Maxent estimation of aquatic *Escherichia coli* stream impairment

Dennis Gilfillan [Corresp., 1] , **Timothy A Joyner** [2] , **Phillip Scheuerman** [1]

[1] Department of Environmental Health Sciences, East Tennessee State University, Johnson City, TN, United States

[2] Department of Geosciences, East Tennessee State University, Johnson City, TN, United States

Corresponding Author: Dennis Gilfillan
Email address: gilfillan@etsu.edu

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_3. 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 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.

1 **Maxent estimation of aquatic *Escherichia coli* stream impairment**

2 Dennis Gilfillan[a] , Timothy Andrew Joyner[b], Phillip Scheuerman[a]

4 [a]Department of Environmental Health, East Tennessee State University, Johnson City, TN,

5 United States of America

7 [b] Department of Geosciences, East Tennessee State University, Johnson City, TN, United States

8 of America

9 Timothy Joyner: Joynert@mail.etsu.edu

11 Corresponding Author:

12 Dennis Gilfillan

13 Email address: Gilfillan@etsu.edu

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

36   variables included water temperature, dissolved oxygen, discharge, and NO_3. The removal of

37   dissolved oxygen improved model performance based on testing AUC, justifying development

38   of two optimized multivariate models; a 5-variable model including all sensitive parameters,

39   and a 4-variable model that excluded dissolved oxygen.

40   **Discussion.** Results suggest that *E. coli* impairment in Sinking Creek is influenced by seasonality

41   and agricultural run-off, stressing the need for multi-month sampling along a stream

42   continuum. Although discharge was not predictive of *E. coli* impairment alone, its interactive

43   effect stresses the importance of both flow dependent and independent processes associated

44   with *E. coli* impairment. This research also highlights the interactions between nutrient and

45   fecal pollution, a key consideration for watersheds with multiple synergistic impairments.

46   Although one indicator cannot mimic the plethora of existing pathogens in water, incorporating

47   modeling can fine tune an indicator's utility, providing information concerning fate, transport,

48   and source of fecal pollution while prioritizing resources and increasing confidence in decision

49   making.

50

51   Abbreviations

52   AUC – Area under the curve

53   BOD – Biochemical Oxygen Demand

54   FIO – Fecal Indicator Organism

55   NLCD- National Land Cover Dataset

56   ROC – Receiver Operating Characteristic

57   TMDL – Total maximum daily load

58

1. Introduction

59

60 Rapid urbanization of rural areas causes deterioration in water quality, rendering many water

61 bodies unfit for their domestic and recreational use. An assortment of contaminants is

62 introduced into aquatic systems, but pathogens represent the major cause of stream

63 impairment in the United States (United States Environmental Protection Agency, 2017).

64 Pathogens are difficult to measure directly because of their sporadic distribution, costly

65 identification, and potential health risks to laboratory workers (Field and Samadpour, 2007).

66 Most pathogens in aquatic systems stem from human and animal fecal wastes, including direct

67 deposition of feces in water (Vidon et al., 2008), run-off from land with fecal deposits (Tyrrel

68 and Quinton, 2003; Jamieson et al., 2004), and sanitary sewer malfunctions (Ferguson et al.,

69 2003; McLellan and Eren, 2014). To address the difficulties in monitoring specific pathogens,

70 fecal indicator organisms (FIOs) are commonly used to assess the presence of fecal pathogens.

71

72 An effective fecal indicator is associated with the presence of specific pathogens, with a

73 straightforward method for enumeration that correlates with magnitude and age of fecal

74 pollution (Savichtcheva and Okabe, 2006; Maier et al., 2009).   The use of FIOs such as fecal

75 coliform bacteria and *Escherichia coli* are traditionally used for determining surface water

76 pathogen impairment (Yates, 2007; US Environmental Protection Agency, 2012).   Although

77 these indicators assist in alerting populations when exposure to pathogens is likely, the current

78 approach is limited by using a single indicator such as *E. coli* for a designated use (Wade et al.,

79 2003; Savichtcheva and Okabe, 2006; Field and Samadpour, 2007). The cosmopolitan nature of

80 *E. coli* in warm-blooded animals makes them impractical for source identification (Field and

81   Samadpour, 2007; Yates, 2007; McLellan and Eren, 2014; Blount, 2015). The ability of *E. coli* to

82   survive in soils (Lasalde et al., 2005; Ishii et al., 2006), algae (Byappanahalli et al., 2003), and

83   sediments (LaLiberte and Grimes, 1982; Alm and Burke, 2003; Drummond et al., 2015) provide

84   a reservoir for continued persistence and potential to naturalize (Winfield and Groisman, 2003;

85   Lasalde et al., 2005; Luo et al., 2011). These characteristics and deficiencies emphasize the

86   difficulty of single standard FIO monitoring for impairment, stressing the need for additional

87   methods to evaluate source and mechanisms of FIO impairment.

88

89   In addition to the above issues, appropriately characterizing FIO impairment for regulation and

90   decision-making is difficult due to complex fate and transport processes (Benham et al., 2006;

91   de Brauwere et al., 2014; Drummond et al., 2015). These complex fate and transport processes

92   include transport through run-off and storm water(Kistemann et al 2002, Lipp et al 2001;

93   McKergow and Davies-Colley, 2010), remobilization from sediments and hyporheic exchange

94   (Drummond 2015; Dwivedi 2016), particle attachment (Characklis, 2005), and UV light exposure

95   (Sinton 2002). Additionally, ecological processes control FIO fate and transport through variable

96   survival patterns of indicators and pathogens (Anderson et al., 2005; Stott et al., 2011),

97   availability of nutrients and organic matter (Surbeck et al. 2010; Perkins et al. 2016), and

98   predation (McCambridge and McMeekin, 1981). Appropriately characterizing the physics and

99   ecology driving fate and transport can better inform management decisions for total maximum

100   daily load (TMDL) development, reduction of pollution, and allocation of resources.

101

102   Modeling provides flexible approaches to infer sources and processes associated with FIOs and

103   other pathogens, overcoming some of the issues of the single indicator paradigm. Various

104   statistical and machine learning models have been used to approach such problems of

105   incorporating age of fecal pollution for source tracking or detection of viruses (Brion et al. 2002;

106   Black et al., 2007); identifying land use, environmental, and water quality parameters

107   associated with FIOs and pathogens (Brion and Lingireddy, 1999; Viau et al., 2011; Wilkes et al.,

108   2011; Gonzalez et al., 2012; Gonzalez and Noble, 2014; Hall et al., 2014; Herrig et al., 2015;

109   Lušić et al., 2017); determining factors influencing particle attachment and virulence

110   (Piorkowski et al., 2013); and optimizing microbial source tracking (Belanche-Muñoz and

111   Blanch, 2008; Ballestè et al., 2010; Smith et al., 2010; Molina et al., 2014). Some other

112   applications of modeling include using turbidity or rainfall to predict *E. coli* concentrations at

113   unmonitored sites (Money et al. 2009, Coulliete et al. 2009), estimating E. coli loads using

114   physical, chemical, and biological factors within a neural network (Dwivedi 2013), and

115   hyporheic-groundwater interactions associated with transport of *E. coli* within sediments

116   porewater (Dwivedi 2016). Modeling can inform decision-makers concerning what drives

117   impairment, addressing some of the shortcomings of a single indicator approach.

118

119   Maxent, a commonly used ecological niche model (Phillips et al., 2004; Phillips and Dudik 2008),

120   identified environmental variables associated with probability of *E. coli* stream impairment,

121   making inferences concerning source and mechanisms driving fecal pollution. Although

122   modeling *E. coli* using a machine learning model such as Maxent is not a novel approach, e.g.

123   Dwivedi et al. 2013, this study is unique in the following ways: it focuses on how the water

124   quality is associated with *E. coli* impairment in lower order streams, uses nonparametric

125   bootstrapping as a probabilistic assessment of model performance based on the area under

126   the curve (AUC) of the receiving operator characteristic (ROC), and uses loss of information as

127   an indicator of sensitive variables. Ecological niche models have been utilized for species

128   distribution (Lozier et al., 2009), conservation of rare species (Guisan et al., 2006), invasive

129   species (Thuiller et al., 2005), and disease vector epidemiology studies (Boeckmann and Joyner,

130   2014), but this is a new application of Maxent to microbial water quality. Additionally,

131   developing models in lower order streams has not been previously reported; this is important

132   because water from low order streams is used for domestic water supply and recreation in

133   many areas of the United States.

134

135   The motivation for using Maxent to predict *E. coli* impairment is to investigate how

136   environment, i.e water quality parameters, shapes the niche of *E. coli* impairment based on a

137   decision boundary; in this case, a water quality standard. A probabilistic procedure for

138   univariate and multivariate model development is presented using nonparametric

139   bootstrapping cross-validation. Univariate models generated action values, or environmental

140   thresholds of impairment, to indicate potential *E. coli* impairment based on a single parameter.

