

East Tennessee State University

## Digital Commons @ East Tennessee State University

---

ETSU Faculty Works

Faculty Works

---

9-13-2018

### Maxent Estimation of Aquatic Escherichia Coli Stream Impairment

Dennis Gilfillan

*East Tennessee State University*

Timothy A. Joyner

*East Tennessee State University, joynert@etsu.edu*

Phillip Scheuerman

*East Tennessee State University, philsche@etsu.edu*

Follow this and additional works at: <https://dc.etsu.edu/etsu-works>



Part of the [Environmental Microbiology and Microbial Ecology Commons](#), and the [Environmental Public Health Commons](#)

---

#### Citation Information

Gilfillan, Dennis; Joyner, Timothy A.; and Scheuerman, Phillip. 2018. Maxent Estimation of Aquatic Escherichia Coli Stream Impairment. *PeerJ*. Vol.6 5610. <https://doi.org/10.7717/peerj.5610> ISSN: 2167-8359

This Article is brought to you for free and open access by the Faculty Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in ETSU Faculty Works by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact [digilib@etsu.edu](mailto:digilib@etsu.edu).

---

# Maxent Estimation of Aquatic Escherichia Coli Stream Impairment

## Copyright Statement

2018 Gilfillan et al. This document was originally published in *PeerJ*.



# Maxent estimation of aquatic *Escherichia coli* stream impairment

Dennis Gilfillan<sup>1</sup>, Timothy A. Joyner<sup>2</sup> and Phillip Scheuerman<sup>1</sup>

<sup>1</sup>Department of Environmental Health Sciences, East Tennessee State University, Johnson City, TN, United States of America

<sup>2</sup>Department of Geosciences, East Tennessee State University, Johnson City, TN, United States of America

## ABSTRACT

**Background.** The leading cause of surface water impairment in United States' rivers and streams is pathogen contamination. Although use of fecal indicators has reduced human health risk, current approaches to identify and reduce exposure can be improved. One important knowledge gap within exposure assessment is characterization of complex fate and transport processes of fecal pollution. Novel modeling processes can inform watershed decision-making to improve exposure assessment.

**Methods.** We used the ecological model, Maxent, and the fecal indicator bacterium *Escherichia coli* to identify environmental factors associated with surface water impairment. Samples were collected August, November, February, and May for 8 years on Sinking Creek in Northeast Tennessee and analyzed for 10 water quality parameters and *E. coli* concentrations. Univariate and multivariate models estimated probability of impairment given the water quality parameters. Model performance was assessed using area under the receiving operating characteristic (AUC) and prediction accuracy, defined as the model's ability to predict both true positives (impairment) and true negatives (compliance). Univariate models generated action values, or environmental thresholds, to indicate potential *E. coli* impairment based on a single parameter. Multivariate models predicted probability of impairment given a suite of environmental variables, and jack-knife sensitivity analysis removed unresponsive variables to elicit a set of the most responsive parameters.

**Results.** Water temperature univariate models performed best as indicated by AUC, but alkalinity models were the most accurate at correctly classifying impairment. Sensitivity analysis revealed that models were most sensitive to removal of specific conductance. Other sensitive variables included water temperature, dissolved oxygen, discharge, and NO<sub>3</sub>. The removal of dissolved oxygen improved model performance based on testing AUC, justifying development of two optimized multivariate models; a 5-variable model including all sensitive parameters, and a 4-variable model that excluded dissolved oxygen.

**Discussion.** Results suggest that *E. coli* impairment in Sinking Creek is influenced by seasonality and agricultural run-off, stressing the need for multi-month sampling along a stream continuum. Although discharge was not predictive of *E. coli* impairment alone, its interactive effect stresses the importance of both flow dependent and independent processes associated with *E. coli* impairment. This research also highlights the interactions between nutrient and fecal pollution, a key consideration for watersheds with multiple synergistic impairments. Although one indicator cannot mimic the

Submitted 8 March 2018  
Accepted 20 August 2018  
Published 13 September 2018

Corresponding author  
Dennis Gilfillan, gilfillan@etsu.edu

Academic editor  
Timothy Scheibe

Additional Information and  
Declarations can be found on  
page 17

DOI 10.7717/peerj.5610

© Copyright  
2018 Gilfillan et al.

Distributed under  
Creative Commons CC-BY 4.0

OPEN ACCESS

plethora of existing pathogens in water, incorporating modeling can fine tune an indicator's utility, providing information concerning fate, transport, and source of fecal pollution while prioritizing resources and increasing confidence in decision making.

**Subjects** Microbiology, Public Health, Computational Science, Data Mining and Machine Learning, Spatial and Geographic Information Science

**Keywords** Fecal indicators, Environmental microbiology, Surface Water quality, Statistical modeling

## INTRODUCTION

Rapid urbanization of rural areas causes deterioration in water quality, rendering many water bodies unfit for their domestic and recreational use. An assortment of contaminants is introduced into aquatic systems, but pathogens represent the major cause of stream impairment in the United States (*United States Environmental Protection Agency, 2017*). Pathogens are difficult to measure directly because of their sporadic distribution, costly identification, and potential health risks to laboratory workers (*Field & Samadpour, 2007*). Most pathogens in aquatic systems stem from human and animal fecal wastes, including direct deposition of feces in water (*Vidon, Campbell & Gray, 2008*), run-off from land with fecal deposits (*Tyrrel & Quinton, 2003; Jamieson et al., 2004*), and sanitary sewer malfunctions (*Ferguson et al., 2003; McLellan & Eren, 2014*). To address the difficulties in monitoring specific pathogens, fecal indicator organisms (FIOs) are commonly used to assess the presence of fecal pathogens.

An effective fecal indicator is associated with the presence of specific pathogens, with a straightforward method for enumeration that correlates with magnitude and age of fecal pollution (*Savichtcheva & Okabe, 2006; Maier, Pepper & Gerba, 2009*). The use of FIOs such as fecal coliform bacteria and *Escherichia coli* are traditionally used for determining surface water pathogen impairment (*Yates, 2007; US Environmental Protection Agency, 2012*). Although these indicators assist in alerting populations when exposure to pathogens is likely, the current approach is limited by using a single indicator such as *E. coli* for a designated use (*Wade et al., 2003; Savichtcheva & Okabe, 2006; Field & Samadpour, 2007*). The cosmopolitan nature of *E. coli* in warm-blooded animals makes them impractical for source identification (*Field & Samadpour, 2007; Yates, 2007; McLellan & Eren, 2014; Blount, 2015*). The ability of *E. coli* to survive in soils (*Lasalde et al., 2005; Ishii et al., 2006*), algae (*Byappanahalli et al., 2003*), and sediments (*LaLiberte & Grimes, 1982; Alm & Burke, 2003; Drummond et al., 2015*) provide a reservoir for continued persistence and potential to naturalize (*Winfield & Groisman, 2003; Lasalde et al., 2005; Luo et al., 2011*). These characteristics and deficiencies emphasize the difficulty of single standard FIO monitoring for impairment, stressing the need for additional methods to evaluate source and mechanisms of FIO impairment.

In addition to the above issues, appropriately characterizing FIO impairment for regulation and decision-making is difficult due to complex fate and transport processes (*Benham et al., 2006; De Brauwere, Ouattara & Servais, 2014; Drummond et al., 2015*).

These complex fate and transport processes include transport through run-off and storm water (Kistemann *et al.*, 2002; Lipp *et al.*, 2001; McKergow & Davies-Colley, 2010), remobilization from sediments and hyporheic exchange (Drummond, 2015; Dwivedi, Mohanty & Lesikar, 2016), particle attachment (Characklis *et al.*, 2005), and UV light exposure (Sinton *et al.*, 2002). Additionally, ecological processes control FIO fate and transport through variable survival patterns of indicators and pathogens (Anderson *et al.*, 2005; Stott *et al.*, 2011), availability of nutrients and organic matter (Surbeck, Jiang & Grant, 2010; Perkins *et al.*, 2016), and predation (McCambridge & McMeekin, 1980). Appropriately characterizing the physics and ecology driving fate and transport can better inform management decisions for total maximum daily load (TMDL) development, reduction of pollution, and allocation of resources.

Modeling provides flexible approaches to infer sources and processes associated with FIOs and other pathogens, overcoming some of the issues of the single indicator paradigm. Various statistical and machine learning models have been used to approach such problems of incorporating age of fecal pollution for source tracking or detection of viruses (Brion, Neelakantan & Lingireddy, 2002; Black, Brion & Freitas, 2007); identifying land use, environmental, and water quality parameters associated with FIOs and pathogens (Brion & Lingireddy, 1999; Viau *et al.*, 2011; Wilkes *et al.*, 2011; Gonzalez *et al.*, 2012; Gonzalez & Noble, 2014; Hall *et al.*, 2014; Herrig *et al.*, 2015; Lušić *et al.*, 2017); determining factors influencing particle attachment and virulence (Piorkowski *et al.*, 2013); and optimizing microbial source tracking (Belanche-Muñoz & Blanch, 2008; Ballestè *et al.*, 2010; Smith, Sterba-Boatwright & Mott, 2010; Molina *et al.*, 2014). Some other applications of modeling include using turbidity or rainfall to predict *E. coli* concentrations at unmonitored sites (Money, Carter & Serre, 2009; Coulliete *et al.*, 2009), estimating *E. coli* loads using physical, chemical, and biological factors within a neural network (Dwivedi, Mohanty & Lesikar, 2013), and hyporheic-groundwater interactions associated with transport of *E. coli* within sediments porewater (Dwivedi, Mohanty & Lesikar, 2016). Modeling can inform decision-makers concerning what drives impairment, addressing some of the shortcomings of a single indicator approach.

