ISEIS

Journal of Environmental Informatics 26(1) 1-13 (2015)

Journal of Environmental Informatics

www.iseis.org/jei

Use of Environmental Parameters to Model Pathogenic Vibrios in Chesapeake Bay

E. A. Urquhart[1,*], B. F. Zaitchik[1], S. D. Guikema[2], B. J. Haley[3], E. Taviani[3], A. Chen[3], M. E. Brown[4],
A. Huq[3], and R. R. Colwell[5]

[1] *Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218, USA*
[2] *Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*
[3] *Maryland Pathogen Research Institute, Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA*
[4] *Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[5] *Center for Bioinformatics and Computational Biology, University of Maryland Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742, USA*

Received 9 September 2013; revised 1 September 2014; accepted 18 September 2014; published online 23 July 2015

ABSTRACT. Annual reports show that human infections caused by *Vibrio* spp. have nearly doubled over the past decade in the Virginia and Maryland waters of the Chesapeake Bay. *Vibrio* spp. are autochthonous to estuarine and coastal waters and follow a seasonal cycle attributed mainly to fluctuations in water temperature and salinity. This study presents the development of empirical algorithms for predicting the probability of *Vibrio vulnificus* and *Vibrio parahaemolyticus* likelihood and abundance in the upper Chesapeake Bay. To model likelihood of occurrence, a set of binary classification models was developed, employing a suite of geophysical predictor variables and statistical methods. Accuracy of results was ~ 68% at 0.40 prediction for *V. vulnificus* and ~ 70% at 0.60 prediction for *V. parahaemolyticus*. For *Vibrio* spp. abundance, regression methods were applied to samples positive for *Vibrio*, showing *Vibrio* abundance can be predicted as a function of sea surface temperature and salinity in Chesapeake Bay, with mean absolute error (MAE) of 3.9 cells 10 ml^{-1} for *V. vulnificus* and 5.8 cells 10 ml^{-1} for *V. parahaemolyticus*. Additionally, for the purpose of operational potential in the Chesapeake Bay, we developed a two-step classification/regression hybrid approach was used to generate estimates of abundance in the absence of bacteriological data on presence of *Vibrio* spp. This hybrid approach predicted *Vibrio* abundance with MAE of 2.8 cells 10 ml^{-1} for *V. vulnificus* and 4.4 cells 10^{-1} ml for *V. parahaemolyticus*. Since the risk of human infection is a function of *Vibrio* spp. pathogenicity and abundance, extending available predictive modeling capabilities to provide concentration, in addition to presence/absence, advances the public health utility of these models significantly.

Keywords: quantitative colony bot hybridization, hybrid modeling, classification, regression, generalized additive model, random forest model

1. Introduction

The microbiology of the Chesapeake Bay includes many species of the family *Vibrionaceae,* some of which are pathogenic to humans and marine animals (Colwell et al., 1977; Hoge et al., 1989; Wright et al., 1996). Cases of human infection are infrequent, but reports from local health departments and the Centers for Disease Control and Prevention indicate the annual number of reported human *Vibrio* infections in the Bay region has nearly doubled in the past decade (Maryland Department of Health and Mental Hygiene, 2013; Virginia Department of Health, 2013). Furthermore, *Vibrio* spp. are frequently detected in oysters and other shellfish harvested for human consumption during the summer months (Constantin de Magny et al., 2009). This seasonality correlates with peak incidence of vibrioses. Soft tissue infections, gastroenteritis, and primary septicemia following consumption of contaminated seafood or exposure to the marine environment are the most common manifestations of *V. vulnificus* disease in humans (Howard and Bennett, 1993; Wright et al., 1996; Strom and Paranjpye, 2000). *V. parahaemolyticus* is an invasive bacterium that typically causes severe diarrhea, but can also cause skin infections if wounds are exposed to seawater or contact with shellfish or crustaceans (Howard and Bennett, 1993; Centers for Disease Control and Prevention, 2013).

Despite the fact that *Vibrio* spp. are known pathogens of global occurrence, the environmental conditions associated with risk of *Vibrio* infection are poorly characterized, with no scientific consensus on the effect of climate change on *Vibrio* populations or risk of *Vibrio* infection. A recent study by Urquhart et al. (2014) examined *V. vulnificus* model sensitivity to climatic variability and change within the upper Chesapeake Bay by assessing model response to a range of temperature and salinity values. The predicted response of *V. vulnificus* proba-

[*] Corresponding author. Tel.: +1 603 8622250; fax: +1 6038621101
E-mail address: erin.urquhart@unh.edu (E. A. Urquhart).

ISSN: 1726-2135 print/1684-8799 online
doi: 10.3808/jei.201500307

bility to high temperatures in the Bay differed systematically between models of differing structure, indicating that the impact of climatic change on the probability of *V. vulnificus* presence in the Chesapeake Bay remains uncertain (Urquhart et al., 2014). Development of regionally customized models for monitoring and predicting risk can empower public health authorities in risk management and controlling vibrioses under evolving climate conditions.

In the Chesapeake Bay, where *Vibrio* spp. are an increasing public health concern, many studies (Kaneko and Colwell, 1973, 1974; Colwell et al., 1977; Kaper et al., 1981; Wright et al., 1996; Parveen et al., 2008) have documented the relationship between *V. vulnificus* and *V. parahaemolyticus* and environmental parameters. In general, abundance of *Vibrio* spp. is greatest when the temperature is greater than 15 °C, with salinity between 5 and 25 ‰, and optimal conditions varying by species and region. Temperature and salinity requirements for growth of *Vibrio* spp. have been shown to be related to the seasonal *Vibrio* cycles in coastal and estuarine environments (Kaper et al., 1981; Motes et al., 1998; Lipp et al., 2001; Jacobs et al., 2010).

Other environmental variables can also influence the abundance and distribution at seasonal and subseasonal scales. Yamazaki and Nwadiuto (2012), showed a positive correlation between the concentration of *Vibrio* spp. in coastal waters off the southeast coast of Florida and rainfall, concluding that the decrease in salinity, increased eutrophication, and increased turbidity from terrestrial runoff after rain events were responsible for the observed increase.

Environmental parameters related to the abundance of *Vibrio* spp. and plankton in Chesapeake Bay have been studied extensively (Wright et al., 1996; Louis et al., 2003; Constantin de Magny et al., 2009; Jacobs et al., 2010; Parveen et al., 2013). With the goal of modeling the presence of *V. cholerae* as a function of environmental factors in the Chesapeake Bay, Louis et al. (2003) developed an empirical habitat model using logistic regression and a binary classification tree. They showed variations in sea surface temperature and salinity contribute to variability in both frequency of bacterial occurrence and geographic distribution of *V. cholerae*. Wright et al. (1996) and Jacobs et al. (2010) developed similar predictive models for presence of *V. vulnificus*, using *in situ* temperature, salinity, and sampling depth data and logistic regression analysis (Wright et al., 1996) in the Bay. Parveen et al. (2013) developed a predictive model, using temperature, salinity, harvest season, and region on the growth rate of *V. parahaemolyticus* in oysters in the Chesapeake Bay.

