity measures are the misclassification error ME, the Gini index GI and the deviance (entropy) D given by Eqs. (5)–(7), respectively.

$$ME = 1 - p_{mk(m)} \tag{5}$$

$$GI = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{6}$$

$$D = -\sum_{k=1}^{K} p_{mk} \log(p_{ml}) \tag{7}$$

In general, these three measures are similar, but GI and D are differentiable and more sensitive to changes in node probabilities than ME (Hastie et al., 2009). Therefore, GI was used in this study for growing the tree while ME was used for pruning. For this purpose, the package tree (Ripley, 2018) in R (version 4.3.3) was implemented.

3. Results

3.1. Best fitting model frequency

For the MS-2 data, the most identified best fitting models were jm2, epd and ep. The poliovirus 1 data had similar best fitting models with jm1, jm2, dep, and epd identified most frequently. Finally, the best fitting models for echovirus 1 were most commonly jm1, jm2, epd, and ln. For all three of the pathogens, jm1, jm2, and epd were some of the highest occurring best fitting models. Jm2 specifically, was the most identified best fitting model for MS-2, poliovirus 1, and echovirus 1, fitting 47.22% (n = 36), 50% (n = 20), and 65% (n = 20) of the datasets, respectively.

When comparing nonfiltered versus filtered samples, the most common best fitting models slightly differed. For the nonfiltered samples: (1) jm2 fit 38.10% (n = 21) of the MS-2 datasets; (2) jm1 fit 63.64% (n = 11) of the poliovirus 1 data; and (3) jm2 fit 54.55% (n = 11) of the echovirus 1 data. For the filtered samples, jm2 fit the data most frequently, fitting at least 60% of the data for MS-2, poliovirus 1, and echovirus 1.

Ten of the models tested were found to be the best fit for at least one dataset. The models that fit with the lowest frequencies were ep, lg1, ln, wb, and gz3. Of note, the exponential model (ep) is a commonly used model in persistence studies (Mitchell and Akram, 2017; Liang et al., 2017; Walters et al., 2009) and was found to be the best fit for only 30.56%, 5%, and 25% of the MS-2, poliovirus 1, and echovirus 1 datasets, respectively. The Weibull (wb) model, another commonly preferred persistence model for pathogens, was identified as the best fit model for 30% or less of the data for all three microorganisms (Bozkurt et al., 2014a, 2014b; Kingsley et al., 2018; Peleg and Cole, 1998). Six of the 17 models were never determined to be the best fitting model – lg2, gz, bi, bi-3, gzm, sB and sA. The best fitting model(s) for each dataset for each pathogen and the respective parameters are provided in Tables 2–4. The 95% confidence interval for each parameter is also provided in these tables.

3.2. Goodness of fit

Based on the lowest BIC value of the models fit to the datasets, one or several best fitting models were determined for each dataset as a relative selection. It was necessary to assess the goodness-of-fit of the model(s) for each dataset as to determine if any or all of the best fitting models provided a statistically poor fit to a particular dataset. A statistically poor fit indicates there is uncertainty about the decay of the pathogen and the calculated decay rate, and that the model should not be used. A likelihood ratio test identified a model with a statistically good fit for 60 of the 76 datasets. Of the 36 MS-2 datasets, 7 did not have a model with a good fit; 3 of 20 poliovirus 1 datasets did not have a good fit; and 6 of 20 echovirus 1 datasets did not have a good fit. There were no clear, consistent factors of the datasets that could explain the lack of fit. Some of the datasets exhibited true shoulders, such as the persistence of echovirus 1 in groundwater from Texas at 13 °C shown in Fig. 1, and tails such as the one shown in Fig. 2.

Table 2
Models and Parameters for MS2 Datasets

Data Set	Model	k1 (95%CI)	k2 (95% CI)	k3 (95%CI)	Good fit?
Arizona, Nonfiltered, 4 °C	jm2	−2.93 (−4.14, −1.71)	2.75 (2.36, 3.13)		Yes
	epd	0.34 (0.29, 0.38)	0.015 (0.011, 0.018)		Yes
	wb	6.12 (6.08, 6.16)	0.60 (0.53, 0.68)		Yes
Arizona, Nonfiltered, 12 °C	jm2	−1.05 (−1.81, −0.29)	2.78 (2.42, 3.13)		Yes
Arizona, Nonfiltered, 23 °C	lg1	1.42 (1.36, 1.47)			Yes
	jm1	1.38 (1.30, 1.46)	1.62 (1.14, 2.11)		Yes
	Wb	2.00 (1.98, 2.02)	1.10 (0.98, 1.23)		Yes
Wisconsin, Nonfiltered, 4 °C	gz3	−1.44 (−1.89, −0.98)	0.03 (−0.15, 0.22)	16.42 (16.21, 16.64)	Yes
	gam	10.44 (10.39, 10.50)	2.80 (1.89, 3.71)		Yes
Wisconsin, Nonfiltered, 12 °C	ep	0.21 (0.20, 0.22)			Yes
	epd	0.23 (0.19, 0.26)	0.002 (−0.002, 0.007)		Yes
Wisconsin, Nonfiltered, 23 °C	epd	0.21 (−0.04, 0.45)	−0.07 (−0.16, 0.02)		No
	lg1	0.57 (0.43, 0.70)			No
	ep	0.50 (0.37, 0.64)			No
	wb	7.09 (7.00, 7.17)	1.75 (0.48, 3.02)		No
California 2, Filtered, 17 °C	ep	0.49 (0.45, 0.52)			Yes
	jm1	0.46 (0.34, 0.59)	0.73 (0.10, 1.36)		Yes
	gam	2.25 (2.22, 2.28)	0.79 (0.41, 1.18)		Yes
	epd	0.52 (0.40, 0.64)	0.004 (−0.01, 0.02)		Yes
California 2, Nonfiltered, 17 °C	jm1	0.17 (0.05. 0.30)	0.28 (−0.26, 0.81)		No
	gam	6.73 (6.69, 6.76)	0.35 (0.17, 0.53)		No
California 1, Filtered, 18 °C	gam	20.12 (6.69, 6.76)	0.34 (0.27, 0.41)		Yes
	jm1	0.06 (−0.07, 0.20)	0.32 (0.08, 0.55)		Yes
California 1, Nonfiltered, 18 °C	gz3	−2.44 (−2.54, −2.34)	0.12 (0.01, 0.22)	3.72 (3.44, 4.00)	Yes
North Carolina 1, Filtered, 4 °C	jm2	−6.93 (−9.19, −4.67)	2.36 (1.76, 2.96)		Yes
	lg1	0.06 (0.051, 0.064)			Yes
	ln	2.93 (2.68, 3.18)	0.76 (0.57, 0.95)		Yes
	ep	0.04 (0.039, 0.051)			Yes
North Carolina 1, Filtered, 12 °C	jm2	−3.76 (−4.45, −3.07)	2.19 (1.99, 2.39)		Yes
North Carolina 1, Filtered, 23 °C	lg1	0.59 (0.51, 0.66)			No
	ep	0.54 (0.46, 0.62)			No
	epd	0.39 (0.20, 0.58)	−0.02 (−0.05, 0.01)		No
	gam	1.42 (1.35, 1.49)	1.99 (−0.11, 4.08)		No
North Carolina 1, Nonfiltered, 4 °C	ep	0.033 (0.029, 0.037)			Yes
	lg1	0.044 (0.040, 0.049)			Yes
North Carolina 1, Nonfiltered, 12 °C	jm2	−11.09 (−12.77, −9.41)	3.66 (3.22, 4.10)		Yes
North Carolina 1, Nonfiltered, 23 °C	epd	0.23 (0.18, 0.29)	−0.03 (−0.04, −0.02)		Yes
North Carolina 2, Filtered, 4 °C	epd	0.01 (−0.18, 0.20)	−0.02 (−0.05, 0.01)		Yes
North Carolina 2, Filtered, 12 °C	ln	2.85 (2.69, 3.01)	0.60 (0.51, 0.69)		Yes
	lg1	0.08 (0.07, 0.08)			Yes
	jm1	0.08 (−0.06, 0.22)	2.21 (1.71, 2.70)		Yes
	jm2	−9.41 (−11.39, −7.42)	3.24 (2.72, 3.77)		Yes
	gam	10.86 (10.81, 10.91)	1.82 (1.17, 2.48)		Yes
North Carolina 2, Filtered, 23 °C	lg1	0.57 (0.50, 0.64)			No
	wb	6.26 (6.21, 6.31)	1.38 (1.02, 1.74)		No
	gam	1.24 (1.17, 1.30)	2.89 (0.32, 5.46)		No
	ln	1.47 (1.07, 1.87)	0.36 (0.25, 0.48)		No
	epd	0.33 (0.17, 0.50)	−0.03 (−0.06, 0.001)		No
North Carolina 2, Nonfiltered, 4 °C	ep	0.033 (0.028, 0.039)			Yes
	lg1	0.045 (0.038, 0.051)			Yes
North Carolina 2, Nonfiltered, 12 °C	jm2	−18.51 (−23.90, −13.12)	7.12 (5.62, 8.61)		No
North Carolina 2, Nonfiltered, 23 °C	jm2	−21.42 (−27.49, −15.34)	11.11 (8.82, 13.41)		No
New York 1, Filtered, 12 °C	epd	0.10 (0.09, 0.12)	0.01 (0.005, 0.009)		Yes
	ln	1.99 (1.82, 2.16)	1.01 (0.91, 1.11)		Yes
New York 1, Nonfiltered, 12 °C	jm2	−5.09 (−5.92, −4.26)	2.30 (2.06, 2.53)		Yes
	ln	2.07 (1.91, 2.22)	0.90 (0.82, 0.99)		Yes
	dep	−0.01 (−0.10, 0.09)	0.09 (−0.02, 0.20)	0.01 (−0.27, 0.28)	Yes
New York 2, Filtered, 12 °C	lg1	0.14 (0.13, 0.14)			Yes
	ep	0.12 (0.11, 0.13)			Yes
	jm2	−12.05 (−14.68, −9.41)	4.73 (4.01, 5.45)		Yes
	jm1	0.13 (0.02, 0.24)	1.48 (0.91, 2.06)		Yes
New York 2, Nonfiltered, 12 °C	jm1	0.08 (−0.07, 0.24)	0.56 (0.07, 1.05)		Yes
	ep	0.10 (0.09, 0.10)			Yes
Texas 1, Filtered 13 °C	epd	0.22 (0.18, 0.25)	0.005 (.001, .009)		Yes
	jm2	−7.79 (−9.63, −5.94)	4.07 (3.50, 4.64)		Yes
	dep	0.09 (−0.02, 0.20)	0.21 (0.18, 0.24)	0.02 (−0.41, 0.44)	Yes
	ep	0.18 (0.17, 0.20)			Yes
Texas 1, Nonfiltered 13 °C	epd	0.10 (0.08, 0.12)	−0.02 (−0.02, −0.01)		Yes
Texas 2, Filtered 13 °C	epd	1.00 (0.87, 1.13)	0.03 (0.02, 0.04)		Yes
	jm2	−2.16 (−3.42, −0.89)	4.34 (3.80, 4.88)		No
Texas 2, Nonfiltered 13 °C	jm1	0.37 (0.23, 0.50)	59.10 (57. 51, 60.69)		No
	jm2	−33.29 (−40.05, −26.52)	11.99 (10.05, 13.93)		No
	gam	2.11 (2.07, 2.14)	4.84 (2.82, 6.87)		No
University of Arizona, Filtered, 4 °C	jm2	−11.48 (−14.52, −8.44)	3.74 (2.94, 4.54)		Yes