141   Multivariate models predicted probability of impairment given a suite of environmental

142   variables, and jack-knife sensitivity analysis removed unresponsive variables in multivariate

143   models to elicit a set of the most responsive water quality parameters. Using Maxent to model

144   how water quality influences *E. coli* impairment aids in inferring source and mechanisms of

145   fecal pollution. This approach allows for estimation of both linear and non-linear effects of

146   water quality, demonstrates a probabilistic method for variable selection, and reframes the

147   question from "How much *E. coli* in our watershed?" to "what factors separate *E. coli*

148   impairment from compliance?," which is useful when evaluating watershed decisions.

149

150   **2. Methods**

151   **2.1 Sampling sites and data collection**

152   Sinking Creek is a 1st to 3rd Strahler order mixed-use stream that is noncompliant for state of

153   Tennessee standards for fecal coliform and *E. coli* (Tennessee Department of Environmental

154   and Conservation, 2006). Starting in August 2004, samples were collected by hand in August,

155   November, February, and May of each year until August 2011 as a long-term monitoring plan at

156   14 sites in Sinking Creek, and samples were analyzed for 10 water quality parameters and

157   populations of *E. coli* (Figure 1).

158

159   Specific conductance (conductivity) and water temperature were measured using an Orion

160   115A+ conductivity meter (Thermo Fisher Scientific, Waltham, MA). The pH was measured

161   using an EL2 portable pH meter (Mettler Toledo, Columbus, OH). Dissolved oxygen was

162   collected using a YSI Model 55 dissolved oxygen meter (YSI Inc., Yellow Springs, OH). Samples

163   for nitrates (NO_3), phosphates (PO_4), biochemical oxygen demand (BOD), alkalinity, and

164   hardness were collected in clean 2 L polyethylene bottles and stored at 4°C until laboratory

165   analysis. A flow meter (Global Water, FP101) was placed in the center of the channel to

166   measure stream velocity. Stream width was calculated where the stream velocity was

167   measured, and depth was averaged over three points across the stream width. Velocity was

168   multiplied by stream width and average depth to estimate discharge.

169

170   NO_3 and PO_4 analyses were performed in triplicate using colorimetric HACH™ methods (Hach,

171   Loveland, CO) and reagents. NO_3 and PO_4 analyses were conducted by adding 10 mL of water

172   to a vial containing NitraVer5 or PhosVer3 for the respective analyses. Vials were shaken to

173   dissolve the reagent and samples were analyzed with a DR890 colorimeter (Hach, Loveland, CO)

174   (HACH Company, 2013). Triplicate sample for alkalinity and hardness were determined using

175   100 mL sample volumes and a digital titrator (Hach, Loveland, CO) (HACH Company, 2013).

176   Phenolphthalein and bromcresol green-methyl red indicators were used, and the sample was

177   titrated with 1.6 N sulfuric acid to a grey-green endpoint (Hach Company, 2006). BOD was

178   measured in triplicate using the 5-day BOD test (American Public Health Association, 2005.

179   Populations of *E. coli* were determined using the Colilert defined substrate test. Briefly, 97 wells

180   were filled with 100 mL of water sample with the Colilert substrate added. Samples were

181   incubated for 24 hours, and wells that fluoresce under a UV light were considered positive for

182   *E. coli,* and a most probable number estimate was made based on the number of positive wells

183   in both the large and small wells (American Public Health Association, 2005). If a sample was in

184   excess of the geometric mean United States recreational water quality criteria, the sample site

185   was considered impaired. Impairment was based on recommendation 1, which is a threshold of

186   126 $\frac{CFU}{100\ mL}$ that corresponds to an illness rate of $\frac{36}{10000}$ people (US Environmental Protection

187   Agency, 2012).

188

189   To get an estimation of land use throughout the Sinking Creek watershed, land cover data were

190   downloaded from the National Land Cover Dataset (NLCD) for 2006 (Fry et al., 2011). Each

191   sampling site's drainage area was delineated using StreamStats version 3 from the United

192   States Geologic Survey (Ries III et al., 2017). Land was grouped into 3 categories; forested,

193   developed, and agricultural. Forested land includes the categories deciduous forest, evergreen

194   forest, and mixed forest. Developed land use includes all developed categories; open space

195   (less than 20% impervious surface), low intensity (20 – 49% impervious surface), medium

196   intensity (50 – 79% impervious surface), and high intensity (80 – 100% impervious surface).

197   Agricultural land included grassland/herbaceous and pasture/hay. The area of each land use

198   was divided by the total area of the drainage area to get the percentage land use shown in

199   table 1, and sampling sites as well as land cover categories are shown in figure 1.

200   **2.2 Modeling background**

201   Maxent is an iterative machine learning model commonly used for mapping species

202   distributions (Phillips, S., Dudík, M., Schapire, 2010). Within the sample space, x, and given a

203   set of environmental features (parameters), $f_1(x)$, $f_2(x)$, …, $f_n(x)$, the Maxent distribution

204   estimates a vector of feature weights, $\beta = (\beta_1\ \beta_2, …, \beta_n)$, that maximizes the entropy of the raw

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

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

207   Where Z_β is the normalization constant that ensures that $q_\beta(x)$ integrates to one over the study

208   area (Phillips, S., Dudík, M., Schapire, 2010). This modeling approach is justified because it

209   provides the maximum information concerning impairment. From a water quality management

210 standpoint, this approach is beneficial because decision-makers and stake-holders are more

211 concerned with factors associated with impairment rather than compliance when approaching

212 fecal pollution monitoring and management.

213

214 Original features (parameters) can be transformed into quadratic, product, hinge, and

215 threshold feature classes so that complex multivariate responses can be modelled (Phillips and

216 Dudík, 2008), but Maxent incorporates L1 regularization to balance satisfying the constraints

217 on the features while minimizing overfitting.   L1 regularization is not unique to Maxent, and is

218 used in many general linear models (Elith et al., 2011). A regularization parameter λ_j smooths

219 probability distributions, generating sparse solutions and removing unnecessary features; this

220 shrinks weights to balance fit and complexity (Elith et al., 2011). Because of regularization,

221 Maxent fits a penalized maximum likelihood model equivalent to minimizing the relative

222 entropy dependent on the error-bound constraints

223 $max_\beta \frac{1}{m}\Sigma^m_{i=1}\ln(q_\beta(x_i)) - \Sigma^n_{j=1}\lambda_j|\beta_j|$ (2)

224 $subject\ to \int q_\beta(x)dx = 1$

225

226 Where m is the number of positive samples, n is the number of features, and x is the feature

227 vector for occurrence point i. Equation 2 provides insight into how Maxent uses background

228 data: the first term is larger for models that distinguish between impairment states the best.

229 The second term represents the regularization, which gets larger as the weights β_j increase,

230 indicating a complex model more likely to over fit. The output of $q_\beta(x)$ is termed the raw

231 distribution, but it is difficult to interpret due to its scale dependence. More background points

232 result in smaller raw values because their sum cannot exceed 1 over a large amount of points

233 (Phillips and Dudık, 2008; Elith et al., 2011;). For this reason, the logistical output of the Maxent

234 model, $P(x)$, will be used because it represents the probability of impairment given the sample

235 space, x. This is a logistic model using the same set of weights β with the intercept of the model

236 determined by the entropy of $q_\beta(x)$, H. The model is shown in Eq. 3 below.

237 $$P(x) = \frac{e^H q_\beta(x)}{1 + e^H q_\beta(x)}$$
(3)

238 **2.2.1 Univariate models**

239 Data were processed using a list-wise deletion process, where individual samples from a site

240 were removed if they were missing a parameter measurement due to laboratory errors,

241 equipment malfunctions, calibration issues, or sites being dry at the time of sampling. A sample

242 of 100 bootstrapped models were developed, and 20% was subsampled for testing validation.

243 Bootstrapping is a nonparametric resampling technique to make inferences about a population

244 based on resampling from a set population, generating population level statistics, while

245 providing an estimate of uncertainty of those statistics (Campolongo, 1997). For this modeling

246 approach, all background points are used in the development of the null model, and the

247 impaired samples are bootstrapped. Although Maxent can incorporate a wide variety of feature

248 classes, only linear and quadratic feature classes were used to develop action values, or

249 thresholds of impairment.   The rationale for using these types of feature classes is for ease of

250 generating action values as well as to assess both linear and non-linear effects of single

251 parameters.

252

253   The AUC was calculated for the training and testing datasets. The AUC is a metric of

254   performance for binary classification. the true positive prediction rate (sensitivity) and false

255   positive prediction rates (1 – specificity) of each sample are plotted as a ROC for different

256   decision boundaries, and the area under that ROC is integrated. An AUC of 0.5 indicates that

257   the model is no better than random chance, and a value of 1.0 indicates perfect model

258   performance (Zweig and Campbell, 1993; Zou et al., 2007).

259

260   The decision boundary (logistic threshold) between impaired and unimpaired samples should

261   maximize accurately predicting impairment (exceedance of the *E. coli* criteria) while balancing

262   correct negative predictions (Bean et al., 2012). Therefore, maximum test sensitivity and

263   specificity was defined as the appropriate decision boundary. A low sensitivity would indicate

264   poor performance in identifying impairment, while a low specificity would indicate an

265   overcautious model in which resources might be wasted in remediation of false positives.