Long-term hindcasts and forecasts from predictive models of *Vibrio* spp. can be useful in understanding how land-use and climate change impact the frequency, distribution, and magnitude of bacteria in the Chesapeake Bay. The information can then be applied to long-term projections of *Vibrio* spp. in the Bay.

Satellite remote sensing, interpolated-satellite (Urquhart et al., 2013), and simulated hydrodynamic model data can be used to achieve temporal and spatial *Vibrio* spp. predictions for the Bay. In fact, a previous study by Constantin de Magny et al. (2009) successfully generated spatially complete predictions of *V. cholerae* likelihood that was based on simulated sea surface temperature and salinity from the numeric model Chesapeake Bay Regional Ocean Modeling System (ChesROMS; (Xu et al., 2012). Hindcast prediction, distribution, and potential hotspot of occurrence of *V. vulnificus* in the Chesapeake Bay has been reported by using a multivariate habitat suitability model stimulated by sea surface temperature and salinity during a period of 1991 and 2005 (Banakar et al., 2011). Banakar et al. (2011) concluded that hindcast prediction should be useful for further understanding of the impact of environmental conditions in the occurrence of *V. vulnificus* and long-term projections of *Vibrio* spp. in the Chesapeake Bay. Thus, satellite and *in situ* observations can be combined in a dynamical model with data assimilation so that observations when available are utilized, and the model dynamics drive forecasts in the absence of observations. Furthermore, a data assimilation system, using ChesROMS, has recently been developed (Hoffman et al., 2012) for the Chesapeake Bay.

Here we present empirical algorithms for predicting the probability of *Vibrio* spp. incidence and abundance in the upper Chesapeake Bay, which represent an advance over existing models in two respects. First, a model for *V. parahaemolyticus* presence and concentration in Chesapeake Bay is provided. Second, concentration of *Vibrio* spp. in areas where they are present can be obtained. Since the risk of human infection is a function of *Vibrio* concentration, extending available predictive models to provide concentration, in addition to presence/absence, advances the public health utility of the models significantly.

Methodologically, this study contributes to environmentally-based pathogen prediction by incorporating a range of statistical modeling options. Most ecological forecasting models rely on a single model structure, usually linear regression. In contrast the current study tests three types of empirical models: Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Random Forest Model (RF). In using the three models, we have taken a multistep approach: first, binary classification is used to model whether or not bacteria are present; second, regression of positive count data is used to estimate bacterial abundance; third, the methods are combined using hybrid classification-regression, estimating total bacterial abundance in a given geographic area predicted to have *Vibrio* spp. present. Thus, the main objectives of this study were to develop a *Vibrio* spp. empirical algorithm capable of producing likelihood of presence maps and develop a *Vibrio* spp. algorithm that estimates bacterial abundance in a given geographical area of the Chesapeake Bay.

2. Materials and Methods

2.1. Sample Collection

Water samples were collected during July and September, 2011, and March through June, 2012, at sites located in Chesapeake Bay (See Figure 1). The Maryland Department of Natural Resources and the NASA GEO-CAPE Field Campaign research vessels were used in the sampling, with surface water samples (0.5 ~ 1 m depth) collected using a combination of flow-through collection systems and overboard bucket sampling. For the latter, sterile polypropylene bottles (1 L) were rinsed, filled, and placed on ice for transport to the laboratory within 1 hour of collection. Surface temperature and salinity were measured at the time of collection of water samples using an YSI Series 6 instrument (Yellow Springs, Ohio). A total of 148 surface water samples were collected for bacteriological analysis that was carried out within 12 hours at the Maryland Pathogen Institute located at the University of Maryland, College Park.

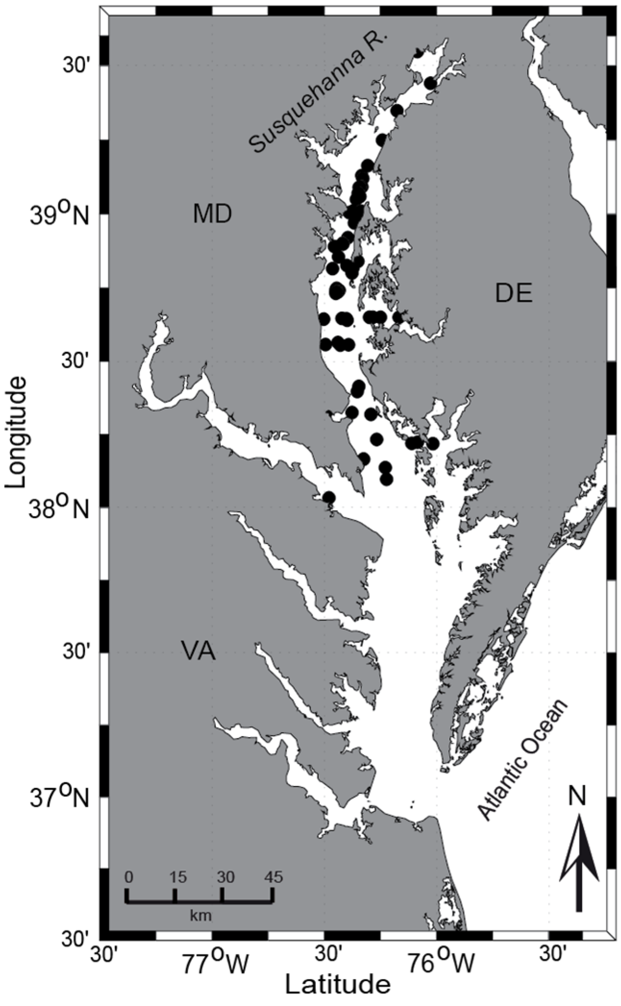

Figure 1. Map of Chesapeake Bay and its tributaries: dark circles represent the sampling stations for this study.

2.2. Laboratory Sample Processing

2.2.1. DNA Extraction and Qualitative Direct PCR

Water samples were shaken and 100 ml passed through a 0.22 μm sterile polycarbonate membrane, then placed in 5 ml of sterile 1X PBS and vortexed. A 1 ml aliquot was removed and boiled for 10 minutes, and iced for 10 minutes before centrifuging at 13,000 rpm for 10 minutes. The supernatant was transferred to a sterile microcentrifuge tube and stored at 20 °C until *toxR* multiplex PCR was employed for qualitative detection of *V. vulnificus*, and *V. parahaemolyticus* (Bauer and Rorvik, 2007). Results were visualized on 1% agarose gel stained with ethidium bromide.