(*continued on next page*)

Table 2 (*continued*)

Data Set	Model	k1 (95%CI)	k2 (95% CI)	k3 (95%CI)	Good fit?
	ln	3.03 (2.82, 3.24)	0.52 (0.41, 0.64)		Yes
University of Arizona, Filtered, 12 °C	jm2	−8.00 (−10.96, −5.04)	3.92 (3.07, 4.77)		No
University of Arizona, Filtered, 23 °C	jm2	−0.76 (−2.15, 0.64)	4.85 (4.11, 5.60)		No
	epd	1.67 (1.39, 1.95)	0.05 (0.03, 0.06)		No
University of Arizona, Nonfiltered, 4 °C	jm2	8.09 (−9.93, −6.24)	2.75 (2.26, 3.24)		Yes
University of Arizona, Nonfiltered, 12 °C	epd	0.14 (0.12, 0.16)	0.01 (0.005, 0.010)		Yes
	dep	0.02 (−0.03, 0.07)	0.12 (0.10. 0.14)	0.01 (−0.26, 0.28)	Yes
University of Arizona, Nonfiltered, 23 °C	ep	0.72 (0.66, 0.78)			Yes
	lg1	0.85 (0.79, 0.92)			Yes
	epd	0.63 (0.44, 0.83)	−0.02 (−0.07, 0.03)		Yes
	gam	1.23 (1.19, 1.27)	1.23 (0.64, 1.82)		Yes
	jm1	0.77 (0.60, 0.95)	1.30 (0.67, 1.92)		Yes

Table 3
Models and Parameters for Poliovirus Datasets.