Maxent, a commonly used ecological niche model (Phillips, Dudík & Schapire, 2004; Phillips & Dudík, 2008), identified environmental variables associated with probability of *E. coli* stream impairment, making inferences concerning source and mechanisms driving fecal pollution. Although modeling *E. coli* using a machine learning model such as Maxent is not a novel approach, e.g., Dwivedi, Mohanty & Lesikar, 2013, this study is unique in the following ways: it focuses on how the water quality is associated with *E. coli* impairment in lower order streams, uses nonparametric bootstrapping as a probabilistic assessment of model performance based on the area under the curve (AUC) of the receiving operator characteristic (ROC), and uses loss of information as an indicator of sensitive variables. Ecological niche models have been utilized for species distribution (Lozier, Aniello & Hickerson, 2009), conservation of rare species (Guisan *et al.*, 2006), invasive species (Thuiller *et al.*, 2005), and disease vector epidemiology studies (Boeckmann & Joyner, 2014), but this is a new application of Maxent to microbial water quality. Additionally, developing models

in lower order streams has not been previously reported; this is important because water from low order streams is used for domestic water supply and recreation in many areas of the United States.

The motivation for using Maxent to predict *E. coli* impairment is to investigate how environment, i.e water quality parameters, shapes the niche of *E. coli* impairment based on a decision boundary; in this case, a water quality standard. A probabilistic procedure for univariate and multivariate model development is presented using nonparametric bootstrapping cross-validation. Univariate models generated action values, or environmental thresholds of impairment, to indicate potential *E. coli* impairment based on a single parameter. Multivariate models predicted probability of impairment given a suite of environmental variables, and jack-knife sensitivity analysis removed unresponsive variables in multivariate models to elicit a set of the most responsive water quality parameters. Using Maxent to model how water quality influences *E. coli* impairment aids in inferring source and mechanisms of fecal pollution. This approach allows for estimation of both linear and non-linear effects of water quality, demonstrates a probabilistic method for variable selection, and reframes the question from “How much *E. coli* in our watershed?” to “what factors separate *E. coli* impairment from compliance?” which is useful when evaluating watershed decisions.

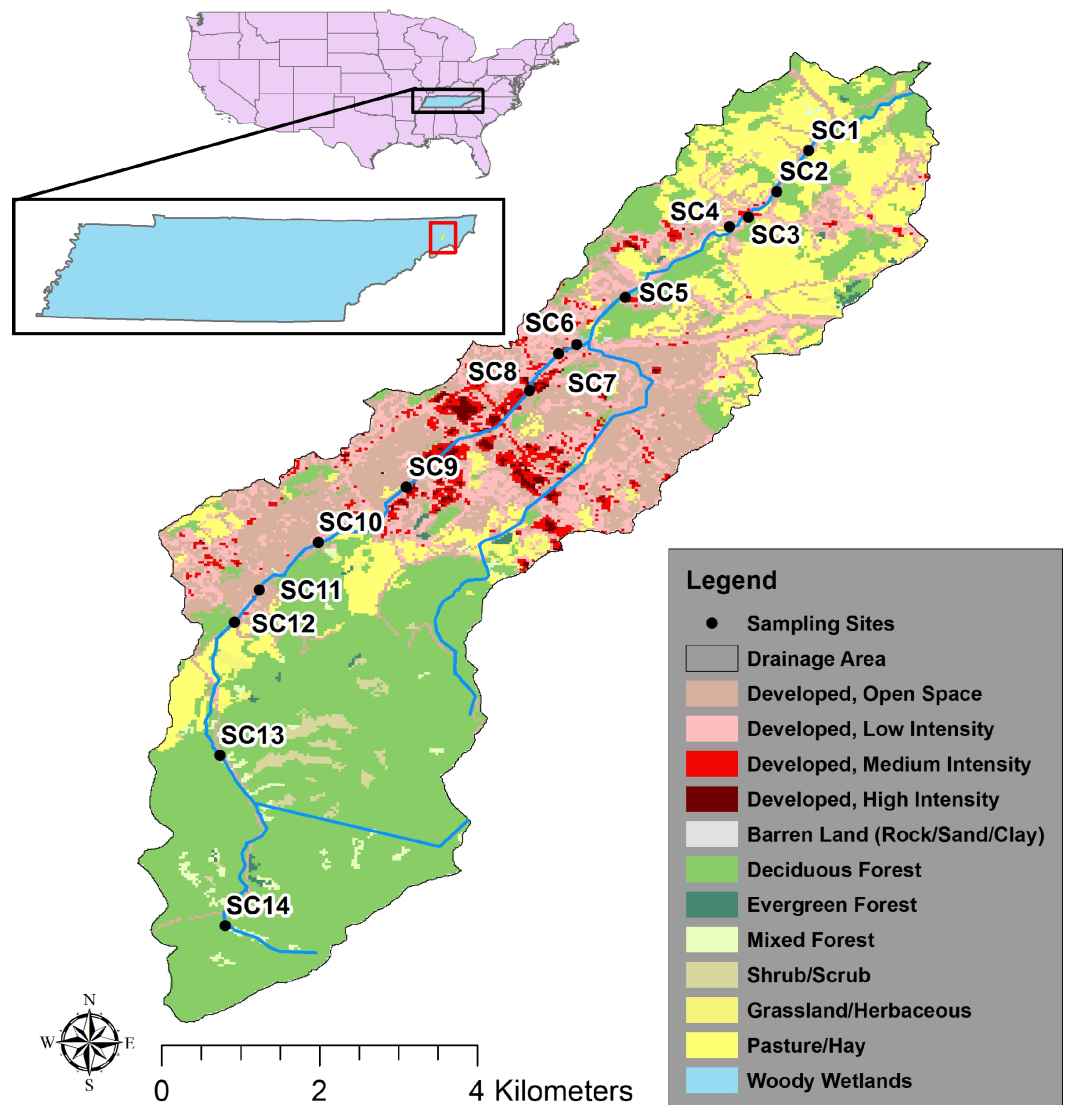
## METHODS

### Sampling sites and data collection

Sinking Creek is a 1st to 3rd Strahler order mixed-use stream that is noncompliant for state of Tennessee standards for fecal coliform and *E. coli* (*Tennessee Department of Environmental and Conservation, 2006*). Starting in August 2004, samples were collected by hand in August, November, February, and May of each year until August 2011 as a long-term monitoring plan at 14 sites in Sinking Creek, and samples were analyzed for 10 water quality parameters and populations of *E. coli* (*Fig. 1*).

Specific conductance (conductivity) and water temperature were measured using an Orion 115A+ conductivity meter (Thermo Fisher Scientific, Waltham, MA). The pH was measured using an EL2 portable pH meter (Mettler Toledo, Columbus, OH). Dissolved oxygen was collected using a YSI Model 55 dissolved oxygen meter (YSI Inc., Yellow Springs, OH). Samples for nitrates ( $\text{NO}_3$ ), phosphates ( $\text{PO}_4$ ), biochemical oxygen demand (BOD), alkalinity, and hardness were collected in clean 2 L polyethylene bottles and stored at 4 °C until laboratory analysis. A flow meter (Global Water, FP101) was placed in the center of the channel to measure stream velocity. Stream width was calculated where the stream velocity was measured, and depth was averaged over three points across the stream width. Velocity was multiplied by stream width and average depth to estimate discharge.

$\text{NO}_3$  and  $\text{PO}_4$  analyses were performed in triplicate using colorimetric HACH™ methods (Hach, Loveland, CO) and reagents.  $\text{NO}_3$  and  $\text{PO}_4$  analyses were conducted by adding 10 mL of water to a vial containing NitraVer5 or PhosVer3 for the respective analyses. Vials were shaken to dissolve the reagent and samples were analyzed with a DR890 colorimeter (Hach, Loveland, CO) (*Hach Company, 2013*). Triplicate sample for alkalinity



**Figure 1** Map of sampling sites and watershed of the study area, Sinking Creek. The inset map shows the United States and the state of Tennessee, and the location of Sinking Creek. Samples were taken from August 2004 to August 2011 during the months of August, November, February, and May. The outline represents the watershed boundary of Sinking Creek, and 2006 NLCD has been clipped to the watershed (Fry *et al.*, 2011). Stream flows from its headwaters at SC14 downstream to SC1.

Full-size [DOI: 10.7717/peerj.5610/fig-1](https://doi.org/10.7717/peerj.5610/fig-1)

and hardness were determined using 100 mL sample volumes and a digital titrator (Hach, Loveland, CO) (Hach Company, 2013). Phenolphthalein and bromocresol green-methyl red indicators were used, and the sample was titrated with 1.6 N sulfuric acid to a grey-green endpoint (Hach Company, 2006). BOD was measured in triplicate using the 5-day BOD test (American Public Health Association, 2005). Populations of *E. coli* were determined using the Colilert defined substrate test. Briefly, 97 wells were filled with 100 mL of water sample with the Colilert substrate added. Samples were incubated for 24 hours, and wells that fluoresce under a UV light were considered positive for *E. coli*, and a most probable number

**Table 1** Sampling sites, land use, and *E. coli* concentrations in Sinking Creek. Percentage of each land cover types (Agricultural, Developed, and Forested) as well as *E. coli* Geometric means (GM), geometric standard deviations (GSD), and maximum and minimum values for each site used in the study.