266   Accuracy for Maxent models was calculated as follows: $\frac{TP + TN}{TP + TN + FP + FN}$, where TP are true

267   positives, TN are true negatives, FP are false positives and FN are false negatives. Significance of

268   the univariate model was determined by calculating the χ^2 statistic for each confusion matrix,

269   with the null hypothesis being that the classifier was no better than random chance.

270

271   Action values (environmental thresholds) are conditions in which a parameter (variable) is at

272   the threshold of impairment, indicating potential exceedance of the *E. coli* standard. Action

273   values were calculated for significant (*p* <0.01) univariate models by averaging bootstrapped

274   weights and estimating the parameter value at which the probability of impairment equals the

275   logistic threshold. Figure 2 demonstrates the concepts of the AUC performance metric,

276   selected decision boundary, and the concept of the action value in relation to the selected

277   decision boundary and Maxent model function (Eq. 3).

278

279   **2.2.2 Multivariate models and sensitivity analysis**

280   Although some authors state that collinearity is not as problematic in Maxent compared to

281   traditional regression approaches, collinearity was explored and subsequently removed using

282   Pearson correlation coefficients (Elith et al., 2011). Variables that were highly correlated ($r >$

283   0.8) were evaluated to determine which variable to include based on expertise, connections to

284   previous models, and accuracy metrics within the analysis. The initial multivariate model

285   included all noncollinear variables and were developed using 100 bootstrapped samples like the

286   univariate models, with the addition of product feature classes to incorporate variable

287   interaction. Average variable contribution was determined by calculating the increase in

288   information gain associated with a change in each feature for each iteration of the model

289   algorithm, normalized to percentages. The permutation importance of a feature is an indicator

290   of variable sensitivity. In each model run, the feature training presence and background data

291   are randomly permutated, and the resulting drop in training AUC is normalized to percentages

292   for each variable and averaged over the 100 bootstrapped runs.

293

294   A jack-knife sensitivity analysis was used to determine the best subset of covariates to include

295   in a trimmed model. Each variable was removed from the analysis, and a comparison was

296   made to determine if the removal of a variable caused a significant ($p<.05$) change in training or

297   testing information gain. Student's *t*-tests were performed on each jack-knifed model to

298   evaluate significance, and variables were included if the information gain from either the

299   testing or training sets decreased; decrease in information would correspond to a significant

300   loss of information, providing criterium for inclusion of the variable in final models.

301   **3.   Results**

302   **3.1 Univariate Model Performance**

303   The sampling program resulted in 29 sampling trips over 14 sites, allowing for a potential of 406

304   samples for analysis. 127 samples were removed due to missing information in the dataset,

305   leaving 279 total samples for model development. This included 95 impairments, identified by

306   exceedance of the *E. coli* recreational water quality standard. Table 1 presents the summary

307   statistics for *E. coli* and the associated land use in each sampling site's drainage area. Each

308   training set included 279 background points, 76 points for training, and 19 points to evaluate

309   performance on testing data.

310

311   Table 2 summarizes the training and testing performance of the univariate Maxent models used

312   to identify *E. coli* impairment based on environmental variables. Water temperature performed

313   best based on AUCs, but had lower accuracy than conductivity, dissolved oxygen, and alkalinity.

314   The plausible explanation of these differences is the latter variables had higher specificity

315   rather than sensitivity at the chosen decision boundary (Table S2). Accuracy was found to be

316   highest for alkalinity and lowest for pH.

317

318   Action values were developed for 8 significant univariate models by solving for the value of the

319   variable when probability of impairment equals the logistic threshold. For example, the action

320   value for alkalinity is 128 mg/L. This means that *E. coli* impairment is likely to occur when

321   alkalinity is observed to be higher than this threshold. Action values and 95% confidence

322   intervals are included in table 2. Action function graphs for each significant univariate model

323   are presented in Figure S1 to aid in interpretation of table 2, and summary statistics for each

324   variable are given in table S1.

325

326   **3.2 Multivariate Model Performance**

327   Pearson correlations ranged from -0.269 to 0.834, with three variables identified as collinear;

328   alkalinity, conductivity, and hardness. Conductivity was selected because of its use in

329   previously developed fecal indicator models (Wilkes et al., 2011; Gonzalez et al., 2012;Gonzalez

330   and Noble, 2014; Piorkowski et al., 2013). The 8-variable model displayed improved accuracy on

331   all univariate models. Variable contribution was dominated by water temperature, conductivity,

332   and discharge in the 8-variable model, with water temperature contributing 36.4% of the

333   information, and conductivity and discharge accounting for 22.6 % and 12.1% of the

334   information, respectively. The permutation importance for the 8-variable model demonstrated

335   a similar pattern. A summary of accuracy metrics is shown in table 2, and table 3 illustrates the

336   contribution of each variable in the multivariate models.

337

338   Conductivity was the most sensitive parameter based on sensitivity analysis, with other

339   sensitive parameters including water temperature, dissolved oxygen, discharge, and NO_3. The

340   removal of dissolved oxygen improved model performance based on testing AUC, justifying

341   development of two optimized multivariate models; a 5-variable model including all sensitive

342   parameters, and a 4-variable model that excluded dissolved oxygen.

343

344   Accuracy of the 5- and 4- variable optimized model was 77.8 %. The patterns of variable

345   contribution were consistent in each model, with water temperature accounting for most of the

346   information gain in each model. Figure 3 shows the variable contribution for the initial

347   multivariate models, each model run during the sensitivity analysis, and the final 4-variable and

348   5-variable models produced. The information gain for each model is also shown within this

349   figure.

350

351   Response surfaces were developed for each of the model runs to assess spatiotemporal trends.

352   Each grid within the surface represents a single sample, with each sampling site representing a

353   single column. The columns are oriented in a downstream fashion, with headwaters sites

354   starting on the left (SC14) and sites further downstream existing on the right (SC1). The

355   temporal scale is represented by the rows, with each row indicating a specific sampling trip.

356   Although the data resolution is coarse, the goal is to demonstrate the potential of visualizing

357   trends in the probability of impairment over space and time. Figure 4 displays the response

358   surface for the estimated probability of impairment for the 4-variable model and the 5- and 8-

359   variable models are shown in Figure S2. Classification performance for the univariate models

360   and multivariate models is shown in table S2. Mean probabilities for the 8-, 5-, and 4-variable

361   model were 0.338 (95%CI: 0.319, 0.358), 0.353 (0.334, 0.373), and 0.359 (0.340, 0.378).

362   Generally, the sites influenced by the greatest amount of developed or agricultural land use

363   (SC5 – SC1) had the highest probability of impairment. August had the highest probability of

364   impairment, followed by May, November, and February. Mean probability of impairment and

365   associated 95% confidence intervals are shown in table S3.

366

367   **4.   Discussion**

368   Over 170,000 miles of U.S. rivers and streams are listed as pathogen impaired based on FIOs. To

369   address these impairments, characterization of sources and transport mechanisms is necessary

370   (United States Environmental Protection Agency, 2017), and statistical models can be used as

371   an inferential tool to overcome these issues. We applied Maxent to identify individual and

372   interacting factors influencing *E. coli* fate and transport that resulted in impairments using

373   univariate and multivariate approaches. In this particular stream, water temperature,

374   conductivity, discharge, and NO_3 were found to be the most influential group of factors driving

375   fecal pollution. The results indicate that seasonality and agricultural run-off are the suggested

376   causes of impairment in this watershed. Seasonality is demonstrated by influence of

377   temperature in the models, whereas the influence of agricultural run-off is suggested by the

378   other variables and the association between land use and *E. coli* in the watershed. Even small

379   increases in agricultural land cause substantial increases in *E. coli* concentrations (Table 1),

380   whereas similar increases in developed land do not have the same pronounced effect. This

381   study highlights the need for multi-month sampling across a stream continuum to truly

382   estimate spatiotemporal variability associated with impairment.

383

384   The fact that water temperature dominated the information in this model suggests that

385   seasonality plays an important role in *E. coli* survival. Although fecal indicators and pathogens

386   have been found to possess diverse temperature-survival relationships (Hofstra, 2011; Sterk et

387   al., 2013), the high August probability for *E. coli* impairment indicates favorable conditions for

388   long-term survival in the summer. Warming due to climate change could exacerbate this

389   condition by increasing those favorable conditions (Weniger et al., 1983; Atherholt et al., 1998;

390   Patz et al., 2000; Guzman Herrador et al., 2015). However, August was not the only month with

391   numerous *E. coli* impairments. Therefore, monitoring for FIOs only in the summer months could

392   distort estimates of impairment in watersheds with year-round users.