Dataset	Model	k1 (95% CI)	k2 (95% CI)	k3 (95%CI)	Good Fit?
Arizona, Nonfiltered, 23 °C	ep	0.85 (0.77, 0.93)			Yes
	jm1	0.80 (0.61, 0.99)	0.71 (−0.18, 1.60)		Yes
	epd	0.95 (0.67, 1.23)	0.01 (−0.02, 0.05)		Yes
Wisconsin, Nonfiltered, 12 °C	gam	10.24 (10.18, 10.30)	0.24 (0.07, 0.41)		No
	jm1	0.12 (−0.12, 0.36)	0.20 (−0.51, 0.92)		No
	dep	0.12 (0.09, 0.15)	0.91 (−0.15, 1.97)	0.20 (0.11, 0.30)	No
	ln	0.48 (−0.26, 1.23)	1.11 (0.81, 1.42)		No
California 1, Filtered, 18 °C	dep	0.07 (0.01, 0.13)	0.82 (0.57, 1.08)	0.04 (−0.05, 0.13)	Yes
	jm2	0.27 (−0.66, 1.19)	1.52 (1.11, 1.93)		No
	epd	0.73 (0.54, 0.92)	0.06 (0.04, 0.07)		No
California 1, Nonfiltered, 18 °C	wb	2.48 (2.42, 2.54)	0.60 (0.43, 0.77)		Yes
	epd	0.93 (0.66, 1.19)	0.05 (0.02, 0.08)		No
	jm1	0.36 (0.06, 0.66)	0.19 (−0.70, 1.08)		No
	jm2	0.21 (−1.44, 1.86)	2.26 (1.39, 3.13)		No
California 2, Filtered, 17 °C	jm2	−1.43 (−2.34, −0.52)	2.31 (1.95, 2.67)		Yes
California 2, Nonfiltered, 17 °C	epd	0.63 (0.49, 0.77)	0.05 (0.04, 0.06)		Yes
	dep	0.56 (0.43, 0.69)	0.03 (−0.01, 0.07)	0.98 (0.97, 0.98)	Yes
	jm2	−0.53 (−1.50, 0.45)	1.73 (1.31, 2.15)		Yes
North Carolina 1, Filtered, 12 °C	jm1	0.39 (0.22, 0.55)	0.18 (−0.48, 0.84)		No
	gam	3.10 (3.06, 3.14)	0.23 (0.08, 0.38)		No
	dep	0.38 (0.32, 0.44)	2.03(0.66, 3.39)	0.17 (0.09, 0.25)	No
North Carolina 1, Nonfiltered, 12 °C	jm1	0.30 (0.09, 0.51)	0.08 (−0.58, 0.74)		No
	gam	4.28 (4.23, 4.32)	0.10 (0.03, 0.17)		No
	dep	0.29 (0.23, 0.36)	2.10 (0.94, 3.26)	0.07 (−0.01, 0.15)	Yes
	wb	1.55 (1.48, 1.63)	0.50 (0.38, 0.62)		No
North Carolina 2, Filtered, 12 °C	ln	0.16 (−0.52, 0.84)	0.85 (0.60, 1.11)		No
	epd	0.69 (0.44, 0.93)	0.03 (0.01, 0.05)		No
	wb	3.35 (3.16, 3.54)	0.67 (−0.40, 1.73)		No
	jm1	0.36 (0.06, 0.66)	0.34 (−0.80, 1.48)		No
	jm2	−1.05 (−3.10, 0.99)	2.77 (1.82, 3.71)		No
North Carolina 2, Nonfiltered, 12 °C	jm1	0.22 (−0.01, 0.46)	0.23 (−0.35, 0.82)		Yes
New York 1, Filtered, 12 °C	epd	0.14 (0.07, 0.20)	0.03 (0.01, 0.06)		Yes
	jm2	−1.96 (−3.16, −0.76)	1.04 (0.58, 1.49)		Yes
	ln	1.89 (1.47, 2.32)	1.59 (0.91, 2.26)		Yes
New York 1, Nonfiltered, 12 °C	jm2	−0.95 (−1.50, −0.40)	0.92 (0.70, 1.14)		Yes
	ln	1.01 (0.65, 1.37)	1.83 (1.43, 2.23)		Yes
New York 2, Filtered, 12 °C	gz3	−1.03 (−1.10, −0.97)	0.22 (−0.09, 0.53)	2.49 (2.20, 2.78)	Yes
New York 2, Nonfiltered, 12 °C	dep	0.01 (−0.04, 0.05)	0.65 (0.47, 0.82)	0.05 (−0.03, 0.12)	Yes
	epd	0.64 (0.50. 0.78)	0.07 (0.05, 0.08)		Yes
Texas 1, Filtered, 13 °C	dep	0.11 (0.07, 0.15)	0.64 (0.31, 0.96)	0.19 (0.10, 0.27)	Yes
	jm2	−0.88 (−1.80, 0.05)	1.65 (1.28, 2.01)		Yes
	wb	7.88 (7.83, 7.92)	0.59 (0.45, 0.73)		Yes
Texas 1, Nonfiltered, 13 °C	jm1	0.04 (−0.46, 0.53)	0.09 (−0.34, 0.52)		Yes
	gam	39.62 (39.49, 39.76)	0.09 (0.05, 0.13)		Yes
	dep	0.07 (0.05, 0.08)	3.30 (−1.07, 7.67)	0.21 (0.17, 0.24)	Yes
Texas 2, Filtered, 13 °C	jm2	−0.25 (−1.16, 0.66)	1.69 (1.32, 2.06)		Yes
Texas 2, Nonfiltered, 13 °C	jm2	−0.52 (−1.44, 0.40)	2.32 (1.84, 2.80)		Yes
	dep	0.05 (−0.12, 0.22)	0.76 (0.59, 0.92)	0.01 (−0.15, 0.17)	Yes
	epd	0.91 (0.69, 1.12)	0.06 (0.04, 0.08)		Yes
Univ. of Arizona, Filtered, 23 °C	gam	0.53 (0.52, 0.54)	2.51 (2.19, 2.84)		Yes
	wb	2.45 (2.37, 2.53)	1.35 (1.24, 1.47)		Yes
Univ. of Arizona, Nonfiltered, 23 °C	gam	0.49 (0.47, 0.50)	2.48 (1.83, 3.13)		Yes
	wb	2.23 (2.21, 2.24)	1.33 (1.20, 1.46)		Yes
	jm1	1.72 (1.64, 1.80)	4.79 (4.19, 5.39)		Yes
	epd	0.92 (0.74, 1.10)	−0.07 (−0.10, −0.04)		Yes

Table 4
Models and Parameters for Echovirus Datasets.

Data Set	Model	k1 (95%CI)	k2 (95% CI)	k3 (95%CI)	Good fit?
Arizona, Nonfiltered, 23 °C	jm2	−2.15 (−3.81, −0.50)	3.60 (2.85, 4.35)		No
	epd	0.81 (0.60, 1.02)	0.03 (0.01, 0.05)		No
Wisconsin, Nonfiltered, 12 °C	dep	0.91 (0.59, 1.22)	0.11 (0.09, 0.13)	0.93 (0.93, 0.94)	Yes
California 1, Filtered, 18 °C	jm2	−1.37 (−2.18, −0.55)	1.57 (1.26, 1.88)		Yes
	ln	0.76 (0.33, 1.18)	1.24 (0.99, 1.49)		Yes
	dep	0.08 (0.04, 0.12)	0.43 (0.26, 0.60)	0.19 (0.09, 0.28)	Yes
California 1, Nonfiltered, 18 °C	jm2	−4.83 (−6.50, −3.17)	3.61 (2.87, 4.36)		Yes
	lg1	0.39 (0.35, 0.42)			Yes
	ln	1.19 (0.94, 1.44)	0.61 (0.48, 0.74)		Yes
California 2, Filtered, 17 °C	jm2	−3.33 (−3.90, −2.76)	2.63 (2.40, 2.87)		Yes
	ln	1.15 (0.96, 1.33)	0.79 (0.68, 0.89)		Yes
California 2, Nonfiltered, 17 °C	jm2	−2.46 (−2.92, −1.99)	2.46 (2.28, 2.63)		Yes
North Carolina 1, Filtered, 12 °C	dep	0.22 (0.19, 0.26)	3.55 (−8.77, 15.88)	0.46 (0.41, 0.51)	Yes
	jm1	0.21 (−0.02, 0.44)	0.36 (−0.19, 0.90)		Yes
North Carolina 1, Nonfiltered, 12 °C	jm1	0.41 (0.22, 0.61)	0.48 (−0.41, 1.36)		No
	ep	0.47 (0.42, 0.52)			No
	gam	2.60 (2.55, 2.65)	0.59 (0.12, 1.06)		No
	epd	0.51 (0.33, 0.69)	0.01 (−0.02, 0.03)		No
North Carolina 2, Filtered, 12 °C	ln	0.10 (−0.29, 0.49)	1.28 (1.06, 1.50)		Yes
	jm2	−0.41(−1.02, 0.21)	1.56 (1.28, 1.85)		Yes
North Carolina 2, Nonfiltered, 12 °C	ep	0.42 (0.39, 0.44)			Yes
	jm1	0.39 (0.29, 0.50)	0.74 (0.27, 1.21)		Yes
New York 1, Filtered, 12 °C	jm2	−2.06 (−2.86, −1.26)	1.45 (1.14, 1.75)		Yes
	ln	1.41 (1.12, 1.71)	1.20 (0.95, 1.46)		Yes
	epd	0.20 (0.14, 0.25)	0.03 (0.01, 0.04)		Yes
New York 1, Nonfiltered, 12 °C	jm2	−3.49 (−4.47, −2.51)	1.87 (1.52, 2.23)		Yes
	ep	0.11 (−0.09, 0.32)			Yes
	ln	1.84 (1.66, 2.03)	0.94 (0.78, 1.10)		Yes
	epd	0.12 (0.08, 0.15)	0.003 (−0.01, 0.02)		Yes
	jm1	0.11 (−0.09, 0.32)	1.05 (0.65, 1.45)		Yes
	gam	8.90 (8.84, 8.96)	1.03 (0.66, 1.41)		Yes
New York 2, Filtered, 12 °C	jm1	0.01 (−1.84, 1.86)	0.09 (−0.79, 0.97)		Yes
	gam	110.14 (109.60, 110.69)	0.09 (0.01, 0.16)		Yes
	wb	26.67 (26.42, 26.93)	0.21 (0.04, 0.39)		Yes
	jm2	0.72 (−0.04, 1.47)	0.44 (0.10, 0.78)		Yes
New York 2, Nonfiltered, 12 °C	jm1	0.09 (−0.57, 0.75)	0.34 (−0.62, 1.30)		No
	gam	14.23 (14.05, 14.41)	0.36 (0.01, 0.71)		No
	wb	14.35 (14.28, 14.42)	0.59 (0.27, 0.92)		No
	epd	0.25 (0.10, 0.39)	0.03 (0.00, 0.06)		No
	ln	0.82 (−0.29, 1.93)	1.41 (0.56, 2.26)		No
	jm2	−1.30 (−3.59, 1.00)	1.34 (0.43, 2.25)		No
Texas 1, Filtered, 13 °C	jm2	−1.56 (−2.99, −0.13)	1.86 (1.31, 2.41)		No
	jm1	0.14 (−0.16, 0.45)	0.38 (−0.38, 1.15)		No
	epd	0.31 (0.19, 0.44)	0.02 (−0.00, 0.04)		No
	dep	0.13 (0.07, 0.18)	0.57 (0.12, 1.01)	0.27 (0.13, 0.40)	No
Texas 1, Nonfiltered, 13 °C	gz3	−2.31 (−2.48, −2.14)	0.30 (0.07, 0.53)	2.39 (2.08, 2.71)	Yes
Texas 2, Filtered, 13 °C	jm2	−3.58 (−5.47, −1.69)	2.85 (2.11, 3.59)		No
	gz3	−1.99 (−2.33, −1.66)	0.45 (−0.99, 1.88)	4.38 (3.82, 4.93)	No
Texas 2, Nonfiltered, 13 °C	jm2	−1.82 (−2.97, −0.67)	1.93 (1.47, 2.39)		Yes
	ln	0.79 (0.34, 1.24)	1.06 (0.81, 1.30)		Yes
	epd	0.34 (0.24, 0.43)	0.03 (0.01, 0.04)		Yes
University of Arizona, Filtered, 23 °C	ep	1.45 (1.37, 1.52)			Yes
	epd	1.39 (1.12, 1.65)	−0.01 (−0.04, 0.02)		Yes
	jm1	1.43 (1.33, 1.54)	0.94 (0.26, 1.62)		Yes
	gam	0.69 (0.67, 0.72)	1.00 (0.53, 1.46)		Yes
University of Arizona, Nonfiltered, 23 °C	jm1	1.37 (1.18, 1.55)	0.39 (−0.70, 1.48)		No
	ep	1.56 (1.40, 1.71)			No
	gam	0.78 (0.74, 0.82)	0.54 (0.03, 1.06)		No