Sampling site	Agricultural land use (%)	Developed land use (%)	Forested land use (%)	<i>E. coli</i> GM (GSD)	Min, Max
SC1	15.6	36.4	47.3	254.5 (3.4)	43.7,2398.8
SC2	14	37.2	48.1	182.3 (6.1)	17.4,39810.7
SC3	9.7	38	51.5	137.1 (4.0)	14.5,1737.8
SC4	9.7	37.9	51.6	169.8 (5.7)	8.5,23988.3
SC5	8.7	38.1	52.4	140.0 (7.2)	4.1,30903.0
SC6	7.1	30.2	61.6	50.2 (8.3)	0.5,8709.6
SC7	7.1	30	61.8	36.7 (9.4)	0.5,10232.9
SC8	7.7	24.3	66.8	73.9 (5.3)	10.7,8709.6
SC9	7.4	19.9	71.4	110.3 (5.8)	14.5,3981.1
SC10	5.2	6.6	86.5	70.6 (5.2)	6.2, 1995.3
SC11	5.6	3.8	89	17.2 (9.9)	0.5,1202.3
SC12	5.8	2.1	90.3	91.3 (3.8)	5.2,812.8
SC13	0	1.1	96.5	7.8 (5.5)	0.5,102.3
SC14	0	0	100	5.0 (6.1)	0.5, 245.5

estimate was made based on the number of positive wells in both the large and small wells (*American Public Health Association, 2005*). If a sample was in excess of the geometric mean United States recreational water quality criteria, the sample site was considered impaired. Impairment was based on recommendation 1, which is a threshold of  $126 \frac{CFU}{100mL}$  that corresponds to an illness rate of  $\frac{36}{10,000}$  people (*US Environmental Protection Agency, 2012*).

To get an estimation of land use throughout the Sinking Creek watershed, land cover data were downloaded from the National Land Cover Dataset (NLCD) for 2006 (*Fry et al., 2011*). Each sampling site's drainage area was delineated using StreamStats version 3 from the United States Geologic Survey (*Ries III et al., 2017*). Land was grouped into 3 categories; forested, developed, and agricultural. Forested land includes the categories deciduous forest, evergreen forest, and mixed forest. Developed land use includes all developed categories; open space (less than 20% impervious surface), low intensity (20–49% impervious surface), medium intensity (50–79% impervious surface), and high intensity (80–100% impervious surface). Agricultural land included grassland/herbaceous and pasture/hay. The area of each land use was divided by the total area of the drainage area to get the percentage land use shown in [Table 1](#), and sampling sites as well as land cover categories are shown in [Fig. 1](#).

## Modeling background

Maxent is an iterative machine learning model commonly used for mapping species distributions (*Phillips, Dudík & Schapire, 2010*). Within the sample space,  $x$ , and given a set of environmental features (parameters),  $f_1(x), f_2(x), \dots, f_n(x)$ , the Maxent distribution estimates a vector of feature weights,  $\beta = (\beta_1, \beta_2, \dots, \beta_n)$ , that maximizes the entropy of the



raw distribution of impairments,  $q_\beta(x)$ , using a Gibbs distribution,

$$q_\beta(x) = \frac{\exp(\sum_{j=1}^n \beta_j f_j(x))}{Z_\beta} \quad (1)$$

where  $Z_\beta$  is the normalization constant that ensures that  $q_\beta(x)$  integrates to one over the study area (Phillips, Dudík & Schapire, 2010). This modeling approach is justified because it provides the maximum information concerning impairment. From a water quality management standpoint, this approach is beneficial because decision-makers and stake-holders are more concerned with factors associated with impairment rather than compliance when approaching fecal pollution monitoring and management.

Original features (parameters) can be transformed into quadratic, product, hinge, and threshold feature classes so that complex multivariate responses can be modelled (Phillips & Dudík, 2008), but Maxent incorporates L1 regularization to balance satisfying the constraints on the features while minimizing overfitting. L1 regularization is not unique to Maxent, and is used in many general linear models (Elith et al., 2011). A regularization parameter  $\lambda_j$  smooths probability distributions, generating sparse solutions and removing unnecessary features; this shrinks weights to balance fit and complexity (Elith et al., 2011). Because of regularization, Maxent fits a penalized maximum likelihood model equivalent to minimizing the relative entropy dependent on the error-bound constraints

$$\max_{\beta} \frac{1}{m} \sum_{i=1}^m \ln(q_\beta(x_i)) - \sum_{j=1}^n \lambda_j |\beta_j| \quad (2)$$

subject to  $\int q_\beta(x) dx = 1$ .

Where  $m$  is the number of positive samples,  $n$  is the number of features, and  $x$  is the feature vector for occurrence point  $i$ . Eq. (2) provides insight into how Maxent uses background data: the first term is larger for models that distinguish between impairment states the best. The second term represents the regularization, which gets larger as the weights  $\beta_j$  increase, indicating a complex model more likely to over fit. The output of  $q_\beta(x)$  is termed the raw distribution, but it is difficult to interpret due to its scale dependence. More background points result in smaller raw values because their sum cannot exceed 1 over a large amount of points (Phillips & Dudok, 2008; Elith et al., 2011). For this reason, the logistical output of the Maxent model,  $P(x)$ , will be used because it represents the probability of impairment given the sample space,  $x$ . This is a logistic model using the same set of weights  $\beta$  with the intercept of the model determined by the entropy of  $q_\beta(x)$ ,  $H$ . The model is shown in Eq. (3) below.

$$P(x) = \frac{e^H q_\beta(x)}{1 + e^H q_\beta(x)}. \quad (3)$$

### Univariate models

Data were processed using a list-wise deletion process, where individual samples from a site were removed if they were missing a parameter measurement due to laboratory errors, equipment malfunctions, calibration issues, or sites being dry at the time of sampling. A

sample of 100 bootstrapped models were developed, and 20% was subsampled for testing validation. Bootstrapping is a nonparametric resampling technique to make inferences about a population based on resampling from a set population, generating population level statistics, while providing an estimate of uncertainty of those statistics (Campolongo, 1997). For this modeling approach, all background points are used in the development of the null model, and the impaired samples are bootstrapped. Although Maxent can incorporate a wide variety of feature classes, only linear and quadratic feature classes were used to develop action values, or thresholds of impairment. The rationale for using these types of feature classes is for ease of generating action values as well as to assess both linear and non-linear effects of single parameters.

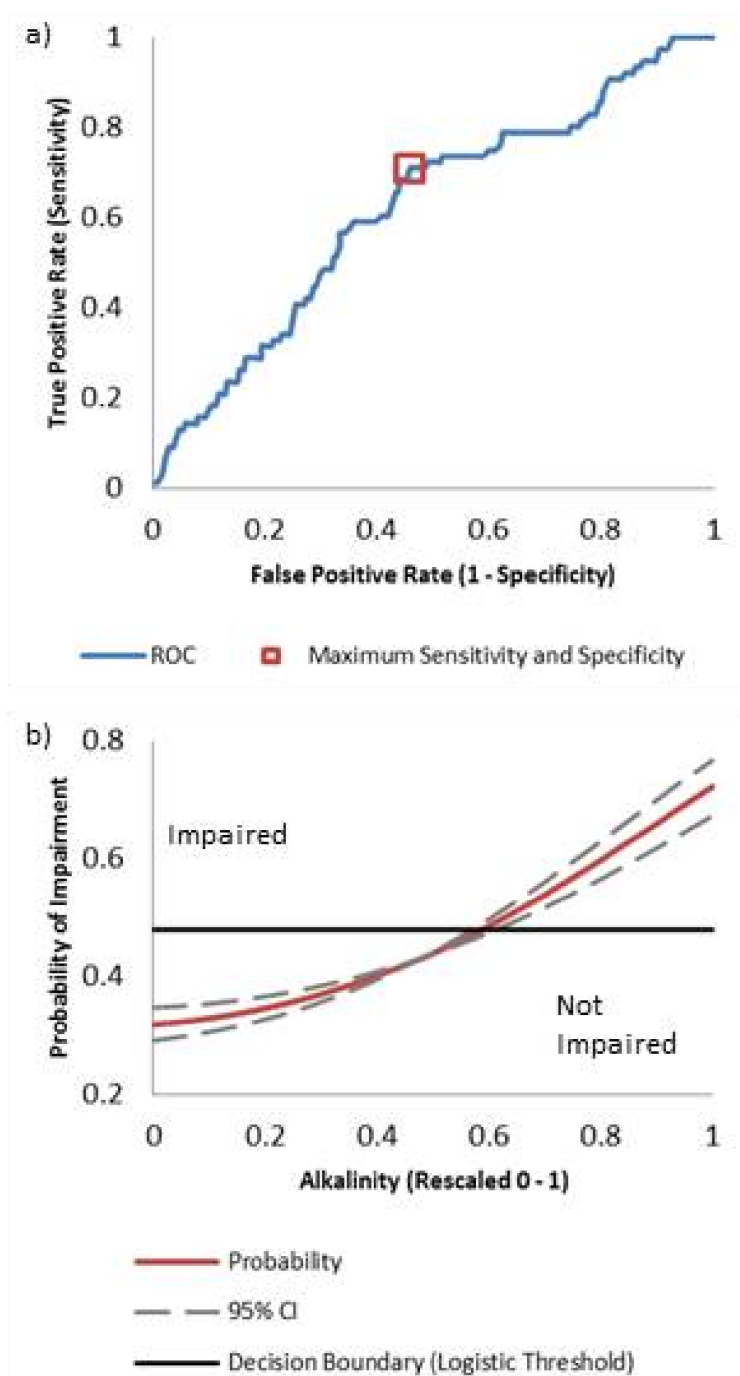
The AUC was calculated for the training and testing datasets. The AUC is a metric of performance for binary classification. The true positive prediction rate (sensitivity) and false positive prediction rates (1-specificity) of each sample are plotted as a ROC for different decision boundaries, and the area under that ROC is integrated. An AUC of 0.5 indicates that the model is no better than random chance, and a value of 1.0 indicates perfect model performance (Zweig & Campbell, 1993; Zou, O'Malley & Mauri, 2007).