393   Although discharge was not predictive of *E. coli* impairment alone, its interactive effect stresses

394   the importance of both flow dependent and independent processes associated with *E. coli*

395   impairment. Dissolved solutes such as NO_3 and ions measured through conductivity are largely

396   discharge-dependent; however, FIOs are not as strongly dependent on discharge. This flow

397   independence is due to additional ecological mechanisms such as nutrient limitation and

398   competition (Surbeck et al., 2006; Drummond et al., 2015). Various forms of nitrogen are

399   associated with increased concentration of FIOs in certain environments (Carrillo et al., 1985;

400   Herrig et al., 2015), and results of the Maxent models suggest that nutrient loading in the form

401   of NO_3 contributes to *E. coli* impairment in Sinking Creek. Other studies have found that

402   dissolved organic carbon can affect magnitude and extent of fecal indicators (Surbeck et al.,

403   2010; Blazewicz et al., 2013; Cloutier et al., 2015), but this was not collected during this

404   sampling program and was found to be insignificant using BOD as a surrogate for organic

405   pollution. This interaction between nutrient levels and fecal pollution highlights the potential

406   for synergistic effects of different sources of pollution, suggesting a limitation of TMDL

407   development when only considering one pollutant at a time.

408   Although machine learning application to microbial water quality problems is not unique, this

409   study presents some beneficial techniques in this area of research. First, it demonstrates the

410   ability to open the black box of Maxent, using action values to predict threshold of impairment

411   based on a single variable. Multivariate action functions can be developed as well, but is not

412   presented in this manuscript. The probabilistic approach to model validation and variable

413   selection allows for inclusion of uncertainty, improving on deterministic methods traditionally

414   used for validation and criteria for variable inclusion. Probabilistic methods have been used in

415   TMDLs (Borsuk et al., 2002), frequency of water quality posting errors (Kim and Grant, 2004),

416   and uncertainty of different fecal indicator methodologies (Gronewold et al., 2008); this paper

417   adds to this framework through identifying the probability of stream impairment given a set of

418   environmental variables. This improves confidence in decision-making for implementation of

419   monitoring, management, and remediation strategies. Modeling microbial water quality is a

420   challenge no matter the method used, but this study demonstrates that Maxent provides a

421   valid approach to understand the factors driving impairment.

422

423   Streams are dynamic systems with multiple flow regimes, confounding an already difficult

424   modeling process. Understanding how models behave in extreme situations is useful for

425   regulation, monitoring, and management of these ecosystems. Over the long-term study

426   periods, samples from both drought and high water conditions were captured. Maxent has

427   been suggested as a strong prediction of extreme values (Petrov et al. 2013), and this study

428   found that Maxent sufficiently predicted impairment during the high flow sampling date of

429   November 11, 2009. Depending on which multivariate model was used, accuracy ranged from

430   72.8 % to 90.9 % for this sampling date. Five sampling dates resulted in at least one site being

431   dry, indicating drought-like conditions. Maxent correctly predicted impairment in these

432   situations 62.2 % to 73.0 % of the time. This suggests that Maxent can be useful for certain

433   extreme situations, but is highly dependent on the environmental variables used for prediction.

434

435   While this study presents proof of concept of using Maxent to infer source and mechanisms of

436   impairment, there are some limitations to this study. Although the dataset has a large time

437   scale (8 years), only collecting from 4 months makes the resolution coarse, reducing the scale at

438   which inferences can be made. The list-wise deletion of samples before univariate modeling

439   removed some data that could inform each of those models; however, using the same series of

440   data in the multivariate models and list-wise deletion are commonly used procedures in

441   statistical models. Future applications of Maxent will improve on the coarse resolution of the

442   data by using monthly and potentially weekly sampling approaches, and research will be

443   developed as to the best approach for handling missing data in Maxent. While AUC scores

444   above 0.70 indicate good model fit, only considering physiochemical water quality parameters

445   limits the potential to accurately predict impairment; however, this study demonstrates that

446   these parameters are informative as a proof of concept for using Maxent as a modeling

447   approach. Future areas of research include using Maxent to optimize water quality monitoring

448   to identify causes of impairment with FIOs and specific pathogens in the most cost-effective

449   way using a variety of microbial, chemical, and physical parameters.

450

451   It is a difficult task to develop and implement remediation strategies in watersheds with many

452   diffuse causes of fecal impairment, but modeling can increase confidence in decision making

453   through inferring mechanisms and sources of fecal pollution. Incorporating environmental

454   variables into models allows for insights into the ecology of fecal indicators, identifying causes

455   of chronic FIO impairment. Although one indicator cannot mimic the plethora of existing

456   pathogens in water, incorporating modeling can fine tune an indicator's utility, ultimately

457   informing the public concerning health risks, and aiding in overcoming the shortcomings of a

458   single indicator monitoring strategy.

459

460   **5. Conclusions**

461

462   Characterizing *E. coli* impairment is essential because of the plethora of streams polluted with

463   fecal wastes. This study used Maxent to identify water quality parameters associated with *E.*

464   *coli* impairment in a low-order, mixed-use watershed. Univariate models generated action

465   values, or thresholds of impairment, based on single parameters, while multivariate models

466   extracted information concerning multivariate interaction. We presented a probabilistic

467 approach to sensitivity analysis, improving confidence in variable selection. Maxent presents a

468 flexible machine learning approach to aid in understanding mechanisms and sources of fecal

469 pollution as well as a host of other complex decision boundary problems. We demonstrated

470 that:

471

472 • Models using alkalinity and water temperature were found to be either the most

473   accurate or best performing univariate models; this stresses the importance of

474   discharge composition and seasonality in *E. coli* impairment. Discharge, however, was

475   not an influential univariate parameters by itself, stressing the importance of flow-

476   independent processes that correlate with impairment.

477 • Sensitivity analysis indicated that the most information was lost when conductivity was

478   removed from the multivariate models, and water temperature, discharge, dissolved

479   oxygen, and NO_3 represent other sensitive parameters sensitive to *E. coli* impairment in

480   this watershed.

481 • Results suggest that *E. coli* impairment in this stream is driven by seasonality and

482   agricultural run-off. This suggests that multi-month sampling along a stream continuum

483   is essential to characterize spatiotemporal variability, importance of flow in relation to

484   other water quality parameters, and the potential synergistic effect of nutrient and fecal

485   pollution.

486 • Incorporating modeling can fine tune an indicator's utility, informing the public

487   concerning human health risks, enhancing our understanding of FIOs, assisting in water

488   quality decision-making, and providing input variables for quantitative microbial risk

489   assessment.

490

491   **Acknowledgements**

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

493

494   **References**

495   Alm, E.W., Burke, J., 2003. Fecal indicator bacteria are abundant in wet sand at freshwater
496   beaches 37, 3978–3982. doi:10.1016/S0043-1354(03)00301-4

497   American Public Health Association, 2005. Standard methods for the examination of water and
498   wastewater, 21st ed. Washington, DC.

499   Anderson, K.L., Whitlock, J.E., Valerie, J., Harwood, V.J., 2005. Persistence and differential
500   survival of fecal indicator bacteria in subtropical waters and sediments. Appl. Environ.
501   Microbiol. doi:10.1128/AEM.71.6.3041

502   Atherholt, T.B., Lechevallier, M.W., Norton, W.D., Rosen, J.S., 1998. Effect of rainfall on Giardia
503   and Cryptosporidium. Am. Water Work. Assoc. 90, 66–80.

504   Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias
505   on threshold selection and accuracy assessment of species distribution models 250–258.
506   doi:10.1111/j.1600-0587.2011.06545.x

507   Belanche-Muñoz, L., Blanch, A.R., 2008. Machine learning methods for microbial source
508   tracking. Environ. Model. Softw. 23, 741–750. doi:10.1016/j.envsoft.2007.09.013

509   Benham, B.L., Baffaut, C., Zeckoski, R.W., Mankin, K.R., Pachepsky, Y.A., Sadeghi, A.M., Brannan,
510   K.M., Soupir, M.L., Habersack, M.J., 2006. Modeling bacteria fate and transport in
511   watersheds to support TMDLs. Trans. ASABE 49, 987–1002.

512   Black, L.E., Brion, G.M., Freitas, S.J., 2007. Multivariate logistic regression for predicting total
513   culturable virus presence at the intake of a potable-water treatment plant : novel
514   application of the atypical coliform / total coliform ratio 73, 3965–3974.
515   doi:10.1128/AEM.02780-06

516   Blazewicz, S.J., Barnard, R.L., Daly, R.A., Firestone, M.K., 2013. Evaluating rRNA as an indicator
517   of microbial activity in environmental communities: limitations and uses. ISME J. 7, 2061–
518   2068. doi:10.1038/ismej.2013.102

519   Blount, Z.D., 2015. The unexhausted potential of E. coli. Elife 4, e05826.
520   doi:10.7554/eLife.05826

521   Boeckmann, M., Joyner, T.A., 2014. Old health risks in new places? An ecological niche model
522   for I. ricinus tick distribution in Europe under a changing climate. Health Place 30, 70–7.
523   doi:10.1016/j.healthplace.2014.08.004

524   Ballestè, E., Bonjoch, X., Belanche, L.A., Blanch, A.R., 2010. Molecular indicators used in the

525   development of predictive models for microbial source tracking. Appl Env. Microbiol 76,
526   1789–1795. doi:10.1128/AEM.02350-09

527   Borsuk, M.E., Stow, C.A., Reckhow, K.H., 2002. Predicting the frequency of water quality
528   standard violations: A probabilistic approach for TMDL development. Environ. Sci. Technol.
529   36.