3.3. T90 comparisons

A T90 value was calculated for every dataset based on the best fitting models and decay rates. The model averaged T90s for the poliovirus 1 and echovirus 1 persistence data were lower than the T90s calculated from the original log-linear fits completed by Yates et al. (1985) for 85% of the datasets, indicating that the log-linear predicted decay was conservative. The model averaged T90s for MS-2, however, resulted in a lower T90 value in only 50% of the datasets. Thus for the other 50% of the datasets, the log-linear fit was underestimating the time it took to achieve one log reduction.

The three-way ANOVA test performed on the MS-2 data showed temperature significantly affected the T90 values with a p-value of 4.70E-05. Meanwhile, location and filter status were not significantly associated with the T90 value and there were no significant interactions between the three independent variables. Box plots demonstrating the variance of the T90 values based on temperature, location, and filter status are shown in Figs. 3–5, respectively. With respect to temperature, the highest T90 values were seen at 4 °C. Fig. 4 exhibits the most heteroscedasticity and Fig. 5 shows the minimal impact of filter status on the T90 value.

Temperature was also significantly associated with the T90 values for the poliovirus 1 data, with a p-value of 0.00256. Unlike the

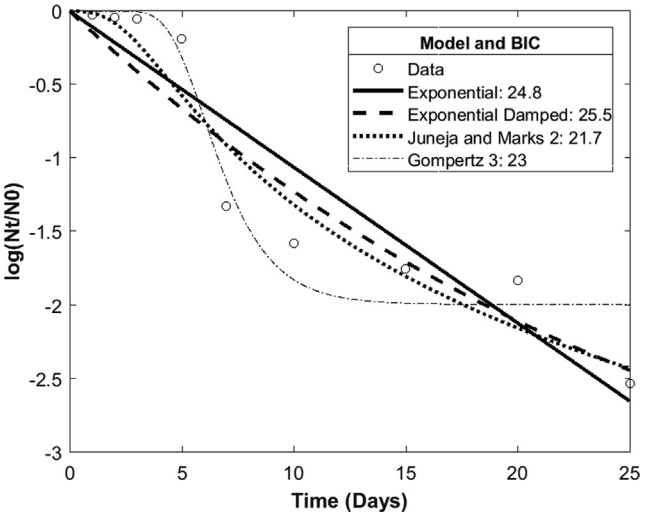


Fig. 1. A true shoulder in the persistence of echovirus 1 in Texas filtered groundwater at 13 °C.

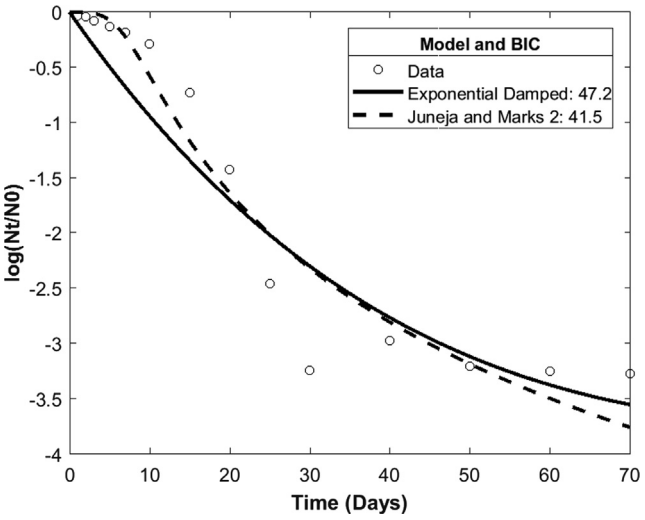


Fig. 2. A tail in the persistence of MS-2 in University of Arizona filtered groundwater at 12 °C.

MS-2 data, location was also significantly associated with the T90 values, with a p-value of 0.000319. Filter status was not significantly associated. The same trends from the MS-2 data are reflected in Figs. 6–8. Variance in the T90 values is the highest for the datasets at the lowest temperature, in this case 12 °C, and subsequently decreases as temperature increases, as shown in Fig. 6. The T90 values based on location are the most heteroscedastic, as shown in Fig. 7. Similar to the MS-2 data, filter status showed little impact as shown in Fig. 8.

Similar to the poliovirus 1 data, the echovirus 1 T90 values were significantly associated with both temperature and location, with p-values of 0.00115 and 0.00325, respectively. Filter status had no significant association. Fig. 9 shows the effect of temperature on the T90 values, Fig. 10 demonstrates the variance associated with sampling location, and Fig. 11 compares the two filter statuses.

3.4. Model selection tool

The final classification trees obtained for each microorganism based on the CART analysis are shown in Figs. 12–14. For MS-2,

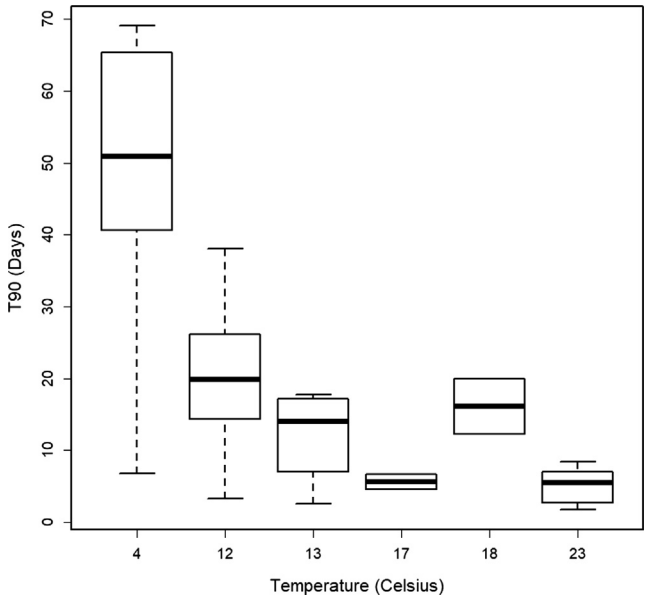


Fig. 3. Variance in the MS-2 T90 values with respect to temperature.