The decision boundary (logistic threshold) between impaired and unimpaired samples should maximize accurately predicting impairment (exceedance of the *E. coli* criteria) while balancing correct negative predictions (Bean, Stafford & Brashares, 2012). Therefore, maximum test sensitivity and specificity was defined as the appropriate decision boundary. A low sensitivity would indicate poor performance in identifying impairment, while a low specificity would indicate an overcautious model in which resources might be wasted in remediation of false positives. Accuracy for Maxent models was calculated as follows:  $\frac{TP+TN}{TP+TN+FP+FN}$ , where TP are true positives, TN are true negatives, FP are false positives and FN are false negatives. Significance of the univariate model was determined by calculating the  $\chi^2$  statistic for each confusion matrix, with the null hypothesis being that the classifier was no better than random chance.

Action values (environmental thresholds) are conditions in which a parameter (variable) is at the threshold of impairment, indicating potential exceedance of the *E. coli* standard. Action values were calculated for significant ( $p < 0.01$ ) univariate models by averaging bootstrapped weights and estimating the parameter value at which the probability of impairment equals the logistic threshold. Figure 2 demonstrates the concepts of the AUC performance metric, selected decision boundary, and the concept of the action value in relation to the selected decision boundary and Maxent model function (Eq. (3)).

### **Multivariate models and sensitivity analysis**

Although some authors state that collinearity is not as problematic in Maxent compared to traditional regression approaches, collinearity was explored and subsequently removed using Pearson correlation coefficients (Elith et al., 2011). Variables that were highly correlated ( $r > 0.8$ ) were evaluated to determine which variable to include based on expertise, connections to previous models, and accuracy metrics within the analysis. The initial multivariate model included all noncollinear variables and were developed using 100 bootstrapped samples like the univariate models, with the addition of product feature



**Figure 2** Theoretical plots to illustrate the concept of the ROC, decision boundaries, and action values. (A) Plot of an ROC curve, where the x-axis represents the false positive rate, or the compliment of the specificity, and the y-axis represents the true positive rate, the sensitivity. The curve is integrated to obtain the AUC, the performance metric for each of the models. The box represents the point at the decision boundary (B) Theoretical plot of a univariate Maxent model function (Eq. (3)) with values for alkalinity rescaled from 0 to 1. The solid red line represents Eq. (3), the dotted lines represent the upper and lower 95% confidence intervals, and the horizontal black line represents the decision boundary. The action values, or environmental thresholds, and associated confidence intervals are the intersections between the results of Eq. (3) and the decision boundary.

Full-size DOI: 10.7717/peerj.5610/fig-2

classes to incorporate variable interaction. Average variable contribution was determined by calculating the increase in information gain associated with a change in each feature for each iteration of the model algorithm, normalized to percentages. The permutation importance of a feature is an indicator of variable sensitivity. In each model run, the feature training presence and background data are randomly permuted, and the resulting drop in training AUC is normalized to percentages for each variable and averaged over the 100 bootstrapped runs.

A jack-knife sensitivity analysis was used to determine the best subset of covariates to include in a trimmed model. Each variable was removed from the analysis, and a comparison was made to determine if the removal of a variable caused a significant ( $p < .05$ ) change in training or testing information gain. Student's  $t$ -tests were performed on each jack-knifed model to evaluate significance, and variables were included if the information gain from either the testing or training sets decreased; decrease in information would correspond to a significant loss of information, providing criterium for inclusion of the variable in final models.

## RESULTS

### Univariate model performance

The sampling program resulted in 29 sampling trips over 14 sites, allowing for a potential of 406 samples for analysis. 127 samples were removed due to missing information in the dataset, leaving 279 total samples for model development. This included 95 impairments, identified by exceedance of the *E. coli* recreational water quality standard. [Table 1](#) presents the summary statistics for *E. coli* and the associated land use in each sampling site's drainage area. Each training set included 279 background points, 76 points for training, and 19 points to evaluate performance on testing data.

[Table 2](#) summarizes the training and testing performance of the univariate Maxent models used to identify *E. coli* impairment based on environmental variables. Water temperature performed best based on AUCs, but had lower accuracy than conductivity, dissolved oxygen, and alkalinity. The plausible explanation of these differences is the latter variables had higher specificity rather than sensitivity at the chosen decision boundary ([Table S2](#)). Accuracy was found to be highest for alkalinity and lowest for pH.

Action values were developed for 8 significant univariate models by solving for the value of the variable when probability of impairment equals the logistic threshold. For example, the action value for alkalinity is 128 mg/L. This means that *E. coli* impairment is likely to occur when alkalinity is observed to be higher than this threshold. Action values and 95% confidence intervals are included in [Table 2](#). Action function graphs for each significant univariate model are presented in [Fig. S1](#) to aid in interpretation of [Table 2](#), and summary statistics for each variable are given in [Table S1](#).

### Multivariate model performance

Pearson correlations ranged from  $-0.269$  to  $0.834$ , with three variables identified as collinear; alkalinity, conductivity, and hardness. Conductivity was selected because of its use in previously developed fecal indicator models ([Wilkes et al., 2011](#); [Gonzalez et](#)

**Table 2** Summary of training and testing performance of Maxent models based AUC metrics, accuracy based on maximum test sensitivity and specificity decision boundary (logistic threshold), and action values with 95% confidence intervals. If an upper bound of a confidence interval exceeds the maximum sampling value for a set of data, the maximum value is given.

Variables	Training AUC ( $\pm$ SE)	Testing AUC ( $\pm$ SE)	Accuracy	Action values (x) $\forall$ (95% CI)
Alkalinity ( $\frac{mg}{L}$ )	0.616 (0.003)	0.620 (0.006)	68.5	$x > 129 \frac{mg}{L}$ (125, 134)
BOD ( $\frac{mg}{L}$ )	0.572 (0.004)	0.554 (0.008)	60.6	$x < 0.976 \frac{mg}{L}$ (0.825, 1.09) $x > 6.19 \frac{mg}{L}$ (4.51, 6.43)
Conductivity ( $\mu$ S)	0.628 (0.003)	0.638 (0.006)	65.6	$x > 306 \mu$ S (287, 315)
Dissolved Oxygen ( $\frac{mg}{L}$ )	0.635 (0.003)	0.640 (0.007)	67.7	$x < 9.39 \frac{mg}{L}$ (8.68, 10.6)
Discharge ( $\frac{m^3}{s}$ )	0.556 (0.004)	0.553 (0.006)	63.8	*
Hardness ( $\frac{mg}{L}$ )	0.632 (0.003)	0.627 (0.006)	59.9	$x > 132 \frac{mg}{L}$ (122, 152)
NO <sub>3</sub> ( $\frac{mg}{L}$ )	0.581 (0.004)	0.579 (0.007)	63.4	$x > 1.78 \frac{mg}{L}$ (1.63, 1.84)
pH	0.571 (0.003)	0.562 (0.006)	55.6	*
PO <sub>4</sub> ( $\frac{mg}{L}$ )	0.581 (0.004)	0.580 (0.008)	63.8	$0.0642 \frac{mg}{L} < x < 7.80 \frac{mg}{L}$ (0.0873, 0.766) $\sim$ (6.27, 9.01) $\wedge$
Water Temperature ( $^{\circ}$ C)	0.666 (0.003)	0.670 (0.005)	65.2	$x > 12.4 ^{\circ}$ C (11.3, 15.5 < x < 20.0)
8-variable model	0.770 (0.002)	0.709 (0.005)	78.5	
5-variable model	0.753 (0.002)	0.723 (0.006)	77.8	
4 variable model	0.750 (0.002)	0.726 (0.005)	77.8	

**Notes.**

\*Model was not significant.