530   Brion, G.M., Lingireddy, S., 1999. A neural network approach to identifying non-point sources of
531   microbial contamination. Water Res. 33, 3099–3106.

532   Brion, G.M., Neelakantan, T.R., Lingireddy, S., 2002. A neural-network-based classification
533   scheme for sorting sources and ages of fecal contamination in water. Water Res. 36, 3765–
534   3774.

535   Byappanahalli, M.N., Shively, D. a, Nevers, M.B., Sadowsky, M.J., Whitman, R.L., 2003. Growth
536   and survival of Escherichia coli and enterococci populations in the macro-alga Cladophora
537   (Chlorophyta). FEMS Microbiol. Ecol. 46, 203–11. doi:10.1016/S0168-6496(03)00214-9

538   Campolongo, F., 1997. Sensitivity analysis of an environmental model: an application of
539   different analysis methods. Reliab. Eng. Syst. Saf. 57, 49–69. doi:10.1016/S0951-
540   8320(97)00021-5

541   Carrillo, M., Estrada, E., Hazen, T.C., 1985. Survival and enumeration of the fecal indicators
542   Bifidobacterium adolescentis and Escherichia coli in a tropical rain forest watershed. Appl.
543   Environ. Microbiol. 50, 468–476.

544   Characklis, G.W., Dilts, M.J., Simmons, O.D., Likirdopulos, C. a., Krometis, L.A.H., Sobsey, M.D.,
545   2005. Microbial partitioning to settleable particles in stormwater. Water Res. 39, 1773–
546   1782. https://doi.org/10.1016/j.watres.2005.03.004

547   Cloutier, D., Alm, E., McLellan, S., 2015. The influence of land-use, nutrients, and geography on
548   microbial communities and fecal indicator abundance at lake michigan beaches. Appl.
549   Environ. Microbiol. 81, 4904–4913. doi:10.1128/AEM.00233-15

550   Coulliete, A., Money, E.S., Serre, M.L., Noble, R.T., 2009. Space/time analysis of fecal pollution
551   and rainfall in an eastern north carolina estuary. Environ. Sci. Technol. 43, 3728–3735.

552   de Brauwere, A., Ouattara, N.K., Servais, P., 2014. Modeling fecal indicator bacteria
553   concentrations in natural surface waters: a review. Crit. Rev. Environ. Sci. Technol. 44,
554   2380–2453.

555   Drummond, J.D., Davies-Colley, R.J., Stott, R., Sukias, J.P., Nagels, J.W., Sharp, A., Packman, A.I.,
556   2015. Microbial transport, retention, and inactivation in streams: a combined

557   experimental and stochastic modeling approach. Environ. Sci. Technol. 49, 7825–33.
558   doi:10.1021/acs.est.5b01414

559   Dwivedi, D., Mohanty, B.P., Lesikar, B.J., 2013. Estimating Escherichia coli loads in streams
560   based on various physical, chemical, and biological factors. Water Resour. Res. 49, 2896–
561   2906. https://doi.org/10.1002/wrcr.20265

562   Dwivedi, D., Mohanty, B.P., Lesikar, B.J., 2016. Impact of the Linked Surface Water-Soil Water-
563   Groundwater System on Transport of E. coli in the Subsurface. Water, Air, Soil Pollut. 227,
564   351. https://doi.org/10.1007/s11270-016-3053-2

565   Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation
566   of MaxEnt for ecologists. Divers. Distrib. 17, 43–57. doi:10.1111/j.1472-4642.2010.00725.x

567   Ferguson, C., Husman, A.M. de R., Altavilla, N., Deere, D., Ashbolt, N., 2003. fate and transport
568   of surface water pathogens in watersheds. Crit. Rev. Environ. Sci. Technol. 33, 299–361.
569   doi:10.1080/10643380390814497

570   Field, K.G., Samadpour, M., 2007. Fecal source tracking, the indicator paradigm, and managing
571   water quality. Water Res. 41, 3517–3538. doi:10.1016/j.watres.2007.06.056

572   Fry, J.A., Xian, G., Jin, S., Dewitz, J.A., Homer, C.G., Limin, Y., Barnes, C.A., Herold, N.D.,
573   Wickham, J.D., 2011. Completion of the 2006 national land cover database for the
574   conterminous United States. Photogramm. Eng. Remote Sensing 77, 858–864.

575   Gonzalez, R.A., Conn, K.E., Crosswell, J.R., Noble, R.T., 2012. Application of empirical predictive
576   modeling using conventional and alternative fecal indicator bacteria in eastern North
577   Carolina waters. Water Res. 46, 5871–5882.
578   doi:http://dx.doi.org/10.1016/j.watres.2012.07.050

579   Gonzalez, R.A., Noble, R.T., 2014. Comparisons of statistical models to predict fecal indicator
580   bacteria concentrations enumerated by qPCR- and culture-based methods. Water Res. 48,
581   296–305. doi:http://dx.doi.org/10.1016/j.watres.2013.09.038

582   Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann,
583   N.E., 2006. Using niche-based models to improve the sampling of rare species. Conserv.
584   Biol. 20, 501–511.

585   Gronewold, A.D., Borsuk, M.E., Wolpert, R.L., Reckhow, K.H., 2008. An assessment of fecal
586   indicator bacteria-based water quality standards. Environ. Sci. Technol. 42, 4676–4682.

587   Guzman Herrador, B.R., de Blasio, B.F., MacDonald, E., Nichols, G., Sudre, B., Vold, L., Semenza,
588   J.C., Nygård, K., 2015. Analytical studies assessing the association between extreme

589   precipitation or temperature and drinking water-related waterborne infections: a review.
590   Environ. Heal. 14, 29. doi:10.1186/s12940-015-0014-y

591   HACH Company, 2013. DR/890 Colorimeter Procedures Manual.

592   Hach Company, 2006. Digital Titrator - Model 16900: Procedure Manual.

593   Hall, K.K., Evanshen, B.G., Maier, K.J., Scheuerman, P.R., 2014. Application of multivariate
594   statistical methodology to model factors influencing fate and transport of fecal pollution in
595   surface waters. J. Environ. Qual. 43, 358–370. doi:10.2134/jeq2013.05.0190

596   Herrig, I.M., Böer, S.I., Brennholt, N., Manz, W., 2015. Development of multiple linear
597   regression models as predictive tools for fecal indicator concentrations in a stretch of the
598   lower Lahn River, Germany. Water Res. 85, 148–157.
599   doi:http://dx.doi.org/10.1016/j.watres.2015.08.006

600   Hofstra, N., 2011. Quantifying the impact of climate change on enteric waterborne pathogen
601   concentrations in surface water. Curr. Opin. Environ. Sustain. 3, 471–479.
602   doi:http://dx.doi.org/10.1016/j.cosust.2011.10.006

603   Ishii, S., Ksoll, W.B., Hicks, R.E., Sadowsky, M.J., 2006. Presence and growth of naturalized
604   Escherichia coli in temperate soils from Lake Superior watersheds. Appl. Environ.
605   Microbiol. 72, 612–21. doi:10.1128/AEM.72.1.612-621.2006

606   Kistemann, T., Claßen, T., Koch, C., Dangendorf, F., Fischeder, R., Gebel, J., Vacata, V., Exner, M.,
607   2002. Microbial load of drinking water reservoir tributaries during extreme rainfall and
608   runoff. Appl. Environ. Microbiol. 68, 2188–2197. doi: 10.1128/AEM.68.5.2188-2197.2002

609   Kim, J.H., Grant, S.B., 2004. Public Mis-Notification of Coastal Water Quality: A Probabilistic
610   Evaluation of Posting Errors at Huntington Beach, California. Environ. Sci. Technol. 38,
611   2497–2504. https://doi.org/10.1021/es034382v

612   Jamieson, R., Gordon, R., Joy, D., Lee, H., 2004. Assessing microbial pollution of rural surface
613   waters: A review of current watershed scale modeling approaches. Agric. Water Manag.
614   70, 1–17. doi:10.1016/j.agwat.2004.05.006

615   LaLiberte, P., Grimes, D.J., 1982. Survival of Escherichia coli in lake bottom sediment. Appl.
616   Environ. Microbiol. 43, 623–628.

617   Lasalde, C., Rodriguez, R., Toranzos, G. a, Smith, H.H., 2005. Heterogeneity of uidA gene in
618   environmental Escherichia coli populations. J. Water Health 3, 297–304.

619   Lozier, J.D., Aniello, P., Hickerson, M.J., 2009. Predicting the distribution of Sasquatch in
620   western North America : anything goes with ecological niche modelling. J. Biogeogr. 1–5.