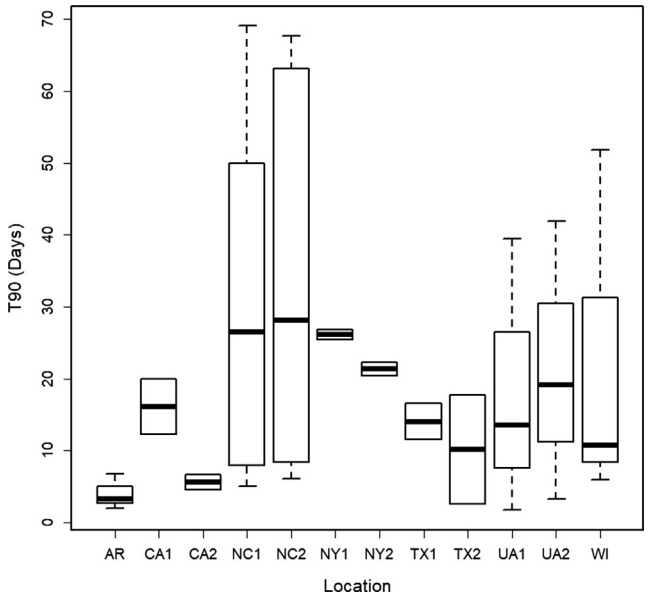


Fig. 4. Variance in the MS-2 T90 values with respect to sampling location.

the relevant predictor variables were temperature, nitrate, and turbidity. The tree for MS-2 (Fig. 12) has six terminal nodes, and only nine of the 36 experiments were misclassified indicating good performance. The trees should be read top down to identify an appropriate model for any given study of the decay of MS-2 in groundwater given the predictor variable value. For example, a decay dataset with turbidity greater than 0.625 NTU and nitrate less than 5.5 mg/L should be fit with the epd model.

The final classification tree for poliovirus 1 (Fig. 13) has four terminal nodes where temperature, pH, and calcium hardness were the only relevant predictor variables; therefore, four possible models of the seventeen tested (jm2, jm1, dep or epd) are likely to fit the decay patterns observed at the sampled water quality conditions. The poliovirus 1 tree only misclassified four out of 20 experiments. For echovirus 1 (Fig. 14), the final tree has four terminal nodes with temperature and pH as relevant predictors. For echovirus 1, 18 of 20 experiments were correctly classified with three possible models (jm2, ep, or jm1) for all datasets.

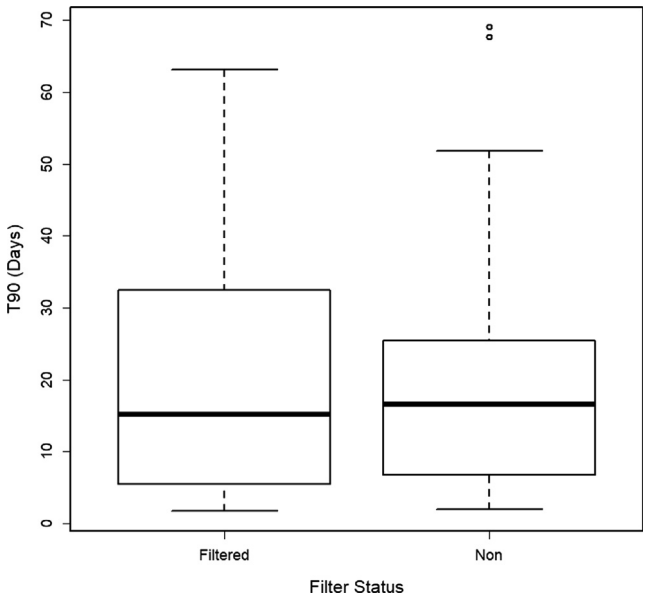


Fig. 5. Variance in the MS-2 T90 values with respect to filter status.

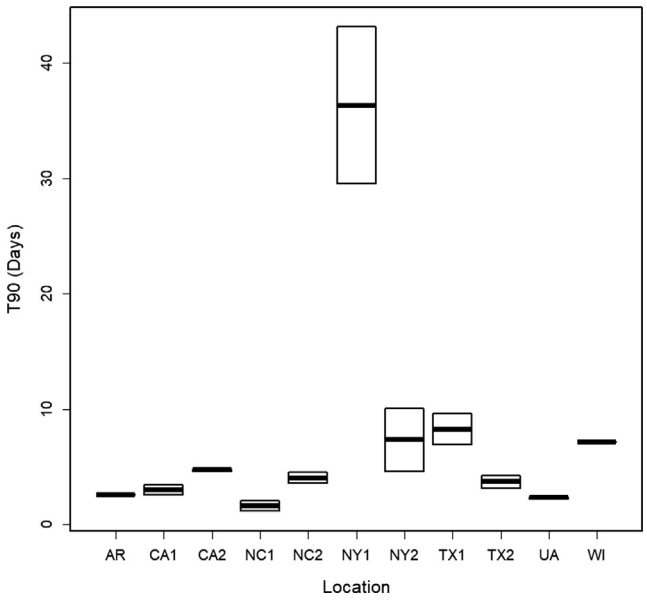


Fig. 7. Variance in the poliovirus 1 T90 values with respect to sampling location.

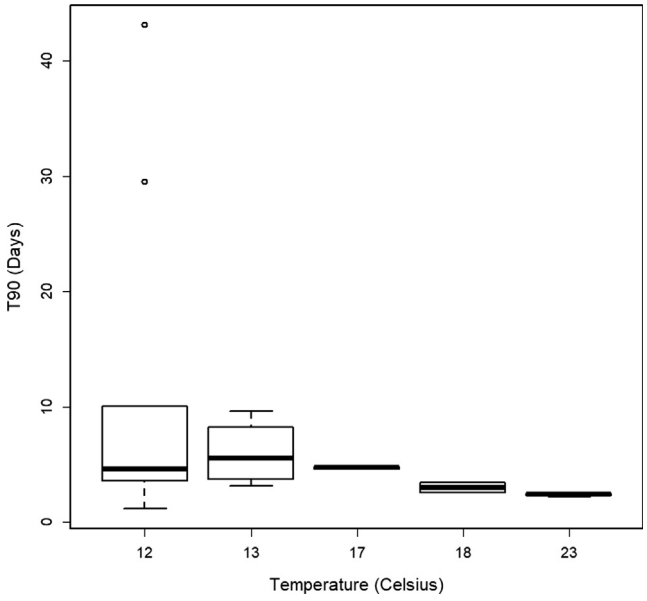


Fig. 6. Variance in the poliovirus 1 T90 values with respect to temperature.

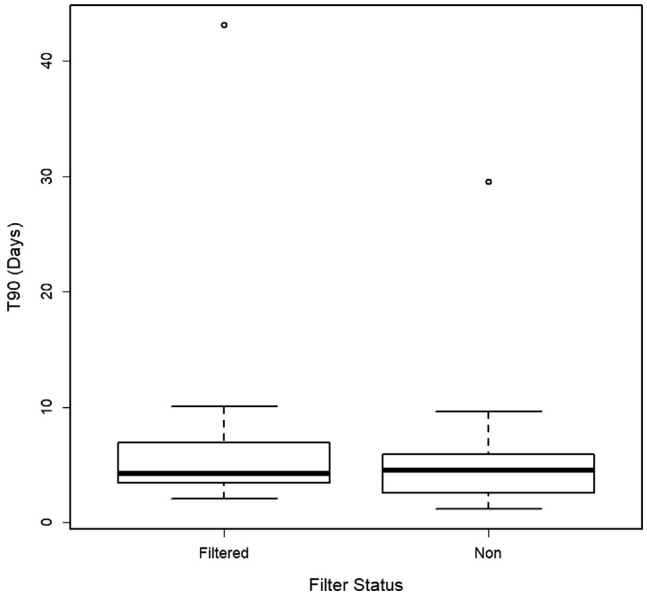


Fig. 8. Variance in the poliovirus 1 T90 values with respect to filter status.