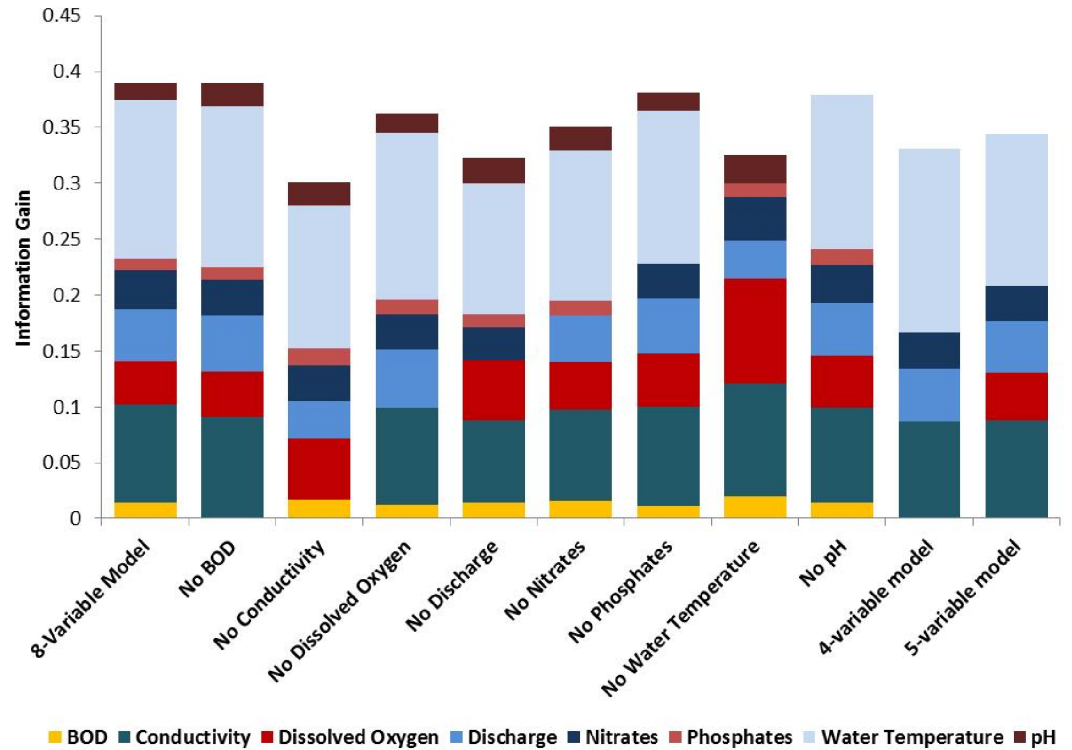
<sup>a</sup>Values of the variables that corresponded to impairment.

<sup>b</sup>95% CI for the lower bound of the action value.

<sup>c</sup>95% CI for the upper bound of the action value.

*al., 2012; Gonzalez & Noble, 2014; Piorkowski et al., 2013*). The 8-variable model displayed improved accuracy on all univariate models. Variable contribution was dominated by water temperature, conductivity, and discharge in the 8-variable model, with water temperature contributing 36.4% of the information, and conductivity and discharge accounting for 22.6% and 12.1% of the information, respectively. The permutation importance for the 8-variable model demonstrated a similar pattern. A summary of accuracy metrics is shown in [Tables 2](#) and [3](#) illustrates the contribution of each variable in the multivariate models.

Conductivity was the most sensitive parameter based on sensitivity analysis, with other sensitive parameters including water temperature, dissolved oxygen, discharge, and NO<sub>3</sub>. The removal of dissolved oxygen improved model performance based on testing AUC, justifying development of two optimized multivariate models; a 5-variable model including all sensitive parameters, and a 4-variable model that excluded dissolved oxygen.



**Figure 3** Bar graph displaying results of jack-knife sensitivity analysis. Each color represents the information gain contributed for each parameter in the model, and features are removed one at a time to assess their importance in the trimmed model.

Full-size DOI: 10.7717/peerj.5610/fig-3

**Table 3** Variable contribution and permutation importance for the multivariate models, normalized to percentages.

Variable	4-variable model		5-variable model		8-variable model	
	Percent contribution	Permutation importance	Percent contribution	Permutation importance	Percent contribution	Permutation importance
BOD					3.6	5.9
Conductivity	26.2	23.0	22.6	22.3	25.6	27.5
Discharge	14.5	22.0	12.1	20.1	13.4	21.6
Dissolved Oxygen			9.9	5.2	12.3	6.6
NO <sub>3</sub>	9.5	8.5	8.9	8.6	8.9	10.3
pH					3.9	1.7
PO <sub>4</sub>					2.7	2.5
Water temperature	49.9	46.5	36.4	33.7	39.7	34.0

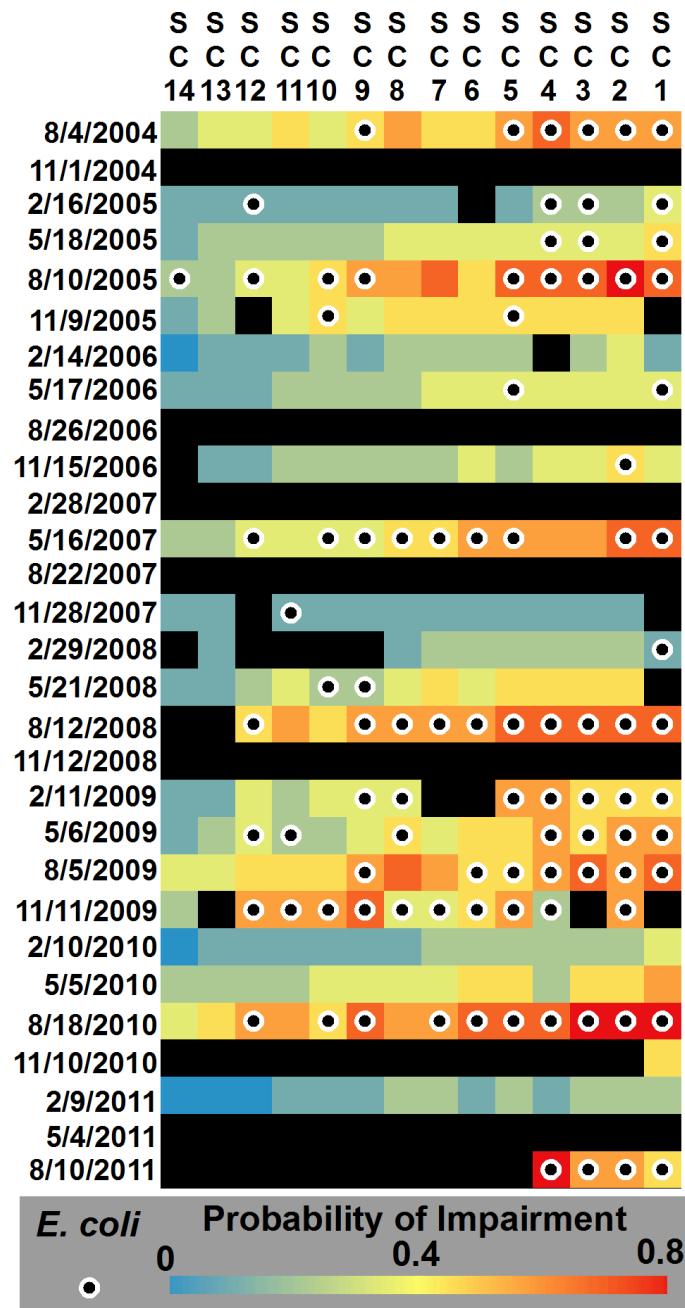
Accuracy of the 5- and 4- variable optimized model was 77.8%. The patterns of variable contribution were consistent in each model, with water temperature accounting for most of the information gain in each model. Figure 3 shows the variable contribution for the initial multivariate models, each model run during the sensitivity analysis, and the final 4-variable and 5-variable models produced. The information gain for each model is also shown within this figure.

Response surfaces were developed for each of the model runs to assess spatiotemporal trends. Each grid within the surface represents a single sample, with each sampling site representing a single column. The columns are oriented in a downstream fashion, with headwaters sites starting on the left (SC14) and sites further downstream existing on the right (SC1). The temporal scale is represented by the rows, with each row indicating a specific sampling trip. Although the data resolution is coarse, the goal is to demonstrate the potential of visualizing trends in the probability of impairment over space and time. [Figure 4](#) displays the response surface for the estimated probability of impairment for the 4-variable model and the 5- and 8-variable models are shown in [Fig. S2](#). Classification performance for the univariate models and multivariate models is shown in [Table S2](#). Mean probabilities for the 8-, 5-, and 4-variable model were 0.338 (95% CI: 0.319, 0.358), 0.353 (0.334, 0.373), and 0.359 (0.340, 0.378). Generally, the sites influenced by the greatest amount of developed or agricultural land use (SC5–SC1) had the highest probability of impairment. August had the highest probability of impairment, followed by May, November, and February. Mean probability of impairment and associated 95% confidence intervals are shown in [Table S3](#).

## DISCUSSION

Over 170,000 miles of US rivers and streams are listed as pathogen impaired based on FIOs. To address these impairments, characterization of sources and transport mechanisms is necessary ([United States Environmental Protection Agency, 2017](#)), and statistical models can be used as an inferential tool to overcome these issues. We applied Maxent to identify individual and interacting factors influencing *E. coli* fate and transport that resulted in impairments using univariate and multivariate approaches. In this particular stream, water temperature, conductivity, discharge, and NO<sub>3</sub> were found to be the most influential group of factors driving fecal pollution. The results indicate that seasonality and agricultural run-off are the suggested causes of impairment in this watershed. Seasonality is demonstrated by influence of temperature in the models, whereas the influence of agricultural run-off is suggested by the other variables and the association between land use and *E. coli* in the watershed. Even small increases in agricultural land cause substantial increases in *E. coli* concentrations ([Table 1](#)), whereas similar increases in developed land do not have the same pronounced effect. This study highlights the need for multi-month sampling across a stream continuum to truly estimate spatiotemporal variability associated with impairment.

The fact that water temperature dominated the information in this model suggests that seasonality plays an important role in *E. coli* survival. Although fecal indicators and pathogens have been found to possess diverse temperature-survival relationships ([Hofstra, 2011](#); [Sterk et al., 2013](#)), the high August probability for *E. coli* impairment indicates favorable conditions for long-term survival in the summer. Warming due to climate change could exacerbate this condition by increasing those favorable conditions ([Weniger et al., 1983](#); [Atherholt et al., 1998](#); [Patz et al., 2000](#); [Guzman Herrador et al., 2015](#)). However, August was not the only month with numerous *E. coli* impairments. Therefore, monitoring for FIOs only in the summer months could distort estimates of impairment in watersheds with year-round users.