621   doi:10.1111/j.1365-2699.2009.02152.x

622   Lipp, E.K., Kurz, R., Vincent, R., Rodriguez-Palacios, C., Farrah, S.R., Rose, J.B., 2001. The effects
623   of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a
624   subtropical estuary. Estuaries 24, 266–276. https://doi.org/10.2307/1352950

625   Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., 2011. Genome sequencing of
626   environmental Escherichia coli expands understanding of the ecology and speciation of the
627   model bacterial species, in: Proceedings of the National Academy of Sciences.
628   doi:10.1073/pnas.1015622108

629   Lušić, D.V., Kranjčević, L., Maćešić, S., Lušić, D., Jozić, S., Linšak, Ž., Bilajac, L., Grbčić, L., Bilajac,
630   N., 2017. Temporal variations analyses and predictive modeling of microbiological
631   seawater quality. Water Res. 119, 160–170.
632   doi:https://doi.org/10.1016/j.watres.2017.04.046

633   Maier, R.M., Pepper, I.L., Gerba, C.P., 2009. Environmental Microbiology, 2nd ed. Elsevier, New
634   York.

635   McCambridge, J., McMeekin, T.A., 1980. Relative effects of bacterial and protozoan predators
636   on survival of Escherichia coli in estuarine water samples. Appl. Environ. Microbiol. 40,
637   907–911.

638   McKergow, L.A., Davies-Colley, R.J., 2010. Stormflow dynamics and loads of Escherichia coli in a
639   large mixed land use catchment. Hydrol. Process. An Int. J. 24, 276–289.

640   McLellan, S.L., Eren, A.M., 2014. Discovering new indicators of fecal pollution. Trends Microbiol.
641   22, 697–706. doi:10.1016/j.tim.2014.08.002

642   Molina, M., Hunter, S., Cyterski, M., Peed, L.A., Kelty, C.A., Sivaganesan, M., Mooney, T., Prieto,
643   L., Shanks, O.C., 2014. Factors affecting the presence of human-associated and fecal
644   indicator real-time quantitative PCR genetic markers in urban-impacted recreational
645   beaches. Water Res. 64, 196–208. doi:https://doi.org/10.1016/j.watres.2014.06.036

646   Money, E.S., Carter, G.P., Serre, M.L., 2009. Modern space/time geostatistics using river
647   distances: data integration of turbidity and E.coli measurements to assess fecal
648   contamination along the Raritan River in New Jersey. Environ. Sci. Technol. 43, 3736–3742.

649   Patz, J.A., Mcgeehin, M.A., Bernard, S.M., Ebi, K.L., Epstein, P.R., Gubler, D.J., Reiter, P., Romieu,
650   I., Rose, J.B., Samet, J.M., 2000. The potential health impacts of climate variability and
651   change for the United States: Executive summary of the report of the health sector of the
652   U.S. National Assessment. Env. Heal. Perspect 108, 367–376.

653   Perkins, T.L., Perrow, K., Rajko-Nenow, P., Jago, C.F., Jones, D.L., Malham, S.K., McDonald, J.E.,
654   2016. Decay rates of faecal indicator bacteria from sewage and ovine faeces in brackish
655   and freshwater microcosms with contrasting suspended particulate matter concentrations.
656   Sci. Total Environ. 572, 1645–1652.
657   https://doi.org/https://doi.org/10.1016/j.scitotenv.2016.03.076

658   Petrov, V., Guedes Soares, C., Gotovac, H., 2013. Prediction of extreme significant wave heights
659   using maximum entropy. Coast. Eng. 74, 1–10.
660   https://doi.org/https://doi.org/10.1016/j.coastaleng.2012.11.009

661   Phillips, S.J., Dudık, M., 2008. Modeling of species distributions with Maxent : new extensions
662   and a comprehensive evaluation 161–175. doi:10.1111/j.2007.0906-7590.05203.x

663   Phillips, S.J., Dudík, M., Schapire, R.E., 2004. A maximum entropy approach to species
664   distribution modeling, in: Proceedings of the Twenty-First International Conference on
665   Machine Learning. ACM, p. 83.

666   Phillips, S., Dudík, M., Schapire, R., 2010. Maxent Software, version 3.3.3k.

667   Piorkowski, G., Jamieson, R., Bezanson, G., Truelstrup, L., Yost, C., 2013. Evaluation of statistical
668   models for predicting Escherichia coli particle attachment in fluvial systems. Water Res. 47,
669   6701–6711. doi:10.1016/j.watres.2013.09.003

670   Ries III, K.G., Newson, J.K., Smith, M.J., Guthrie, J.D., Steeves, P.A., Haluska, T.L., Kolb, K.R.,
671   Thompson, R.F., Santoro, R.D., Vraga, H.W., 2017. StreamStats, version 4, USGS Fact Sheet.
672   Reston, VA. doi:10.3133/fs20173046

673   Savichtcheva, O., Okabe, S., 2006. Alternative indicators of fecal pollution: Relations with
674   pathogens and conventional indicators, current methodologies for direct pathogen
675   monitoring and future application perspectives. Water Res. 40, 2463–2476.
676   doi:10.1016/j.watres.2006.04.040

677   Sinton, L.W., Hall, C.H., Lynch, P.A., Davies-Colley, R.J., 2002. Sunlight inactivation of fecal
678   indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and
679   saline waters. Appl. Environ. Microbiol. 68, 1122–1131.

680   Smith, A., Sterba-Boatwright, B., Mott, J., 2010. Novel application of a statistical technique,
681   Random Forests, in a bacterial source tracking study. Water Res. 44, 4067–4076.
682   doi:https://doi.org/10.1016/j.watres.2010.05.019

683   Sterk, A., Schijven, J., de Nijs, T., de Roda Husman, A.M., 2013. Direct and indirect effects of
684   climate change on the risk of infection by water-transmitted pathogens. Environ. Sci.
685   Technol. 47, 12648–12660.

686 Stott, R., Davies-Colley, R., Nagels, J., Donnison, A., Ross, C., Muirhead, R., 2011. Differential
687   behaviour of Escherichia coli and Campylobacter spp. in a stream draining dairy pasture. J.
688   Water Health 9, 59–69. doi:10.2166/wh.2010.061

689 Surbeck, C.Q., Jiang, S.C., Ahn, J.H., Grant, S.B., 2006. Flow fingerprinting fecal pollution and
690   suspended solids in stormwater runoff from an urban coastal watershed. Environ. Sci.
691   Technol. 40, 4435–4441. doi:10.1021/es060701h

692 Surbeck, C.Q., Jiang, S.C., Grant, S.B., 2010. Ecological control of fecal indicator bacteria in an
693   urban stream. Environ. Sci. Technol. 44, 631–637. doi:10.1021/es903496m

694 Tennessee Department of Environmental and Conservation, 2006. Proposed total maximum
695   daily load (tmdl) for E. Coli in the Water River Watershed (HUC 06010103).

696 Thuiller, W., Richardson, D.M., PYŠEK, P., Midgley, G.F., Hughes, G.O., Rouget, M., 2005. Niche
697   based modelling as a tool for predicting the risk of alien plant invasions at a global scale.
698   Glob. Chang. Biol. 11, 2234–2250.

699 Tyrrel, S.F., Quinton, J.N., 2003. Overland flow transport of pathogens from agricultural land
700   receiving faecal wastes. J. Appl. Microbiol. 94, 87–93. doi:10.1046/j.1365-2672.94.s1.10.x

701 United States Environmental Protection Agency, 2017. National summary of impaired waters
702   and TMDL information [WWW Document]. URL
703   http://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T (accessed
704   8.1.17).

705 US Environmental Protection Agency, 2012. Recreational Water Quality Criteria.

706 Viau, E.J., Goodwin, K.D., Yamahara, K.M., Layton, B.A., Sassoubre, L.M., Burns, S.L., Tong, H.-I.,
707   Wong, S.H.C., Lu, Y., Boehm, A.B., 2011. Bacterial pathogens in Hawaiian coastal streams—
708   Associations with fecal indicators, land cover, and water quality. Water Res. 45, 3279–
709   3290. doi:https://doi.org/10.1016/j.watres.2011.03.033

710 Vidon, P., Campbell, M.A., Gray, M., 2008. Unrestricted cattle access to streams and water
711   quality in till landscape of the Midwest. Agric. water Manag. 95, 322–330.

712 Wade, T.J., Pai, N., Eisenberg, J.N.S., Colford, J.M., 2003. Do U.S. Environmental Protection
713   Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A
714   systematic review and meta-analysis. Environ. Health Perspect. 111, 1102–1109.