4. Discussion

This study analyzed 76 persistence datasets consisting of groundwater samples taken from across the U.S. for three microorganisms – MS-2, poliovirus 1 and echovirus 1. The goal of the study was to determine which models best described the decay patterns observed. However, it is important to note that this study did not aim to prove that the most widely used decay rate models were inadequate, but rather to determine if other available models could reduce uncertainties associated with model structure and parameter estimation. Furthermore, the results of this analysis were evaluated within a decision framework to establish a systematic approach for the selection of a model structure given specific water quality characteristics of the data known to impact the inactivation of pathogens.

4.1. Model structure and fit

Across all 76 datasets, the Juneja and Marks 1 (jm1) (Juneja et al., 2001), Juneja and Marks 2 (jm2) (Huang et al., 2006), and Exponential Damped (epd) (Cavalli-Sforza et al., 1983) were selected as the best fitting models most often. The Juneja and Marks models are mechanistically motivated models, representing biological behavior (i.e. stages of inactivation of a cell) behind the fits, versus the empirical models that are based on the observed relationship with the experimental data. This analysis suggests that these three models may better predict virus persistence in groundwater when compared to the more commonly used exponential and Weibull models.

Previous persistence modeling also frequently identified the two Juneja and Marks models as the best fit for the persistence of bacteria, bacteriophages, viruses and bacteroidales in different

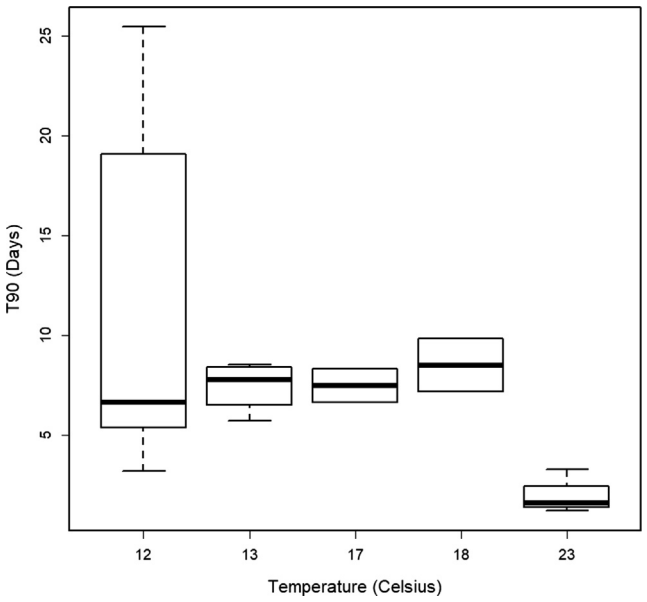


Fig. 9. Variance in the echovirus 1 T90 values with respect to temperature.

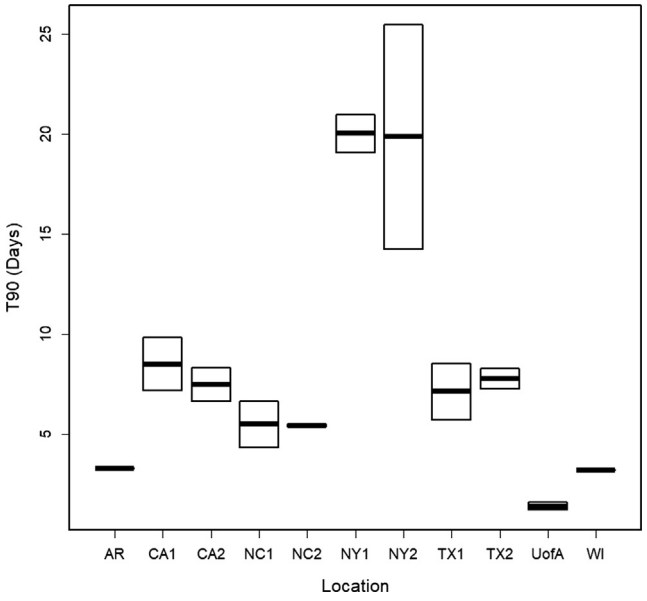


Fig. 10. Variance in the echovirus 1 T90 values with respect to sampling location.

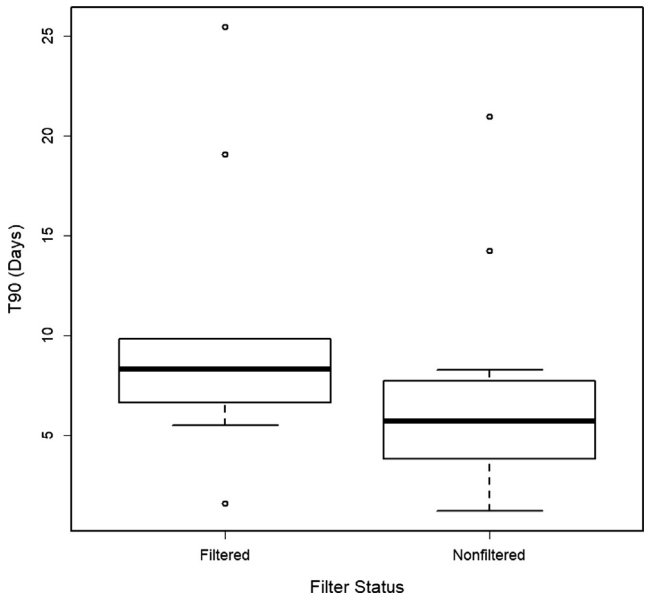


Fig. 11. Variance in the echovirus 1 T90 values with respect to filter status.

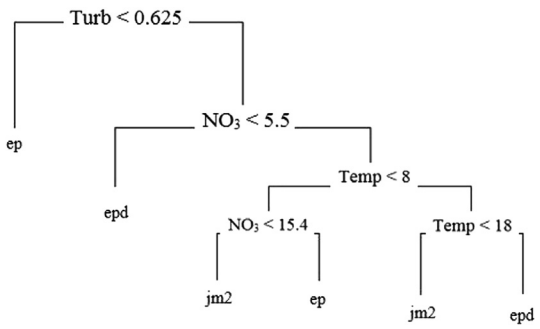


Fig. 12. Final classification tree for MS-2.

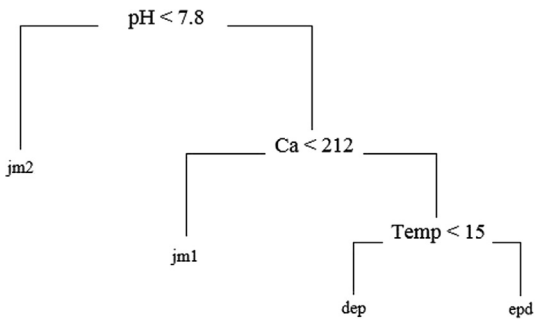


Fig. 13. Final classification tree for poliovirus 1.

water matrices (Mitchell and Akram, 2017). However, Brooks et al. (2015) fit the exponential, biphasic exponential, Juneja and Marks 1, Juneja and Marks 2, and Gompertz-3 models to DNA targets for bacterial indicators and the exponential and biphasic exponential model fit 73% of the experimental data (Brooks et al., 2015). This suggests that although there may be some consistency with model type for the persistence of viruses and virus surrogates, there may be some discrepancy between model types for the persistence of bacteria and the DNA targets for bacterial indicators. Of note, the exponential model was the best fitting model more frequently for MS-2 (30.6%) than for poliovirus 1 (5%) or echovirus 1 (25%) herein. It is possible that the persistence of indicators may be captured by first-order decay rates more frequently than their pathogenic counterparts.

It is important to note that although the Juneja and Marks models and the Exponential Damped model were most frequently identified as the best fitting model for the data analyzed in this study, for 21% of the 76 datasets none of the models fit to the data had a statistically good fit. With the inherent variability and nature of microbial data, perhaps a significance level of 0.05 was too stringent. A different significance level for the likelihood ratio tests would have led to a greater number of datasets with models that had a statistically good fit.