2.2.2. Quantitative Colony Blot Hybridization

To quantify culturable *V. parahaemolyticus* and *V. vulnificus* plates, 1 ml water samples were spread in duplicate onto T1N3 agar and *Vibrio vulnificus* agar (VVA) plates, respectively, and the plates were incubated overnight at 37 °C. Colonies were lifted onto Whatman #541 filters and species-specific probe hybridization was done (DePaola et al., 1997; McCarthy et al., 1999).

2.3. Statistical Model

Three statistical modeling methods were used, Generalized Linear Modeling (GLM), Generalized Additive Modeling (GAM) and Random Forest models (RF) to predict three characteristics of *Vibrio* spp. distribution, namely probability of presence (hereafter: "LIKELIHOOD"), abundance at sites with confirmed presence (hereafter: "ABUNDANCE"), and abundance at all sites in the absence of prior bacteriological data on presence (hereafter: "HYBRID," because it requires a two-step classification/regression approach). ABUNDANCE models assume perfect prior information on presence/absence and were included to determine how models would perform in addressing presence versus absence and quantitative prediction of bacterial abundance. The HYBRID models provide prediction and offer realistic operational potential.

All statistical computations were carried out using R Statistical Package 2.14 on an Intel Xeon W3580 Processor, 3.33 GHz machine with 12 GB RAM. Computation time for all likelihood statistical models within the holdout validation test was less than three minutes.

2.3.1. Statistical Methods

The GLM, GAM, and RF modeling methods were used to develop LIKELIHOOD, ABUNDANCE, and HYBRID models. For LIKELIHOOD models, each method was implemented in logistic form and trained using observational data transformed to binary presence/absence: cell count > 0 cells 10 ml^{-1} ≡ presence, cell count = 0 cells 10 ml^{-1} ≡ absence. For ABUNDANCE models, cell count was predicted as a continuous variable. The ABUNDANCE models were developed using data only from samples with cell counts > 0, and a log link function was applied in GAM and GLM, using a Poisson likelihood function. HYBRID modeling was carried out using a two-step technique described by Guikema and Quiring (2012): (1) binary classification based on the best LIKELIHOOD model, (2) concentration prediction based on the best ABUNDANCE model.

2.3.1.1. Generalized Linear Model (GLM)

The Generalized Linear Model is an extension of the Ordinary Least Squares (OLS) linear model that allows for non-Gaussian probability distributions and the use of both continuous and count data (Nelder and Wedderburn, 1972; Fox, 2008). GLM achieves flexibility by including a link function that relates linear predictor to a function of the explanatory variables (Cameron and Trivedi, 2013). For binary data, one such function is the "logit" link function and it transforms expectation of response to the linear predictor:

$$log\,[p\,/\,(1-p)] = \beta_0 + \Sigma_j \beta_j x_j \tag{1}$$

where $p\,/\,(1-p)$ is the odds ratio of *Vibrio* spp. presence, β_0 is the intercept, β_j is the regression coefficient for variable x. Furthermore, solving for p, the probability of *Vibrio* presence is then:

$$P_{\text{presence}} = e^{(logit)}/[e^{(logit)} + 1] \tag{2}$$

The GLM algorithm was implemented by the *stats* (version 2.14.0) R package (Hastie and Pregibon, 1992).

2.3.1.2. Generalized Additive Model (GAM)

A Generalized Additive Model extends GLM by allowing for nonlinear relationships between explanatory variables and response variable (Hastie and Tibshirani, 1990). This is achieved by replacing the linear predictor $\alpha + \Sigma_j \beta_j x_j$ of a GLM with an additive predictor $\alpha + \Sigma_j f_j(x_j)$ where $f_j(x_j)$ is a non-parametric smoothing function. The smoothing function provides information about the relationship between explanatory variables and response variable not revealed using a traditional linear model (Hastie and Tibshirani, 1986). For this study, the standard smoothing approach, a cubic regression spline, was used. Again, for bacterial presence data, the "logit" link function was used to establish the relationship between response variable and smoothed function of the explanatory variables. The GAM algorithm was implemented by the *mgcv* (version 1.7-16) R package (Wood, 2006).

2.3.1.3. Random Forest (RF) Model

A Random Forest model is an algorithm that fits many classification trees to a dataset, and then uses an ensemble of tree-structure predictions (Breiman, 2001). The algorithm begins with selection of *n* bootstrapped samples (e.g., 500) with replacement from the original dataset. Observations from the original dataset not included in the bootstrap sample are referred to as out-of-bag (OOB) sample, and are used in model cross-validation. A classification tree is fit to each bootstrap sample. To ensure that each of the trees in the ensemble is independent, each tree uses a small number (*m*) of randomly selected predictor variables for split construction at each node. The trees are fully grown and each individual tree is used to estimate the OOB sample. The predicted class is calculated by a majority vote of the OOB predictions for that sample. The RF algorithm in this study was implemented by the *randomForest* (version 4.6.-6) R package (Liaw and Wiener, 2002).

2.3.2. Model Evaluation

2.3.2.1. LIKELIHOOD Model Validation

Predictions from the LIKELIHOOD models come in the form of probabilities, such that a probability threshold or prediction point is needed to transform probability into bacterial presence/absence data. A prediction point is also required to assess model performance using various indices derived from a confusion matrix. Rather than subjectively setting probability to an arbitrary value of 0.50 (50%), which has no ecological basis, the threshold was selected empirically to maximize agreement between observed and predicted distributions in the out of bag data. To ensure correct binary classification, we optimized this prediction point relative to four model assessment indices: true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and accuracy (ACC). In addition, area under the curve (AUC) was calculated for each threshold probability. The indices listed above require information from the confusion matrix, which consists of four elements: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The indices used to assess the predictive performance of the various LIKELIHOOD models are described below:

$$TPR = TP\,/\,(TP + FN) \tag{3}$$

where true positives represent bacterial presence predictions and false negatives represent bacteria present but predicted by the model as absent:

$$TNR = TN\,/\,(FP + TN) \tag{4}$$

where true negative is correctly predicted bacteria presence, and false positives are bacteria absences classified as present by the model. TPR and TNR are widely referred to as sensitivity and specificity; both are used in the Receiver Operator Characteristic (ROC) curve (i.e. sensitivity vs. 1-specificity) whose tangent slope is equal to 1 (Hanley and McNeil, 1982):

$$FPR = FP\,/\,(FP + TN) \tag{5}$$

where FPR is equivalent to "fall out" which in binary classification is equal to (1-specificity):

$$ACC = (TP + TN)\,/\,(P + N) \tag{6}$$

where P is the number of actual presence instances and N is the number of absence instances. Selection of the final prediction points was based on a combination of the indices and is explained in detail below.