**Figure 4** Response surface for the 4-variable Maxent model. Surface shows the probability of impairment for each sample for the monitoring program. This represents the mean probability of 100 bootstrapped runs. Rows are oriented by each sampling period, while columns represent each sampling site over the length of the stream; left to right indicates flow direction. Black cells denote samples in which a parameter was missing and were excluded from analysis, while circles with black centers represent samples in which a stream would be identified as impaired in the study.

Full-size DOI: 10.7717/peerj.5610/fig-4



Although discharge was not predictive of *E. coli* impairment alone, its interactive effect stresses the importance of both flow dependent and independent processes associated with *E. coli* impairment. Dissolved solutes such as NO<sub>3</sub> and ions measured through conductivity are largely discharge-dependent; however, FIOs are not as strongly dependent on discharge. This flow independence is due to additional ecological mechanisms such as nutrient limitation and competition (Surbeck *et al.*, 2006; Drummond *et al.*, 2015). Various forms of nitrogen are associated with increased concentration of FIOs in certain environments (Carrillo, Estrada & Hazen, 1985; Herrig *et al.*, 2015), and results of the Maxent models suggest that nutrient loading in the form of NO<sub>3</sub> contributes to *E. coli* impairment in Sinking Creek. Other studies have found that dissolved organic carbon can affect magnitude and extent of fecal indicators (Surbeck, Jiang & Grant, 2010; Blazewicz *et al.*, 2013; Cloutier, Alm & McLellan, 2015), but this was not collected during this sampling program and was found to be insignificant using BOD as a surrogate for organic pollution. This interaction between nutrient levels and fecal pollution highlights the potential for synergistic effects of different sources of pollution, suggesting a limitation of TMDL development when only considering one pollutant at a time.

Although machine learning application to microbial water quality problems is not unique, this study presents some beneficial techniques in this area of research. First, it demonstrates the ability to open the black box of Maxent, using action values to predict threshold of impairment based on a single variable. Multivariate action functions can be developed as well, but is not presented in this manuscript. The probabilistic approach to model validation and variable selection allows for inclusion of uncertainty, improving on deterministic methods traditionally used for validation and criteria for variable inclusion. Probabilistic methods have been used in TMDLs (Borsuk, Stow & Reckhow, 2002), frequency of water quality posting errors (Kim & Grant, 2004), and uncertainty of different fecal indicator methodologies (Gronewold *et al.*, 2008); this paper adds to this framework through identifying the probability of stream impairment given a set of environmental variables. This improves confidence in decision-making for implementation of monitoring, management, and remediation strategies. Modeling microbial water quality is a challenge no matter the method used, but this study demonstrates that Maxent provides a valid approach to understand the factors driving impairment.

Streams are dynamic systems with multiple flow regimes, confounding an already difficult modeling process. Understanding how models behave in extreme situations is useful for regulation, monitoring, and management of these ecosystems. Over the long-term study periods, samples from both drought and high water conditions were captured. Maxent has been suggested as a strong predictor of extreme values (Petrov, Guedes Soares & Gotovac, 2013), and this study found that Maxent sufficiently predicted impairment during the high flow sampling date of November 11, 2009. Depending on which multivariate model was used, accuracy ranged from 72.8% to 90.9% for this sampling date. Five sampling dates resulted in at least one site being dry, indicating drought-like conditions. Maxent correctly predicted impairment in these situations 62.2% to 73.0% of the time. This suggests that Maxent can be useful for certain extreme situations, but is highly dependent on the environmental variables used for prediction.

While this study presents proof of concept of using Maxent to infer source and mechanisms of impairment, there are some limitations to this study. Although the dataset has a large time scale (8 years), only collecting from 4 months makes the resolution coarse, reducing the scale at which inferences can be made. The list-wise deletion of samples before univariate modeling removed some data that could inform each of those models; however, using the same series of data in the multivariate models and list-wise deletion are commonly used procedures in statistical models. Future applications of Maxent will improve on the coarse resolution of the data by using monthly and potentially weekly sampling approaches, and research will be developed as to the best approach for handling missing data in Maxent. While AUC scores above 0.70 indicate good model fit, only considering physiochemical water quality parameters limits the potential to accurately predict impairment; however, this study demonstrates that these parameters are informative as a proof of concept for using Maxent as a modeling approach. Future areas of research include using Maxent to optimize water quality monitoring to identify causes of impairment with FIOs and specific pathogens in the most cost-effective way using a variety of microbial, chemical, and physical parameters.

It is a difficult task to develop and implement remediation strategies in watersheds with many diffuse causes of fecal impairment, but modeling can increase confidence in decision making through inferring mechanisms and sources of fecal pollution. Incorporating environmental variables into models allows for insights into the ecology of fecal indicators, identifying causes of chronic FIO impairment. Although one indicator cannot mimic the plethora of existing pathogens in water, incorporating modeling can fine tune an indicator's utility, ultimately informing the public concerning health risks, and aiding in overcoming the shortcomings of a single indicator monitoring strategy.

## CONCLUSIONS

Characterizing *E. coli* impairment is essential because of the plethora of streams polluted with fecal wastes. This study used Maxent to identify water quality parameters associated with *E. coli* impairment in a low-order, mixed-use watershed. Univariate models generated action values, or thresholds of impairment, based on single parameters, while multivariate models extracted information concerning multivariate interaction. We presented a probabilistic approach to sensitivity analysis, improving confidence in variable selection. Maxent presents a flexible machine learning approach to aid in understanding mechanisms and sources of fecal pollution as well as a host of other complex decision boundary problems. We demonstrated that:

- Models using alkalinity and water temperature were found to be either the most accurate or best performing univariate models; this stresses the importance of discharge composition and seasonality in *E. coli* impairment. Discharge, however, was not an influential univariate parameters by itself, stressing the importance of flow-independent processes that correlate with impairment.

- Sensitivity analysis indicated that the most information was lost when conductivity was removed from the multivariate models, and water temperature, discharge, dissolved oxygen, and  $\text{NO}_3$  represent other sensitive parameters sensitive to *E. coli* impairment in this watershed.
- Results suggest that *E. coli* impairment in this stream is driven by seasonality and agricultural run-off. This suggests that multi-month sampling along a stream continuum is essential to characterize spatiotemporal variability, importance of flow in relation to other water quality parameters, and the potential synergistic effect of nutrient and fecal pollution.
- Incorporating modeling can fine tune an indicator's utility, informing the public concerning human health risks, enhancing our understanding of FIOs, assisting in water quality decision-making, and providing input variables for quantitative microbial risk assessment.

### Abbreviations

AUC	Area under the curve
BOD	Biochemical Oxygen Demand
FIO	Fecal Indicator Organism
NLCD	National Land Cover Dataset
ROC	Receiver Operating Characteristic
TMDL	Total maximum daily load

## ACKNOWLEDGEMENTS

The authors thank Brian Evanshen for oversight of sample collection and data management.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This work was supported by a contract with the Tennessee Valley Authority (Award # 0025252) and a grant from the East Tennessee State University's school of Graduate Studies and Graduate Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:

Tennessee Valley Authority: # 0025252.

East Tennessee State University's school of Graduate Studies and Graduate Council.

### Competing Interests

The authors declare there are no competing interests.

## Author Contributions

- Dennis Gilfillan and Timothy A. Joyner conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Phillip Scheuerman authored or reviewed drafts of the paper, approved the final draft, allowed access to data, mentored, etc.

## Data Availability

The following information was supplied regarding data availability:

The raw data are provided in the [Supplemental Files](#).

## Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.5610#supplemental-information>.

## REFERENCES

- Alm EW, Burke J. 2003.** Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water Research* **37**:3978–3982 DOI [10.1016/S0043-1354\(03\)00301-4](https://doi.org/10.1016/S0043-1354(03)00301-4).
- American Public Health Association. 2005.** *Standard methods for the examination of water and wastewater*. 21st edition. Washington, D.C.: American Public Health Association.
- Anderson KL, Whitlock JE, Valerie J, Harwood VJ. 2005.** Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. *Applied and Environmental Microbiology* **71**:3041–3048 DOI [10.1128/AEM.71.6.3041-3048.2005](https://doi.org/10.1128/AEM.71.6.3041-3048.2005).
- Atherholt TB, Lechevallier MW, Norton WD, Rosen JS. 1998.** Effect of rainfall on Giardia and Cryptosporidium. *American Water Works Association* **90**:66–80.
- Ballestè E, Bonjoch X, Belanche LA, Blanch AR. 2010.** Molecular indicators used in the development of predictive models for microbial source tracking. *Applied and Environmental Microbiology* **76**:1789–1795 DOI [10.1128/AEM.02350-09](https://doi.org/10.1128/AEM.02350-09).
- Bean WT, Stafford R, Brashares JS. 2012.** The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography* **35**:250–258 DOI [10.1111/j.1600-0587.2011.06545.x](https://doi.org/10.1111/j.1600-0587.2011.06545.x).
- Belanche-Muñoz L, Blanch AR. 2008.** Machine learning methods for microbial source tracking. *Environmental Modelling & Software* **23**:741–750 DOI [10.1016/j.envsoft.2007.09.013](https://doi.org/10.1016/j.envsoft.2007.09.013).
- Benham BL, Baffaut C, Zeckoski RW, Mankin KR, Pachepsky YA, Sadeghi AM, Brannan KM, Soupier ML, Habersack MJ. 2006.** Modeling bacteria fate and transport in watersheds to support TMDLs. *Transactions of the ASABE* **49**:987–1002 DOI [10.13031/2013.21739](https://doi.org/10.13031/2013.21739).
- Black LE, Brion GM, Freitas SJ. 2007.** Multivariate logistic regression for predicting total culturable virus presence at the intake of a potable-water treatment plant: novel