4.2. T90 analysis

When comparing the T90 values from this study and those previously completed, it is important to discuss the relevance of the

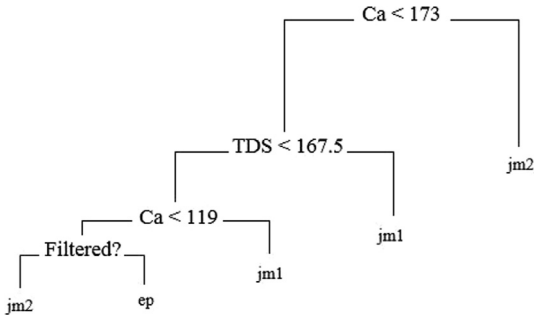


Fig. 14. Final classification tree for echovirus 1.

T90 value. A T90 value indicates the amount of time for a one log reduction in concentration, which means it rarely accounts for the tailing off effect that occurs over time. To treat water contaminated with viruses, the EPA requires systems to achieve a 4-log reduction (U.S. Environmental Protection Agency, 2016). If this treatment goal is based on the time required for the first log reduction (T90) which occurs in the steepest portion of the curve, the total time required for a 4-log reduction may be underestimated, resulting in inadequate treatment.

There was some discrepancy between MS-2 and the two viruses, echovirus and poliovirus, when comparing the model averaged T90s to the original log-linear fit T90s. In this study, the model averaged T90s for MS-2 were lower than the log-linear produced T90s for only half of the datasets. For the two viruses, the model averaged T90s were lower for 85% of the datasets. Thus, the log-linear fits made more conservative decay estimates for the two viruses compared to the best fitting models in this study. For 50% of the MS-2 datasets, however, the log-linear fits estimated faster decay than the models fit in this study. Additionally, for 35% of the sampled locations, MS-2 had a lower model averaged T90 than at least one of the two viruses. If treatment decisions were based on the MS-2 persistence in these locations, the actual virus concentration in the groundwater could be underestimated and pose a risk to the public.

4.3. CART analysis

In general, the CART results demonstrated the same five models (jm1, jm2, ep, epd, and dep) being selected based on variations of different variables for MS-2, poliovirus and echovirus. Similarly to Yates et al. (1984), the ANOVA analysis conducted in this study identified temperature as the only common significant predictor of persistence for all three targets: MS-2, poliovirus, and echovirus. However, the CART analysis identified other water quality characteristics, such as calcium hardness, pH, and nitrate, as relevant predictors for the selection of the model that best describes said persistence. Additionally, the results indicated that explanatory or predictor variables such as ammonia and total and magnesium hardness were not relevant for determining the most suitable type of inactivation model for the viruses or MS-2. The CART results reflect the more subtle influences of environmental and sampling conditions on decay rates and provide insights for future researchers as they suggest that resources may be best spent on certain water quality measurements over others in relation to decay rates.

4.4. Implications

The findings of this study provide meaningful consideration of more systematic approaches to model structure selection in the evaluation of persistence data. Given the implications in the broader context of risk assessment where erroneous predictions of concentrations over time can result in overestimating or underestimating risks, these findings imply that it may be necessary to shift the thinking about which models are most appropriate for new datasets.

5. Conclusion

The classification trees in this study provide a methodology for future research. Based on the water quality and laboratory test conditions, it is possible to follow the branches of the classification tree for the pathogen of concern and determine which model would be the best selection. In this study, we presented a procedure that can be replicated for more pathogens of concern to determine more appropriate mathematical models to fit persistence data under varying conditions. Most studies do not consider model uncertainty, which can contribute to erroneous conclusions and residual risks. The classification trees developed in this study should be applied to future persistence data to evaluate their range of applicability and to further investigate the models that provided the best fit most frequently in this study, Juneja and Marks 1, Juneja and Marks 2 and the exponential damped.

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.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abbaszadegan, M., Alum, A., 2016. Waterborne enteric viruses: diversity, distribution, and detection. Manual of Environmental Microbiology,. ASM Press, Washington, D.C.

Abraham, G., Debray, E., Candau, Y., Piar, G., 1990. Mathematical model of thermal destruction of *Bacillus stearothermophilus* spores. Appl. Environ. Microbiol. 56 (10), 3073–3080.

Aragao, G.M.F., Corradini, M.G., Normand, M.D., Peleg, M., 2007. Evaluation of the Weibull and log normal distribution functions as survival models of *Escherichia coli* under isothermal and non isothermal conditions. Int. J. Food Microbiol. 119 (3), 243–257. https://doi.org/10.1016/j.ijfoodmicro.2007.08.004.

Benedict, K.M., Reses, H., Vigar, M., et al., 2017. Surveillance for waterborne disease outbreaks associated with drinking water-United States, 2013–2014. Morb. Mortal. Wkly Rep. 66 (44), 1216–1221.

Bozkurt, H., D'Souza, D.H., Davidson, P.M., 2014a. A comparison of the thermal inactivation kinetics of human norovirus surrogates and hepatitis A virus in buffered cell culture medium. Food Microbiol. 42, 212–217. https://doi.org/10.1016/j.fm.2014.04.002.

Bozkurt, H., Leiser, S., Davidson, P.M., D'Souza, D.H., 2014. Thermal inactivation kinetic modeling of human norovirus surrogates in blue mussel (*Mytilus edulis*) homogenate. Int. J. Food Microbiol. 172, 130–136. https://doi.org/10.1016/j.ijfoodmicro.2013.11.026.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall/CRC, Boca Raton, FL.

Brooks, L.E., Field, K.G., 2016. Bayesian meta-analysis to synthesize decay rate constant estimates for common fecal indicator bacteria. Water Res. 104, 262–271. https://doi.org/10.1016/j.watres.2016.08.005.

Brooks, Y., Aslan, A., Tamrakar, S., Murali, B., Mitchell, J., Rose, J.B., 2015. Analysis of the persistence of enteric markers in sewage polluted water on a solid matrix and in liquid suspension. Water Res. 76, 201–212. https://doi.org/10.1016/j.watres.2015.02.039.

Cavalli-Sforza, L.T., Menozzi, P., Strata, A., 1983. A model and program for study of a tolerance curve: application to lactose absorption tests. Int. J. Biomed. Comput. 14 (1), 31–41. https://doi.org/10.1016/0020-7101(83)90084-3.

Chick, H., 1908. An investigation of the laws of disinfection. J. Hygiene 8 (1), 92–158.

Dziak, J.J., Coffman, D.L., Lanza, S.T., Runze, L., 2012. Sensitivity and Specificity of Information Criteria: Technical Report Series #12-119. The Methodology Center, pp. 1–30. Technical Report (814), Retrieved from http://methodology.psu.edu/media/techreports/12-119.pdf.

Enger, K.S., Mitchell, J., Murali, B., Birdsell, D.N., Keim, P., Gurian, P.L., Wagner, D.M., 2018. Evaluating the long-term persistence of Bacillus spores on common