2.3.2.2. ABUNDANCE and HYBRID Model Evaluation

The predictive accuracy of *Vibrio* spp. ABUNDANCE models was assessed using random holdout validation analysis. Datasets for each species were randomly partitioned into a training dataset containing 80% of the original records and a validation dataset containing the remaining 20%. The models described above were developed using the training dataset and

Table 1. Correlation Coefficients for *Vibrio* spp. Counts and List of Selected Environmental Variables

	Lat	Lon	Month	Temp	Saln	Inter
V. vulnificus (cells 10 ml^{-1})*	- 0.03	- 0.11	0.25	0.28	0.09	0.22
V. parahaemolyticus (cells 10 ml^{-1})*	- 0.10	0.01	0.13	0.17	0.09	0.19
Latitude		0.15	0.03	0.06	- 0.75	- 0.56
Longitude			0.06	0.06	- 0.32	- 0.15
Month				0.96	- 0.04	0.54
Temperature (°C)*					- 0.04	0.57
Salinity (‰)*						0.76
Interaction term*						

*Included in final model development. Significant correlations at the alpha=0.05 level are highlighted in bold.

Table 2. Best-Fit Likelihood Algorithms for *V. vulnificus* and *V. parahaemolyticus*

	V. vulnificus	*V. parahaemolyticus*
Model	$P_{presence} = e(^{logit}) / [e(^{logit}) + 1]$**	$P_{presence} = e(^{logit}) / [e(^{logit}) + 1]$**
GLM	$Logit = \beta_0 + \beta_1[T] + \beta_2[S] + \beta_3[(T*S)]$	$Logit = \beta_0 + \beta_1[T] + \beta_2[S]$
GAM	$Logit = \beta_0 + S_1[T] + S_2[S] + S_3[[T*S)]$	$Logit = \beta_0 + S_1[T] + S_2[S]$
RF	$P_{presence}$ = randomForest($T + S + (T*S)$)	$P_{presence}$ = randomForest($T + S$)

**Not applicable to RF model. Probability of presence ($P_{presence}$) is a function of logit.

subsequently employed to predict cell number using the holdout dataset. This process was repeated 100 times with a different random partition each time. Mean error (ME) and mean absolute error (MAE) were used to compare estimated bacterial abundance to observed abundance, identify outliers in each model fit, and evaluate comparative model performance.

HYBRID models were evaluated with the same presence-only validation dataset used to assess the ABUNDANCE models. The hybrid models were assessed with presence-only holdout dataset, denoted “HYBRID/P”. To measure hybrid method performance in predicting *Vibrio* spp. abundance at all sample locations, without bacteriological data input, the original validation dataset containing both zero and non zero records was used. “HYBRID” denotes hybrid models evaluated with original holdout dataset. Additionally, unweighted model averages were calculated for both species. All hybrid analyses employed the LIKELIHOOD model structures shown in Table 3.

2.3.2.3. Mean Model

ABUNDANCE and HYBRID models were compared to a mean statistical null model, i.e., the average value of the response variable, *Vibrio* spp. For validation, empirical models including the mean model were input to the holdout analysis.

3. Results

3.1. Observations

Over the eight months during which *Vibrio* spp. counts in the water samples were obtained, 46% contained *V. vulnificus* and 68% contained *V. parahaemolyticus*. In samples positive for *V. vulnificus*, the median and mean counts were 4 and 6 cells 10 ml^{-1} respectively, and concentrations ranged from 1 to 30 cells 10 ml^{-1} (Figure 2). For *V. parahaemolyticus*, the median and mean count was 7 and 9.5 cells 10 ml^{-1}, respectively, and concentrations ranged from 1 to 50 cells 10 ml^{-1} (Figure 2). Counts were obtained for samples collected at temperatures ranging from 8 to 31 °C and 0 to 14 ‰ salinity. The highest number of *Vibrio* spp. were in water samples at 28 °C and salinity of 11.5 ‰ (Figure 3). These results are consistent with those reported for *Vibrio* spp. in Chesapeake Bay by Jacobs et al. (2010).

3.2. Modeling Occurrence and Abundance of *Vibrio* spp. in Chesapeake Bay

Descriptive correlation analyses relating environmental predictors to *Vibrio* spp. distribution and results of LIKELIHOOD, ABUNDANCE, and HYBRID predictive models are presented as follows.

3.2.1. Correlation of *Vibrio* spp. with Environmental Predictors

The predictive potential of environmental parameters was examined using univariate correlation analysis for *Vibrio* counts in samples containing given the *Vibrio* spp. Statistically significant correlations were found between bacteria count and surface water temperature, month, and salinity *x* temperature interaction (Table 1). Although statistically significant, the correlation coefficients were low. It is important to note that correlations observed for month and, potentially, interaction may derive from cross-correlation with surface water temperature. Furthermore, an insignificant correlation between bacteria count and surface salinity was observed, which can likely be attributed to the limited range of salinity observations used in *Vibrio* spp. model development. For the purpose of comparison and consistency to pre-existing *Vibrio* spp. models (Wright et al., 1996; Louis et al., 2003; Constantin de Magny et al., 2009; Jacobs et al., 2010) in the Chesapeake Bay, we decided to include salinity in final model development. For most sampling locations, total bacterial count followed a seasonal pattern following the temperature. Linear correlations between *Vibrio* count and salinity, latitude, or longitude were not statistically significant (Table 1). Since freshwater discharge impacts both nutrient inflow and sediment transport, it can influence bacterial abundance.

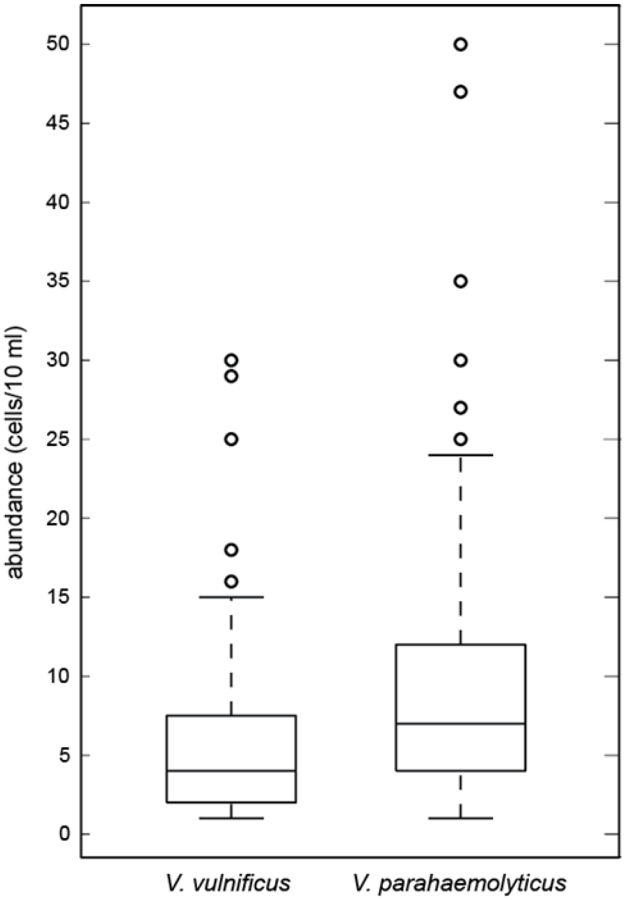

Figure 2. Boxplot showing concentration (cells 10 ml^{-1} > 0) for *V. vulnificus* (n = 68) and *V. parahaemolyticus* (n = 100).