- application of the atypical coliform / total coliform ratio. *Applied and Environmental Microbiology* 73:3965–3974 DOI 10.1128/AEM.02780-06.
- Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. 2013.** Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME Journal* 7:2061–2068 DOI 10.1038/ismej.2013.102.
- Blount ZD. 2015.** The unexhausted potential of *E. coli*. *Elife* 4:e05826 DOI 10.7554/eLife.05826.
- Boeckmann M, Joyner TA. 2014.** Old health risks in new places? An ecological niche model for *I. ricinus* tick distribution in Europe under a changing climate. *Health Place* 30:70–77 DOI 10.1016/j.healthplace.2014.08.004.
- Borsuk ME, Stow CA, Reckhow KH. 2002.** Predicting the frequency of water quality standard violations: a probabilistic approach for TMDL development. *Environmental Science and Technology* 36:2109–2115.
- Brion GM, Lingireddy S. 1999.** A neural network approach to identifying non-point sources of microbial contamination. *Water Research* 33:3099–3106 DOI 10.1016/S0043-1354(99)00025-1.
- Brion GM, Neelakantan TR, Lingireddy S. 2002.** A neural-network-based classification scheme for sorting sources and ages of fecal contamination in water. *Water Research* 36:3765–3774.
- Byappanahalli MN, Shively DA, Nevers MB, Sadowsky MJ, Whitman RL. 2003.** Growth and survival of *Escherichia coli* and enterococci populations in the macro-alga *Cladophora* (Chlorophyta). *FEMS Microbiology Ecology* 46:203–211 DOI 10.1016/S0168-6496(03)00214-9.
- Campolongo F. 1997.** Sensitivity analysis of an environmental model: an application of different analysis methods. *Reliability Engineering & System Safety* 57:49–69 DOI 10.1016/S0951-8320(97)00021-5.
- Carrillo M, Estrada E, Hazen TC. 1985.** Survival and enumeration of the fecal indicators *Bifidobacterium adolescentis* and *Escherichia coli* in a tropical rain forest watershed. *Applied and Environmental Microbiology* 50:468–476.
- Characklis GW, Dilts MJ, Simmons OD, Likirdopulos CA, Krometis LAH, Sobsey MD. 2005.** Microbial partitioning to settleable particles in stormwater. *Water Research* 39:1773–1782 DOI 10.1016/j.watres.2005.03.004.
- Cloutier D, Alm E, McLellan S. 2015.** The influence of land-use, nutrients, and geography on microbial communities and fecal indicator abundance at lake michigan beaches. *Applied and Environmental Microbiology* 81:4904–4913 DOI 10.1128/AEM.00233-15.
- Coulliete A, Money ES, Serre ML, Noble RT. 2009.** Space/time analysis of fecal pollution and rainfall in an eastern north carolina estuary. *Environmental Science and Technology* 43:3728–3735.
- De Brauwere A, Ouattara NK, Servais P. 2014.** Modeling fecal indicator bacteria concentrations in natural surface waters: a review. *Critical Reviews Environmental Science Technology* 44:2380–2453.

- Drummond JD, Davies-Colley RJ, Stott R, Sukias JP, Nagels JW, Sharp A, Packman AI.** 2015. Microbial transport, retention, and inactivation in streams: a combined experimental and stochastic modeling approach. *Environmental Science and Technology* 49:7825–7833 DOI 10.1021/acs.est.5b01414.
- Dwivedi D, Mohanty BP, Lesikar BJ.** 2013. Estimating *Escherichia coli* loads in streams based on various physical, chemical, and biological factors. *Water Resources Research* 49:2896–2906 DOI 10.1002/wrcr.20265.
- Dwivedi D, Mohanty BP, Lesikar BJ.** 2016. Impact of the linked surface water-soil water-groundwater system on transport of *E. coli* in the subsurface. *Water, Air, and Soil Pollution* 227:351 DOI 10.1007/s11270-016-3053-2.
- Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ.** 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17:43–57 DOI 10.1111/j.1472-4642.2010.00725.x.
- Ferguson C, Husman AM de R, Altavilla N, Deere D, Ashbolt N.** 2003. fate and transport of surface water pathogens in watersheds. *Critical Reviews Environmental Science Technology* 33:299–361 DOI 10.1080/10643380390814497.
- Field KG, Samadpour M.** 2007. Fecal source tracking, the indicator paradigm, and managing water quality. *Water Research* 41:3517–3538 DOI 10.1016/j.watres.2007.06.056.
- Fry JA, Xian G, Jin S, Dewitz JA, Homer CG, Limin Y, Barnes CA, Herold ND, Wickham JD.** 2011. Completion of the 2006 national land cover database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing* 77:858–864.
- Gonzalez RA, Conn KE, Crosswell JR, Noble RT.** 2012. Application of empirical predictive modeling using conventional and alternative fecal indicator bacteria in eastern North Carolina waters. *Water Research* 46:5871–5882 DOI 10.1016/j.watres.2012.07.050.
- Gonzalez RA, Noble RT.** 2014. Comparisons of statistical models to predict fecal indicator bacteria concentrations enumerated by qPCR- and culture-based methods. *Water Research* 48:296–305 DOI 10.1016/j.watres.2013.09.038.
- Gronewold AD, Borsuk ME, Wolpert RL, Reckhow KH.** 2008. An assessment of fecal indicator bacteria-based water quality standards. *Environmental Science and Technology* 42:4676–4682.
- Guisan A, Broennimann O, Engler R, Vust M, Yoccoz NG, Lehmann A, Zimmermann NE.** 2006. Using niche-based models to improve the sampling of rare species. *Conservation Biology* 20:501–511.
- Guzman Herrador BR, De Blasio BF, MacDonald E, Nichols G, Sudre B, Vold L, Semenza JC, Nygård K.** 2015. Analytical studies assessing the association between extreme precipitation or temperature and drinking water-related waterborne infections: a review. *Environmental Health* 14:29 DOI 10.1186/s12940-015-0014-y.
- Hach Company.** 2006. Digital Titrator—model 16900. In: *Procedure manual*. Loveland: Hach Company.
- Hach Company.** 2013. DR/890 colorimeter. In: *Procedures manual*. Loveland: Hach Company.

- Hall KK, Evanshen BG, Maier KJ, Scheuerman PR. 2014. Application of multivariate statistical methodology to model factors influencing fate and transport of fecal pollution in surface waters. *Journal of Environmental Quality* **43**:358–370 DOI [10.2134/jeq2013.05.0190](https://doi.org/10.2134/jeq2013.05.0190).
- Herrig IM, Böer SI, Brennholt N, Manz W. 2015. Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. *Water Research* **85**:148–157 DOI [10.1016/j.watres.2015.08.006](https://doi.org/10.1016/j.watres.2015.08.006).
- Hofstra N. 2011. Quantifying the impact of climate change on enteric waterborne pathogen concentrations in surface water. *Current Opinion in Environmental Sustainability* **3**:471–479 DOI [10.1016/j.cosust.2011.10.006](https://doi.org/10.1016/j.cosust.2011.10.006).
- Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ. 2006. Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior watersheds. *Applied and Environmental Microbiology* **72**:612–621 DOI [10.1128/AEM.72.1.612-621.2006](https://doi.org/10.1128/AEM.72.1.612-621.2006).
- Jamieson R, Gordon R, Joy D, Lee H. 2004. Assessing microbial pollution of rural surface waters: a review of current watershed scale modeling approaches. *Agricultural Water Management* **70**:1–17 DOI [10.1016/j.agwat.2004.05.006](https://doi.org/10.1016/j.agwat.2004.05.006).
- Kim JH, Grant SB. 2004. Public mis-notification of coastal water quality: a probabilistic evaluation of posting errors at Huntington Beach, California. *Environmental Science and Technology* **38**:2497–2504 DOI [10.1021/es034382v](https://doi.org/10.1021/es034382v).
- Kistemann T, Claßen T, Koch C, Dangendorf F, Fischeder R, Gebel J, Vacata V, Exner M. 2002. Microbial load of drinking water reservoir tributaries during extreme rainfall and runoff. *Applied and Environmental Microbiology* **68**:2188–2197 DOI [10.1128/AEM.68.5.2188-2197.2002](https://doi.org/10.1128/AEM.68.5.2188-2197.2002).
- LaLiberte P, Grimes DJ. 1982. Survival of *Escherichia coli* in lake bottom sediment. *Applied and Environmental Microbiology* **43**:623–628.
- Lasalde C, Rodriguez R, Toranzos GA, Smith HH. 2005. Heterogeneity of uidA gene in environmental *Escherichia coli* populations. *Journal of Water and Health* **3**:297–304.
- Lipp EK, Kurz R, Vincent R, Rodriguez-Palacios C, Farrah SR, Rose JB. 2001. The effects of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a subtropical estuary. *Estuaries* **24**:266–276 DOI [10.2307/1352950](https://doi.org/10.2307/1352950).
- Lozier JD, Aniello P, Hickerson MJ. 2009. Predicting the distribution of *Sasquatch* in western North America: anything goes with ecological niche modelling. *Journal of Biogeography* **36**:1–5 DOI [10.1111/j.1365-2699.2009.02152.x](https://doi.org/10.1111/j.1365-2699.2009.02152.x).
- Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences* DOI [10.1073/pnas.1015622108](https://doi.org/10.1073/pnas.1015622108).
- Lušić DV, Kranjčević L, Maćešić S, Lušić D, Jozić S, Linšak Ž, Bilajac L, Grbčić L, Bilajac N. 2017. Temporal variations analyses and predictive modeling of microbiological seawater quality. *Water Research* **119**:160–170 DOI [10.1016/j.watres.2017.04.046](https://doi.org/10.1016/j.watres.2017.04.046).
- Maier RM, Pepper IL, Gerba CP. 2009. *Environmental microbiology*. 2nd edition. New York: Elsevier.