715 Weniger, B.G., Blaser, M.J., Gedrose, J., Lippy, E.C., Juranek, D.D., 1983. An outbreak of
716   waterborne giardiasis associated with heavy water runoff due to warm weather and
717   volcanic ashfall. Am. J. Public Health 73, 868–872. doi:10.2105/AJPH.73.8.868

718 Wilkes, G., Edge, T.A., Gannon, V.P.J., Jokinen, C., Lyautey, E., Neumann, N.F., Ruecker, N.,
719   Scott, A., Sunohara, M., Topp, E., Lapen, D.R., 2011. Associations among pathogenic
720   bacteria, parasites, and environmental and land use factors in multiple mixed-use
721   watersheds. Water Res. 45, 5807–5825. doi:10.1016/j.watres.2011.06.021

722 Winfield, M.D., Groisman, E.A., 2003. Role of nonhost environments in the lifestyles of
723   Salmonella and Escherichia coli. Appl. Environ. Microbiol. 69, 3687–3694.
724   doi:10.1128/AEM.69.7.3687

725 Yates, M. V, 2007. Classical indicators in the 21st century—far and beyond the coliform. Water
726   Environ. Res. 79, 279–286.

727 Zou, K.H., O'Malley, A.J., Mauri, L., 2007. Receiver-operating characteristic analysis for
728   evaluating diagnostic tests and predictive models. Circulation 115, 654–657.

729 Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental
730   evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.

731

732

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.

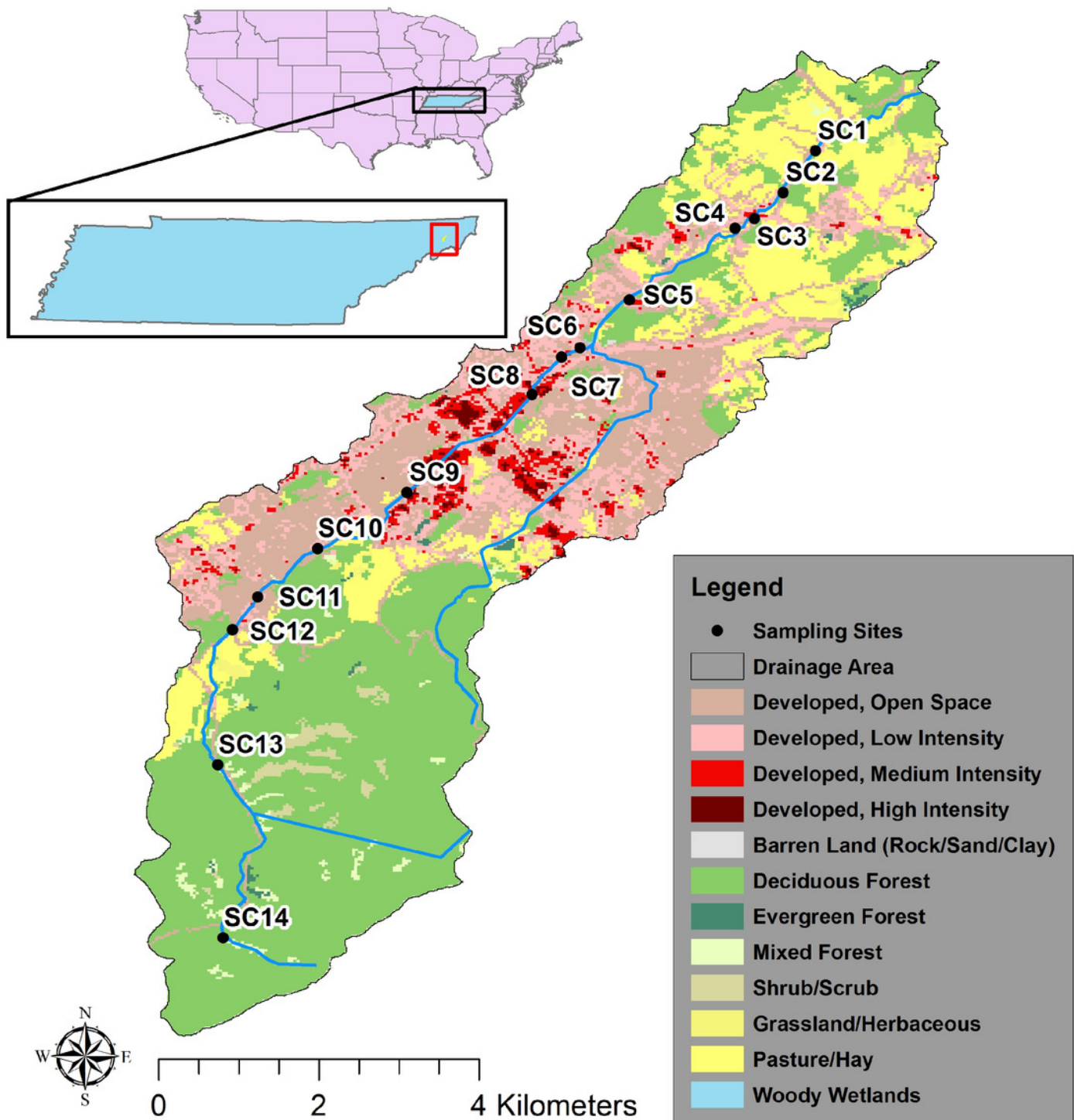



Figure 2

Theoretical plots to illustrate the concept of the ROC, decision boundaries, and action values.

a) Plot of a 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.

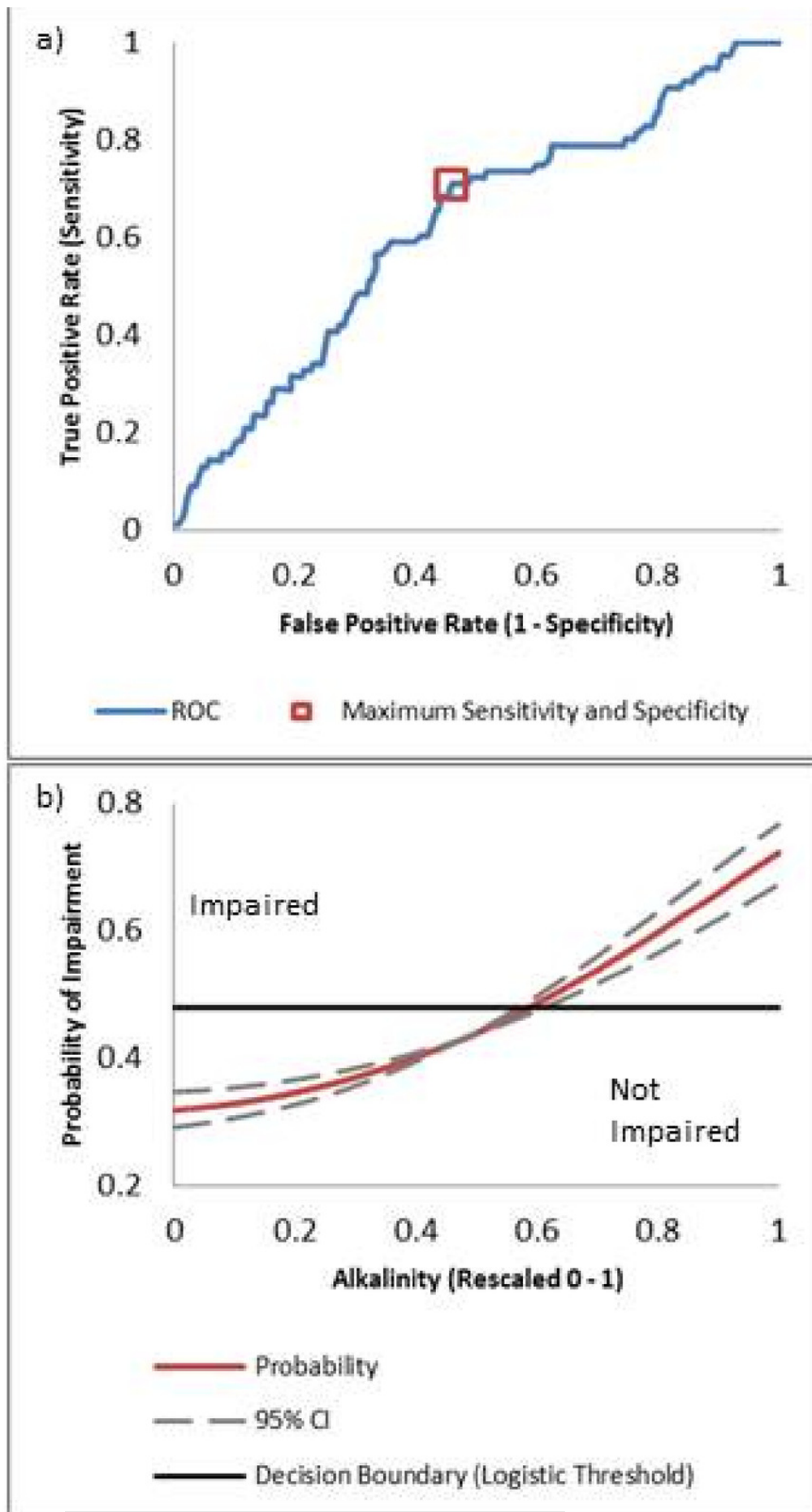



Figure 3

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.