3.2.2. LIKELIHOOD models

A stepwise selection process was used to select a LIKELIHOOD model, whereby each explanatory variable was entered sequentially into each model. The entire suite of models was tested, and selected variables retained only if significant. For the model evaluation, significance was set at an alpha level of 0.05. GLM and GAM logistic regression for both *V. vulnificus* and *V. parahaemolyticus* showed temperature and salinity, and for *V. vulnificus* interaction between the two variables, were core explanatory parameters for the three LIKELIHOOD models. Table 2 presents the best-fit models developed for *V. vulnificus* and *V. parahaemolyticus*, where probability of bacteria presence (P_{presence}) is defined in Equation 2.

Figure 4 illustrates the probability of *V. vulnificus* being present as predicted by best-fit a) GLM, b) GAM and c) RF LIKELIHOOD models (Table 2). Likelihood of presence was split into absence (n = 48; *median prob.* = 0.47, 0.26 and 0.27) and presence (n = 50; *median prob.* = 0.56, 0.67 and 0.57) observations. Figure 5 shows the probability of *V. parahaemolyticus* predicted by best-fit a) GLM, b) GAM and c) RF LIKELIHOOD models (Table 2). Likelihood of occurrence was split into absence (n = 82; *median prob.* = 0.52, 0.55 and 0.48) and presence (n = 40; *median prob.* = 0.69, 0.78 and 0.87) observations. Points falling outside of the 95th percentile in the boxplots represent outliers.

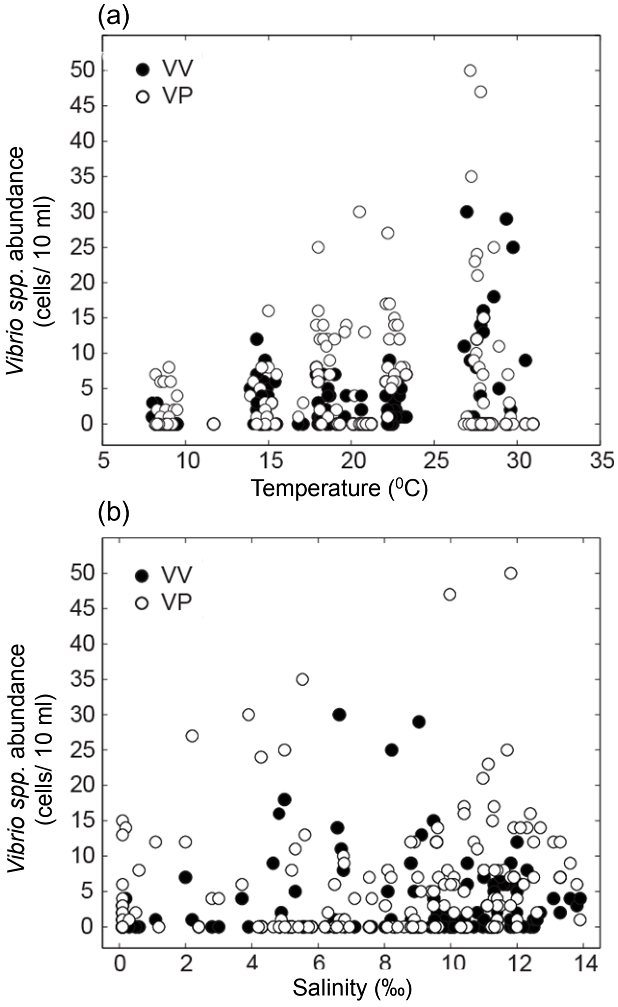

Figure 3. Plots showing the relationship between counts of *Vibrio* and (a) temperature (ºC) and (b) salinity (‰).

LIKELIHOOD GLM, GAM and RF models used to predict bacterial presence required selection of an optimal prediction point or threshold. Rather than setting a prediction point 0.5 arbitrarily, the prediction point for each species was based on four performance indices: TPR, FPR, TNR and ACC. With the goal of maximized model prediction skill and binary classification, information from each of these metrics (Figure 6), as well mean and median statistics from predicted probabilities (Figure 4), was used to select the optimal prediction point for each species. Because no significant difference was observed

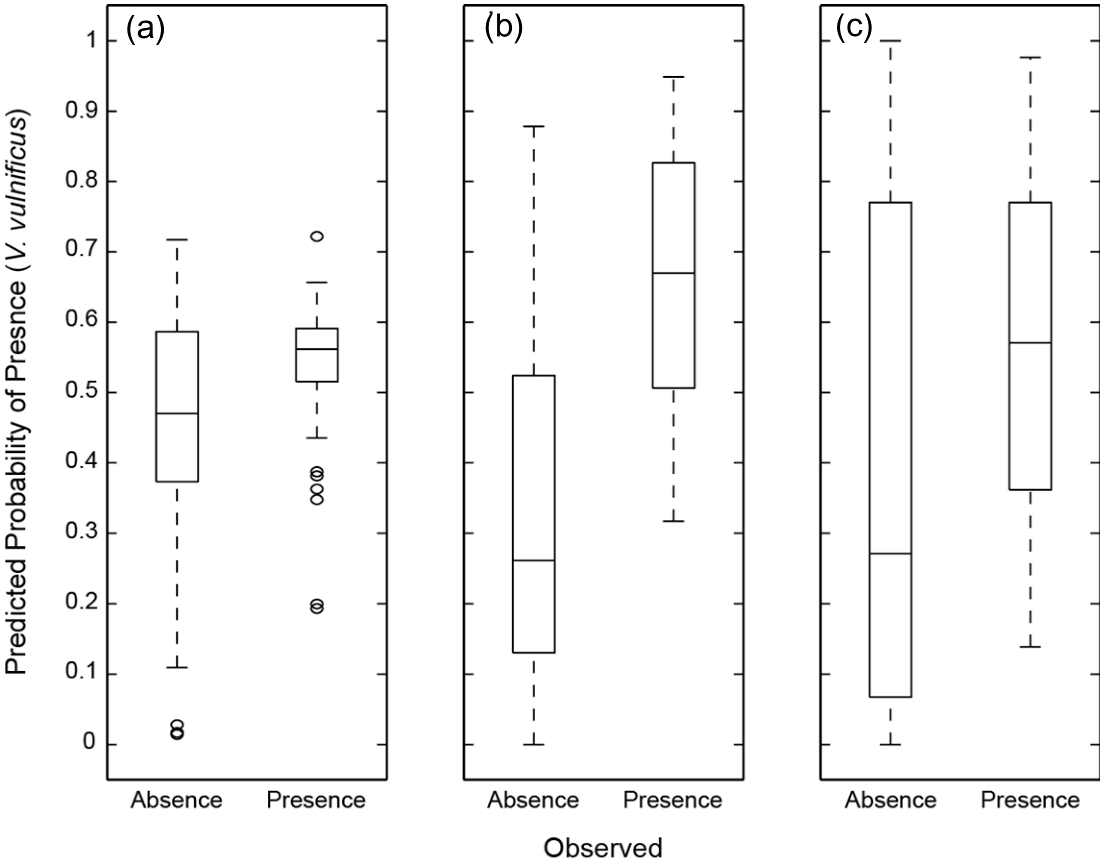

Figure 4. Performance of (a) GLM, (b) GAM and (c) RF *Vibrio vulnificus* classification models (Table 2), presented as boxplots comparing presence and absence with modeled probability, where threshold for presence is cell 10 ml^{-1}.