- McCambridge J, McMeekin TA. 1980.** Relative effects of bacterial and protozoan predators on survival of *Escherichia coli* in estuarine water samples. *Applied and Environmental Microbiology* **40**:907–911.
- McKergow LA, Davies-Colley RJ. 2010.** Stormflow dynamics and loads of *Escherichia coli* in a large mixed land use catchment. *Hydrological Process* **24**:276–289.
- McLellan SL, Eren AM. 2014.** Discovering new indicators of fecal pollution. *Trends in Microbiology* **22**:697–706 DOI [10.1016/j.tim.2014.08.002](https://doi.org/10.1016/j.tim.2014.08.002).
- Molina M, Hunter S, Cyterski M, Peed LA, Kelty CA, Sivaganesan M, Mooney T, Prieto L, Shanks OC. 2014.** Factors affecting the presence of human-associated and fecal indicator real-time quantitative PCR genetic markers in urban-impacted recreational beaches. *Water Research* **64**:196–208 DOI [10.1016/j.watres.2014.06.036](https://doi.org/10.1016/j.watres.2014.06.036).
- Money ES, Carter GP, Serre ML. 2009.** Modern space/time geostatistics using river distances: data integration of turbidity and *E.coli* measurements to assess fecal contamination along the Raritan River in New Jersey. *Environmental Science and Technology* **43**:3736–3742.
- Patz JA, Mcgeehin MA, Bernard SM, Ebi KL, Epstein PR, Gubler DJ, Reiter P, Romieu I, Rose JB, Samet JM. 2000.** The potential health impacts of climate variability and change for the United States: executive summary of the report of the health sector of the US National Assessment. *Environmental Health Perspectives* **108**:367–376.
- Perkins TL, Perrow K, Rajko-Nenow P, Jago CF, Jones DL, Malham SK, McDonald JE. 2016.** Decay rates of faecal indicator bacteria from sewage and ovine faeces in brackish and freshwater microcosms with contrasting suspended particulate matter concentrations. *Science of the Total Environment* **572**:1645–1652 DOI [10.1016/j.scitotenv.2016.03.076](https://doi.org/10.1016/j.scitotenv.2016.03.076).
- Petrov V, Guedes Soares C, Gotovac H. 2013.** Prediction of extreme significant wave heights using maximum entropy. *Coastal Engineering* **74**:1–10 DOI [10.1016/j.coastaleng.2012.11.009](https://doi.org/10.1016/j.coastaleng.2012.11.009).
- Phillips SJ, Dudík M. 2008.** Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**:161–175 DOI [10.1111/j.0906-7590.2008.5203.x](https://doi.org/10.1111/j.0906-7590.2008.5203.x).
- Phillips SJ, Dudík M, Schapire RE. 2004.** A maximum entropy approach to species distribution modeling. In: *Proceedings of the twenty-first international conference on machine learning*. ACM, 83.
- Phillips S, Dudík M, Schapire R. 2010.** Maxent software. Version 3.3.3k. Available at [http://biodiversityinformatics.amnh.org/open\\_source/maxent/](http://biodiversityinformatics.amnh.org/open_source/maxent/).
- Piorkowski G, Jamieson R, Bezanson G, Truelstrup L, Yost C. 2013.** Evaluation of statistical models for predicting *Escherichia coli* particle attachment in fluvial systems. *Water Research* **47**:6701–6711 DOI [10.1016/j.watres.2013.09.003](https://doi.org/10.1016/j.watres.2013.09.003).
- Ries III KG, Newson JK, Smith MJ, Guthrie JD, Steeves PA, Haluska TL, Kolb KR, Thompson RF, Santoro RD, Vraga HW. 2017.** *StreamStats, version 4*. Reston, VA: USGS Fact Sheet DOI [10.3133/fs20173046](https://doi.org/10.3133/fs20173046).
- Savichtcheva O, Okabe S. 2006.** Alternative indicators of fecal pollution: Relations with pathogens and conventional indicators, current methodologies for direct pathogen



- monitoring and future application perspectives. *Water Research* **40**:2463–2476 DOI 10.1016/j.watres.2006.04.040.
- Sinton LW, Hall CH, Lynch PA, Davies-Colley RJ. 2002.** Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. *Applied and Environmental Microbiology* **68**:1122–1131.
- Smith A, Sterba-Boatwright B, Mott J. 2010.** Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Research* **44**:4067–4076 DOI 10.1016/j.watres.2010.05.019.
- Sterk A, Schijven J, De Nijs T, De Roda Husman AM. 2013.** Direct and indirect effects of climate change on the risk of infection by water-transmitted pathogens. *Environmental Science and Technology* **47**:12648–12660.
- Stott R, Davies-Colley R, Nagels J, Donnison A, Ross C, Muirhead R. 2011.** Differential behaviour of *Escherichia coli* and *Campylobacter* spp. in a stream draining dairy pasture. *Journal of Water and Health* **9**:59–69 DOI 10.2166/wh.2010.061.
- Surbeck CQ, Jiang SC, Ahn JH, Grant SB. 2006.** Flow fingerprinting fecal pollution and suspended solids in stormwater runoff from an urban coastal watershed. *Environmental Science and Technology* **40**:4435–4441 DOI 10.1021/es060701h.
- Surbeck CQ, Jiang SC, Grant SB. 2010.** Ecological control of fecal indicator bacteria in an urban stream. *Environmental Science and Technology* **44**:631–637 DOI 10.1021/es903496m.
- Tennessee Department of Environmental and Conservation. 2006.** *Proposed total maximum daily load (tmdl) for E. Coli in the Water River Watershed (HUC 06010103)*. TN: TDEC Nashville.
- Thuiller W, Richardson DM, PYŠEK P, Midgley GF, Hughes GO, Rouget M. 2005.** Niche based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology* **11**:2234–2250.
- Tyrrel SF, Quinton JN. 2003.** Overland flow transport of pathogens from agricultural land receiving faecal wastes. *Journal of Applied Microbiology* **94**:87–93 DOI 10.1046/j.1365-2672.94.s1.10.x.
- United States Environmental Protection Agency. 2017.** National summary of impaired waters and TMDL information [WWW Document]. Available at [http://iaspub.epa.gov/waters10/attains\\_nation\\_cy.control?p\\_report\\_type=T](http://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T) (accessed on 08 January 2017).
- US Environmental Protection Agency. 2012.** *Recreational water quality criteria*. Washington, D.C.: Office of Water.
- Viau EJ, Goodwin KD, Yamahara KM, Layton BA, Sassoubre LM, Burns SL, Tong H-I, Wong SHC, Lu Y, Boehm AB. 2011.** Bacterial pathogens in Hawaiian coastal streams—Associations with fecal indicators, land cover, and water quality. *Water Research* **45**:3279–3290 DOI 10.1016/j.watres.2011.03.033.
- Vidon P, Campbell MA, Gray M. 2008.** Unrestricted cattle access to streams and water quality in till landscape of the Midwest. *Agricultural Water Management* **95**:322–330.

- Wade TJ, Pai N, Eisenberg JNS, Colford JM. 2003.** Do US Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environmental Health Perspectives* **111**:1102–1109.
- Weniger BG, Blaser MJ, Gedrose J, Lippy EC, Juranek DD. 1983.** An outbreak of waterborne giardiasis associated with heavy water runoff due to warm weather and volcanic ashfall. *American Journal of Public Health* **73**:868–872 DOI [10.2105/AJPH.73.8.868](https://doi.org/10.2105/AJPH.73.8.868).
- Wilkes G, Edge TA, Gannon VPJ, Jokinen C, Lyautey E, Neumann NF, Ruecker N, Scott A, Sunohara M, Topp E, Lapen DR. 2011.** Associations among pathogenic bacteria, parasites, and environmental and land use factors in multiple mixed-use watersheds. *Water Research* **45**:5807–5825 DOI [10.1016/j.watres.2011.06.021](https://doi.org/10.1016/j.watres.2011.06.021).
- Winfield MD, Groisman EA. 2003.** Role of nonhost environments in the lifestyles of Salmonella and Escherichia coli. *Applied and Environmental Microbiology* **69**:3687–3694 DOI [10.1128/AEM.69.7.3687](https://doi.org/10.1128/AEM.69.7.3687).
- Yates MV. 2007.** Classical indicators in the 21st century—far and beyond the coliform. *Water Environment Research* **79**:279–286.
- Zou KH, O'Malley AJ, Mauri L. 2007.** Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* **115**:654–657.
- Zweig MH, Campbell G. 1993.** Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**:561–577.