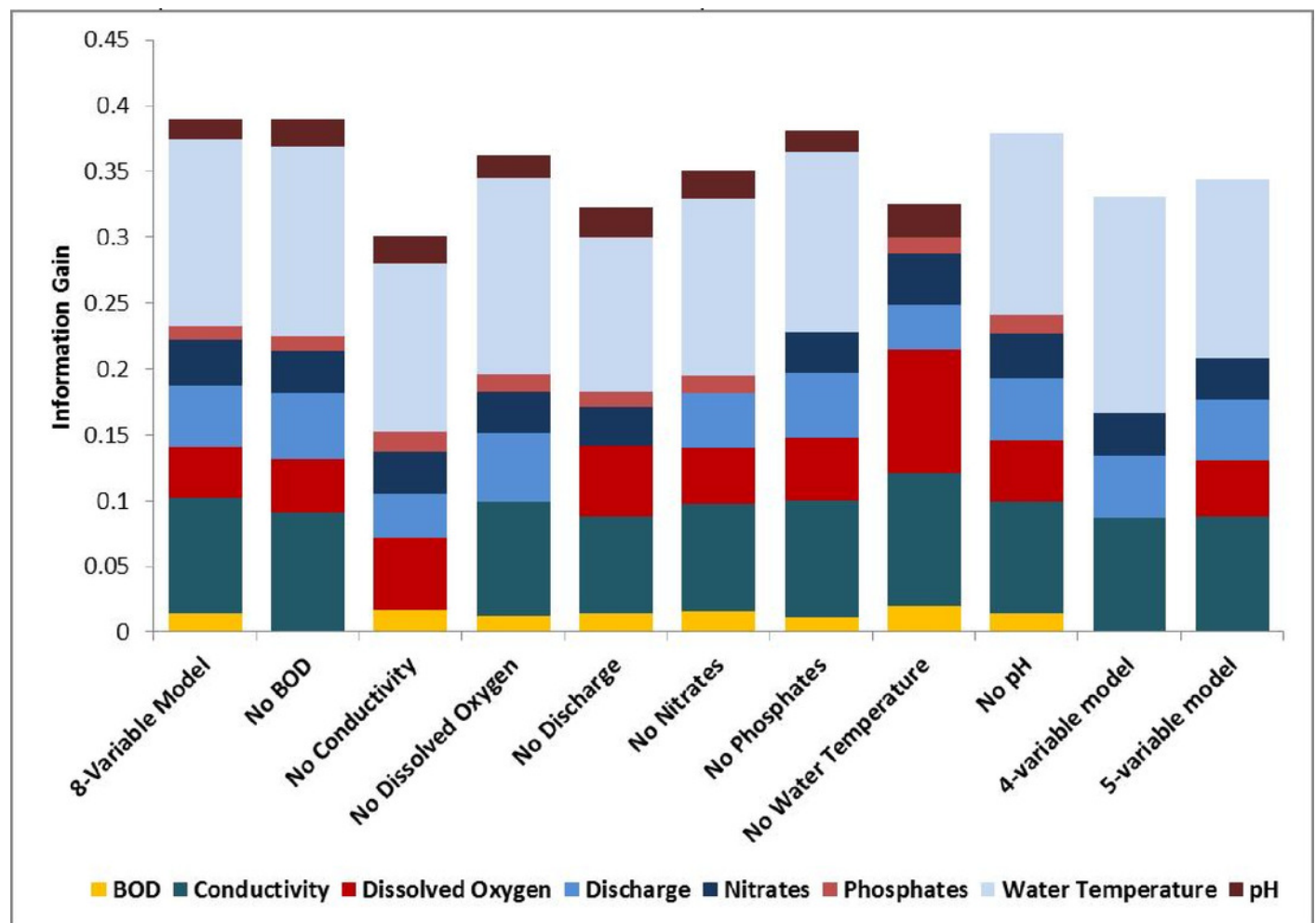
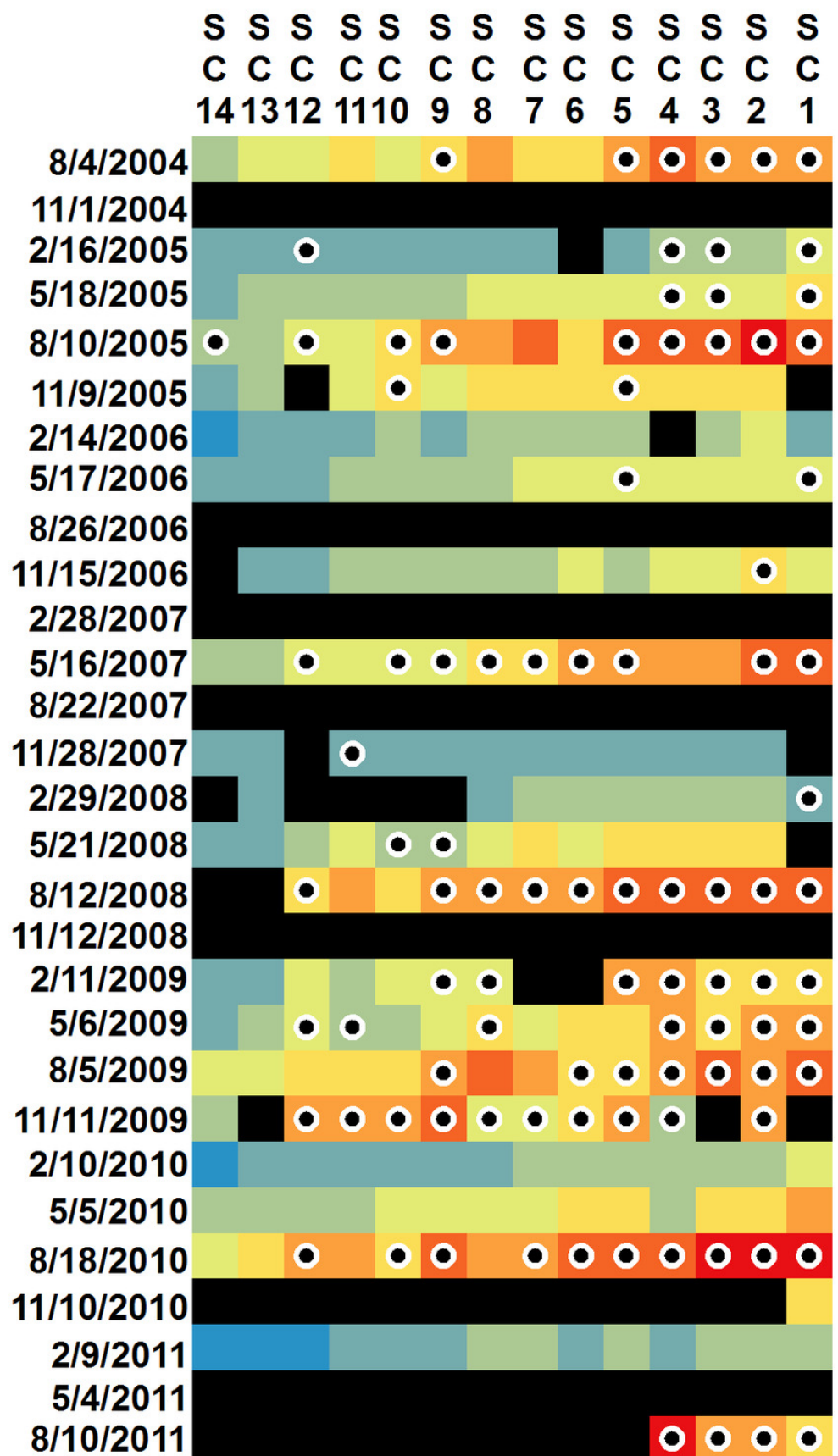



Figure 4

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, and while circles with black centers represent samples in which a stream would be identified as impaired in the study.

E. coli **Probability of Impairment**

0   0.4   0.8

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.

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 (%)	E. coli 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

Summary of training and testing performance of Maxent models.

Performance is based on AUC metrics, accuracy is based on the 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. *Model was not significant #Values of the variables that corresponded to impairment $ 95% CI for the lower bound of the action value &95% CI for the upper bound of the action value

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. *Model was not significant
#Values of the variables that corresponded to impairment
$ 95% CI for the lower bound of the action value
&95% CI for the upper bound of the action value

Variables	Training AUC (+/-SE)	Testing AUC (+/-SE)	Accuracy	Action Values (x) ¥ (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 (μS)	0.628 (0.003)	0.638 (0.006)	65.6	x>306 μS (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$_3$ ($\frac{mg}{L}$)	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	*

$PO_4\ (\frac{mg}{L})$	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) ⌣ (6.27,9.01) ⌢
Water Temperature (^{O}C)	0.666 (0.003)	0.670 (0.005)	65.2	x>12.4 ^{O}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	

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

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_3	9.5	8.5	8.9	8.6	8.9	10.3
pH					3.9	1.7
PO_4					2.7	2.5
Water Temperature	49.9	46.5	36.4	33.7	39.7	34.0