between the accuracy index for 0.4, 0.5, and 0.6 prediction points for *V. parahaemolyticus*, each threshold was tested in the holdout analysis, yielding greater accuracy, with an optimal threshold of 0.6. With this information, maximum ACC and TPR were selected, yielding an optimal threshold of 0.4 for *V. vulnificus* (ACC: 0.63 for GLM, 0.72 for GAM, and 0.68 for RF; Table 3), and 0.6 for *V. parahaemolyticus* (ACC: 0.62 for GLM, 0.65 for GAM, and 0.67 for RF; Table 3).

3.2.3. ABUNDANCE Models

ABUNDANCE models described in section 2.3.1 were applied to all samples with *Vibrio* greater than 0 cells 10 ml^{-1}, using repeated random holdout validation tests. Results indicate RF offered better prediction when the bacterial counts were high and GAM and GLM offer better prediction when the numbers were low. Based on these performance patterns, unweighted model average predictions of GAM and RF were tested. For each species, four ABUNDANCE models were then applied: (1) GLM, (2) GAM, (3) RF, (4) model average. Each model was also compared to the mean prediction model in the holdout test to determine how well each model performed relative to assuming the mean *Vibrio* bacterial count for each species, which provides an estimate of the degree to which each empirical model offers an improvement over using the historic mean as the future prediction. This resulted in 10-pair wise tests. Applying the Bonferroni correction for multiple hypothesis tests, a p-value below 0.005 (p = 0.05 overall) indicates statistical significance for any given test.

Table 3. *V. vulnificus* and *V. parahaemolyticus* (Likelihood) Performance Metrics at Prediction Point 0.40 for *V.vulnificus* and 0.60 for *V. parahaemolyticus*

	V. vulnificus			*V. parahaemolyticus*		
	GLM	GAM	RF	GLM	GAM	RF
AUC	0.68	0.78	0.73	0.63	0.70	0.71
FPR	0.44	0.35	0.37	0.30	0.24	0.22
TPR	0.81	0.81	0.76	0.42	0.48	0.50
TNR	0.56	0.65	0.63	0.70	0.76	0.78
ACC	0.63	0.72	0.68	0.62	0.65	0.67

As shown in Table 4, the RF ABUNDANCE model provides the best predictive accuracy for *V. vulnificus,* with lowest MAE (3.87 cells 10 ml^{-1}) followed by average ABUNDANCE MAE (3.94 cells 10 ml^{-1}). The MAE values were statistically significantly lower than GLM and GAM MAE (p < 0.005). The model average and RF model had lower error than the mean model by a statistically significant amount (p = 0.005).

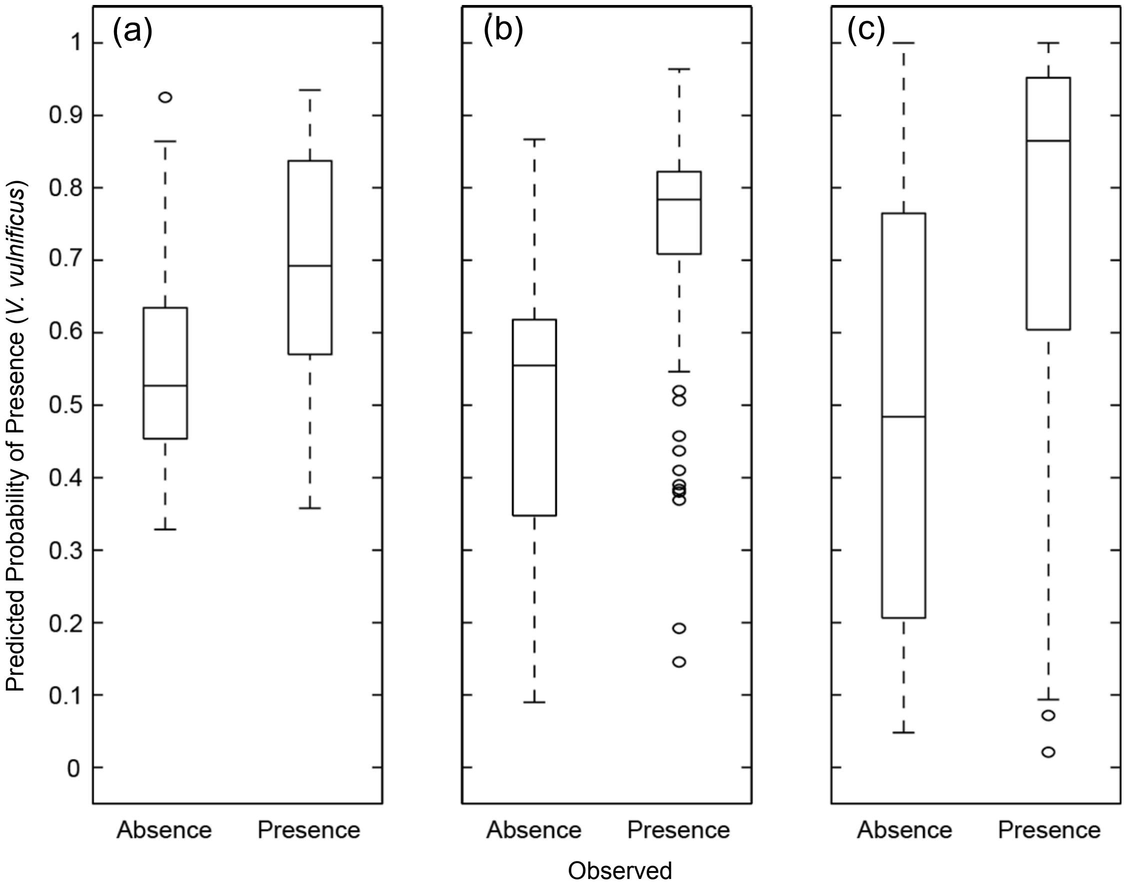

Figure 5. Performance of (a) GLM, (b) GAM and (c) RF for *Vibrio parahaemolyticus* classification models (Table 3), presented as boxplots comparing presence and absence with modeled probabilities where the threshold for presence is 1 cell 10 ml^{-1}.