surfaces. Microb. Biotechnol. 11 (6), 1048–1059. https://doi.org/10.1111/1751-7915.13267.
Gerba, C.P., Rose, J.B. (1990). Viruses in Source and Drinking Water. In: McFeters G. A. (Eds.) Drinking Water Microbiology. Brock/Springer Series in Contemporary Bioscience. Springer, New York, NY.
Gil, M.M., Miller, F.A., Brandão, T.R.S., Silva, C.L.M., 2011. On the use of the gompertz model to predict microbial thermal inactivation under isothermal and non-isothermal conditions. Food Eng. Rev. 3 (1), 17–25. https://doi.org/10.1007/s12393-010-9032-2.
Gronewold, A.D., Myers, L., Swall, J.L., Noble, R.T., 2011. Addressing uncertainty in fecal indicator bacteria dark inactivation rates. Water Res. 45 (2), 652–664. https://doi.org/10.1016/j.watres.2010.08.029.
Haas, C., Rose, J., Gerba, C., 2014. Quantitative Microbial Risk Assessment. John Wiley & Sons Inc, Hoboken, New Jersey.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction,. Springer, New York. https://doi.org/10.1007/b94608.
Huang, L., Thippareddi, H., Juneja, V., 2006. Predictive model for growth of Clostridium perfringens in cooked cured pork. Int. J. Food Microbiol. 110 (1), 85–92. https://doi.org/10.1016/j.ijfoodmicro.2006.01.038.
Jodrá, P., 2009. A closed-form expression for the quantile function of the Gompertz-Makeham distribution. Math. Comput. Simul 79 (10), 3069–3075. https://doi.org/10.1016/j.matcom.2009.02.002.
Juneja, V.K., Eblen, B.S., Marks, H.M., 2001. Modeling non-linear survival curves to calculate thermal inactivation of Salmonella in poultry of different fat levels. Int. J. Food Microbiol. 70 (1–2), 37–51. https://doi.org/10.1016/S0168-1605(01)00518-9.
Kamau, D.N., Doores, S., Pruitt, K.M., 1990. Enhanced thermal destruction of *Listeria monocytogenes* and *Staphylococcus aureus* by the lactoperoxidase system. Appl. Environ. Microbiol. 56 (9), 2711–2716. https://doi.org/10.4315/0362-028X-53.12.1010.
Kingsley, D.H., Chen, H., Meade, G.K., 2018. Persistence of MS-2 bacteriophage within eastern oysters. Food Environ. Virol. 10 (1), 83–88. https://doi.org/10.1007/s12560-017-9315-3.
Kocwa-Haluch, R., 2001. Waterborne enteroviruses as a hazard for human health. Polish J. Environ. Studies 10 (6), 485–487.
Kott, Y., Roze, N., Sperber, S., Betzer, N., 1974. Bacteriophages as viral pollution indicators. Water Res. 8 (3), 165–171. https://doi.org/10.1016/0043-1354(74)90039-6.
Kreutzwiser, R., de Loe, R., Imgrund, K., Conboy, M.J., Simpson, H., Plummer, R., 2011. Understanding stewardship behaviour: factors facilitating and constraining private water well stewardship. J. Environ. Manage. 92, 1104–1114.
Leclerc, H., Edberg, S., Pierzo, V., Delattre, J.M., 2000. Bacteriophages as indicators of enteric viruses and public health risk in groundwaters. J. Appl. Microbiol. 88 (1), 5–21. https://doi.org/10.1046/j.1365-2672.2000.00949.x.
Liang, L., Goh, S.G., Gin, K.Y.H., 2017. Decay kinetics of microbial source tracking (MST) markers and human adenovirus under the effects of sunlight and salinity. Sci. Total Environ. 574, 165–175. https://doi.org/10.1016/j.scitotenv.2016.09.031.
Macler, B.A., Merkle, J.C., 2000. Current knowledge on groundwater microbial pathogens and their control. Hydrogeol. J. 8 (1), 29–40. https://doi.org/10.1007/PL00010972.
Mitchell, J. and Akram, S. 2017. Pathogen Specific Persistence Modeling Data. In: J.B. Rose and B. Jiménez-Cisneros, (Eds.) Global Water Pathogens Project, http://www.waterpathogens.org (M. Yates (Eds.) Part 4 Management of Risk from Excreta and Wastewater) http://www.waterpathogens.org/book/pathogen-specific-persistence-modeling-data Michigan State University, E. Lansing, MI, UNESCO. https://doi.org/10.14321/waterpathogens.53.
Muggeo, V.M.R., 2003. Estimating regression models with unknown break-points. Stat. Med. 22 (19), 3055–3071. https://doi.org/10.1002/sim.1545.
Murphy, H.M., Prioleau, M.D., Borchardt, M.A., Hynds, P.D., 2017. Revue: Preuves épidémiologiques de la contribution des eaux souterraines aux maladies entériques au niveau mondial entre 1948 et 2015. Hydrogeol. J. 25 (4), 981–1001. https://doi.org/10.1007/s10040-017-1543-y.
Nkwe, K., Ateba, C., Sithebe, N., Bezuidenhout, C., 2015. Enumeration of somatic and F-RNA phages as an indicator of fecal contamination in potable water from rural areas of the North West Province. Pathogens 4 (4), 503–512. https://doi.org/10.3390/pathogens4030503.
Peleg, M., 1995. A model of microbial survival after exposure to pulsed electric fields. J. Sci. Food Agric. 67 (1), 93–99. https://doi.org/10.1002/jsfa.2740670115.
Peleg, M., Cole, M.B., 1998. Reinterpretation of microbial survival curves. Crit. Rev. Food Sci. Nutr. 38 (5), 353–380. https://doi.org/10.1080/10408699891274246.
Peleg, M., 2003. Calculation of the non-isothermal inactivation patterns of microbes having sigmoidal isothermal semi-logarithmic survival curves. Crit. Rev. Food Sci. Nutr. 43 (6), 645–658. https://doi.org/10.1080/10408690390251156.
Peleg, M., 2006. Advanced Quantitative Microbiology for Foods and Biosystems: Models for Predicting Growth and Inactivation. Taylor & Francis, Boca Raton.
Ratjar, B., Majek, M., Polanski, L., Polz-Dacewicz, M., 2008. Enteroviruses in water environment-a potential threat to public health. Ann. Agric. Environ. Med. 15, 199–203.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ripley, B. D. 1996. Pattern {R}ecognition and {N}eural {N}etworks, (1995), 1–13. https://doi.org/10.1017/CBO9780511812651.
Ripley, B., 2018. tree: Classification and Regression Trees. R package version 1.0-38 https://CRAN.R-project.org/package=tree.
U.S. Environmental Protection Agency, 2006. National Primary Drinking Water Regulations: Ground Water Rule, Final Rule. Federal Register 71 (216), 65574–65660.
U.S. Environmental Protection Agency, 2018. Private Drinking Water Wells. United States Environmental Protection Agency. Retrieved from https://www.epa.gov/privatewells.
U.S. Environmental Protection Agency, 2016. Drinking Water Contaminant List 4-Final. Federal Register 81 (222), 81099–81114.
van Gerwen, S.J., Zwietering, M.H., 1998. Growth and inactivation models to be used in quantitative risk assessments. J. Food Prot. 61 (11), 1541–1549. https://doi.org/10.4315/0362-028X-61.11.1541.
Wallender, E.K., Ailes, E.C., Yoder, J.S., Roberts, V.A., Brunkard, J.M., 2014. Contributing factors to disease outbreaks with untreated groundwater. Groundwater 52 (6), 886–897.
Walters, S.P., Yamahara, K.M., Boehm, A.B., 2009. Persistence of nucleic acid markers of health-relevant organisms in seawater microcosms: implications for their use in assessing risk in recreational waters. Water Res. 43 (19), 4929–4939. https://doi.org/10.1016/j.watres.2009.05.047.
World Health Organization, 2017. Guidelines for drinking-water quality: fourth edition incorporating first addendum, 4th ed + 1st add. World Health Organization. https://appswho.int/iris/handle/10665/254637.
Wu, J.W., Hung, W.L., Tsai, C.H., 2004. Estimation of parameters of the Gompertz distribution using the least squares method. Appl. Math. Comput. 158 (1), 133–147. https://doi.org/10.1016/j.amc.2003.08.086.
Xiong, R., Xie, G., Edmondson, A.E., Sheard, M.A., 1999. A mathematical model for bacterial inactivation. Int. J. Food Microbiol. 46 (1), 45–55. https://doi.org/10.1016/S0168-1605(98)00172-X.
Yates, M. V. 1984. Virus persistence in ground water, PhD Dissertation. The University of Arizona. Retrieved from https://repository.arizona.edu/handle/10150/191091.
Yates, M.V., Gerba, C.P., Kelley, L.M., 1985. Virus persistence in groundwater. Appl. Environ. Microbiol. 49 (4), 778–781.
Yates, M.V., Stetzenbach, L.D., Gerba, C.P., Sinclair, N.A., 1990. The effect of indigenous bacteria on virus survival in ground water. J. Environ. Sci. Health Part A 25 (1), 81–100. https://doi.org/10.1080/10934529009375541.