For *V. parahaemolyticus*, the model average (5.62 cells 10 ml^{-1}) and RF (5.76 cells 10 ml^{-1}) had the lowest MAE values, and were lower than MAEs of the GLM and GAM by a statistically significant amount ($p < 0.005$). The difference between MAE for model average and RF were not statistically significant ($p = 0.38$). While all four models were statistically different from the mean predictions, only the model average and RF model outperform the mean model ($p < 0.005$).

The prediction accuracy of each model was examined whereby the predictions were binned based on the actual cell number obtained from the validation datasets (cells 10 ml^{-1} = 1, 2 ~ 4, 5 ~ 10 and >10 for *V. vulnificus*) (Figures 7a and 7c) and (cells 10 ml^{-1} = 1, 2 ~ 4, 5 ~ 10, 11 ~ 15 and > 15 for *V. parahaemolyticus*) (Figures 7b and 7d). While the RF model had a lower overall MAE than GLM and GAM for *V. vulnificus*, it exhibited a larger MAE in the lower concentration bins (Figure 7c) due to over prediction in those bins (high ME values) (Figure 7a). For *V. parahaemolyticus*, overall ME values showed all models, except RF, under predicted the cell count because of significant under prediction at high concentrations (Figure 7b). The GLM and GAM exhibited lower MAE values at lower concentrations than the RF (Figure 7d). However, at counts higher than 5 cells 10 ml^{-1}, the RF model outperformed both GLM and GAM. Averaging model predictions reduced overall RF MAE, but increased the MAE when counts were high.

Table 4. Comparison of Holdout ABUNDANCE MAEs (Cells 10 ml^{-1}) Based on 100 Random Holdout Samples for *V. vulnificus* and *V. parahaemolyticus*

Model	Mean MAE	GAM	RF	AVG	MEAN
		V. vulnificus			
GLM	4.69	0.61	< 0.01	< 0.01	0.09
GAM	4.79		< 0.01	< 0.01	0.02
RF	3.87			0.61	< 0.01
AVG	3.94				< 0.01
MEAN	4.39				
		V. parahaemolyticus			
GLM	7.43	0.70	< 0.01	< 0.01	< 0.01
GAM	7.51		< 0.01	< 0.01	< 0.01
RF	5.76			0.38	< 0.01
AVG	5.62				< 0.01
MEAN	6.34				

[*] p-Values in bold represent statistically significant differences between models at the alpha = 0.005 level.

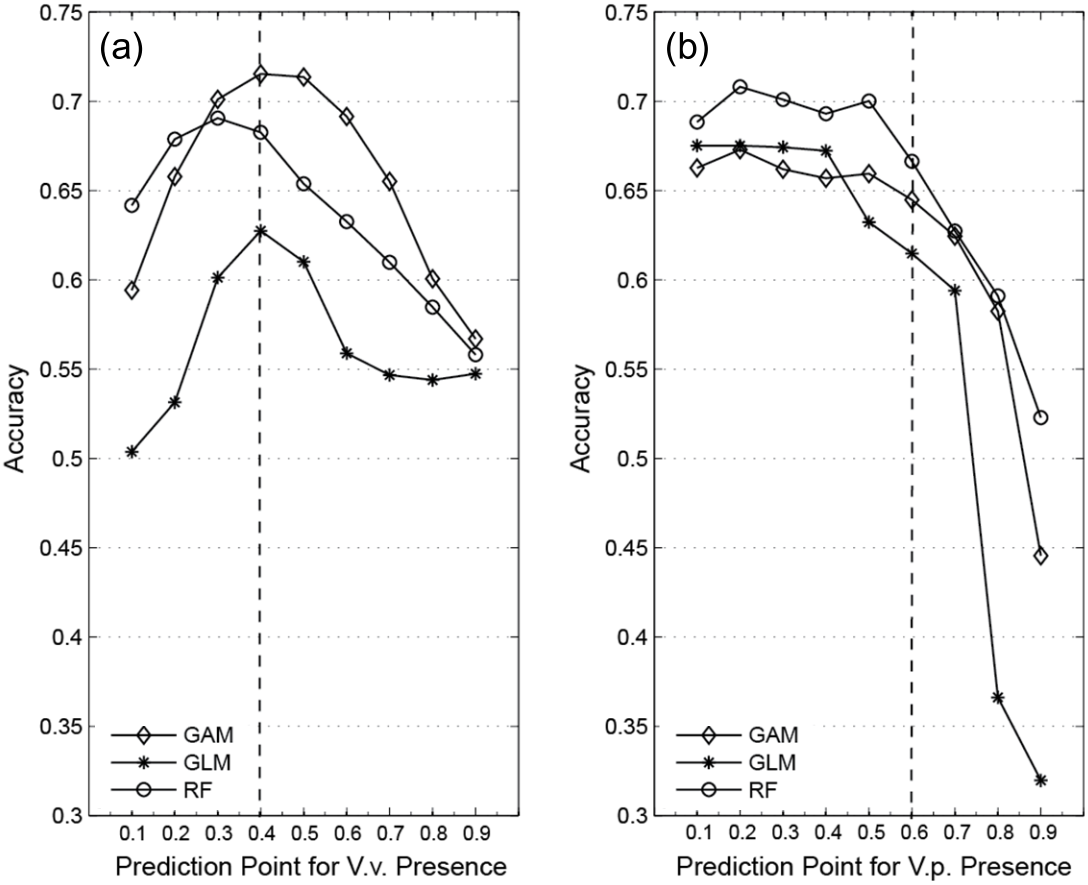

Figure 6. Optimization of prediction point (expressed as decimal fraction) to determine p for (a) *V. vulnificus* (0.40) and (b) *V. parahaemolyticus* (0.60).

Table 5. Comparison of MEs and MAEs for ABUNDANCE and HYBRID Models for *V. vulnificus* and *V. parahaemolyticus*

V. vulnificus			*V. parahaemolyticus*		
Error Metric	Error	MEAN	Error Metric	Error	MEAN
ME.ABUNDANCE	- 0.05	- 0.05	ME.ABUNDANCE	0.14	0.16
ME.HYBRID/P	- 1.58	- 3.25	ME.HYBRID/P	- 1.93	- 2.98
ME.HYBRID	- 0.28	0.19	ME.HYBRID	- 1.91	- 0.11
MAE.ABUNDANCE	3.87	4.39	MAE.ABUNDANCE	5.76	6.34
MAE.HYBRID/P	2.79	4.30	MAE.HYBRID/P	4.36	5.83
MAE.HYBRID	2.94	3.44	MAE.HYBRID	5.26	6.12

3.2.4. HYBRID Models

Based on error results of both LIKELIHOOD and ABUNDANCE models, a two-step GAM classification/RF regression HYBRID modeling approach was used. Other classification/regression model combinations (e.g., GLM/GLM, GAM/GAM, and RF/RF) were also tested, but GAM/RF was the best performing hybrid combination. The GAM/RF combination exhibited significantly lower error ($p < 0.005$) than other HYBRID combinations for *V. parahaemolyticus*, and similar error for *V. vulnificus*, although the difference was not statistically significant ($p = 0.005$). To assess the prediction accuracy of our HYBRID approach we evaluated the model using two different holdout datasets: (1) a presence-only validation dataset, and (2) the original validation dataset irrelevant of presence or absence. Using the same presence-only validation dataset that was used to evaluate the ABUNDANCE models allowed direct comparison of the prediction accuracy of the HYBRID and ABUNDANCE models. Model evaluation using the original holdout dataset allowed an estimation of *Vibrio* counts without the bacteriological data.

Table 5 compares ME and MAE from the ABUNDANCE RF model with those of the HYBRID model, using the presence-only validation dataset (HYBRID/P), and the model validated with the original dataset (HYBRID) for *V. vulnificus* and *V. parahaemolyticus*. The RF ABUNDANCE, HYBRID/P, and HYBRID all exhibited negative bias (ME) due to under predi-

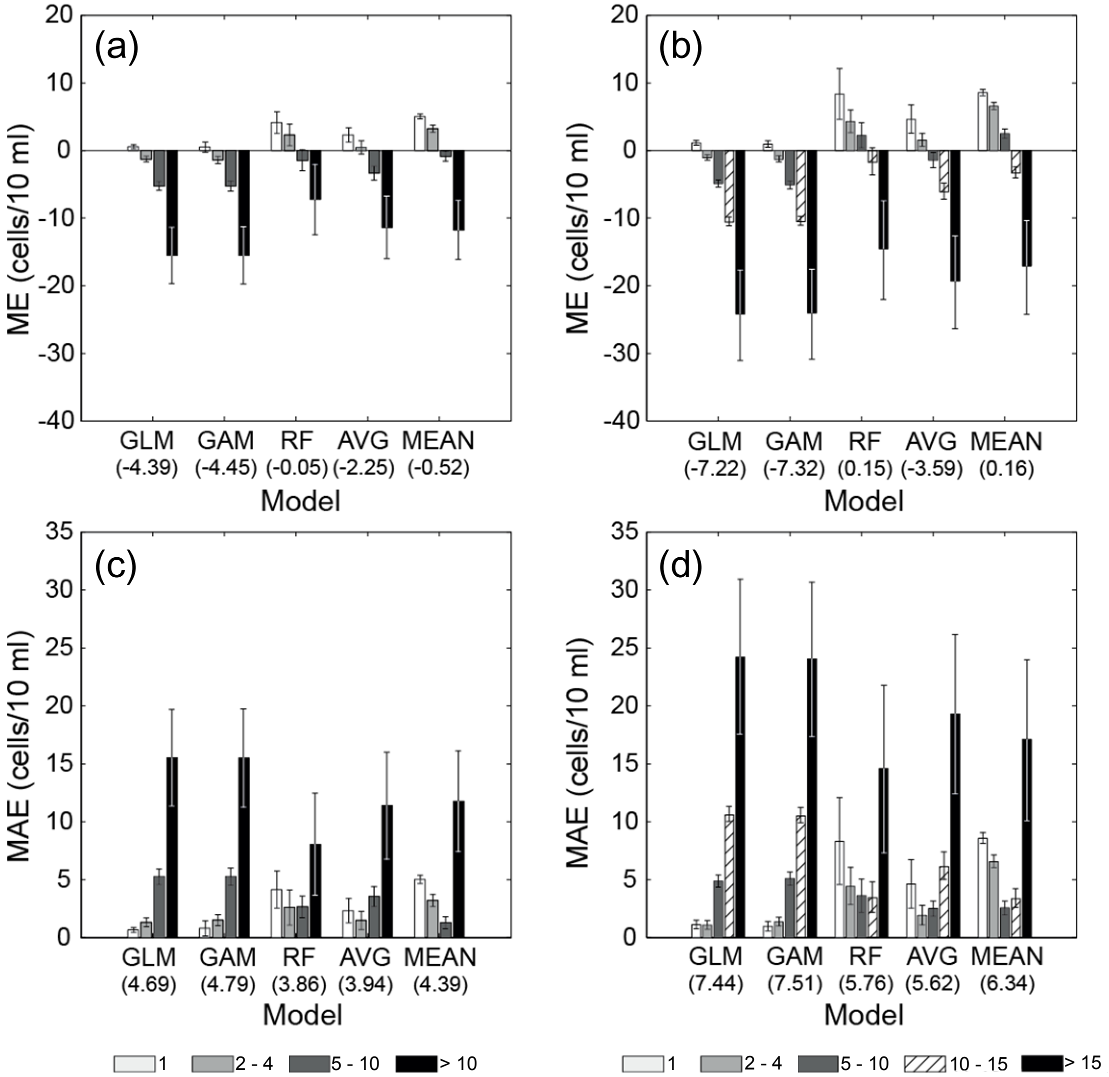

Figure 7. Binned ME and MAE values (cells 10 ml^{-1}) of each ABUNDANCE model for *V. vulnificus* (a) and (c) and *V. parahaemolyticus* (b) and (d), shown as bar graphs with error bars (standard deviation).

ction of counts at high concentrations. Results of the hypothesis tests using the MAE as the error measure showed that when using presence-only data, significant error reduction from the ABUNDANCE RF model MAE (3.9 cells 10 ml^{-1}) to the HYBRID/P MAE (2.8 cells 10 ml^{-1}) was observed. When the HYBRID approach was used to predict the original zero and nonzero dataset, an improvement in error relative to ABUNDANCE model was also noted. These two predictions are not exactly comparable, as the ABUNDANCE model was trained and evaluated using only samples with confirmed *Vibrio* counts, while the HYBRID prediction applied to all data, without bacteriological laboratory data.

For *V. parahaemolyticus*, both HYBRID/P and HYBRID exhibited negative bias largely due to under estimation when counts were high, and a positive bias for RF ABUNDANCE model. Both HYBRID/P (4.4 cells 10 ml^{-1}) and HYBRID (5.26 cells 10 ml^{-1}) offer an improvement in MAE relative to ABUNDACNE (5.8 cells 10 ml^{-1}) model. All of the HYBRID models offer improvement over using the mean of the validation dataset.

4. Discussion and Conclusions

The empirical models presented in this study demonstrate significant skill in estimating probability of occurrence of *Vibrio* spp., as well as bacterial counts in Chesapeake Bay water samples when bacteriological count data are included. When the HYBRID approach was used to generate estimates of *Vibrio* counts, an overall reduction in error was observed compared to presence-only ABUNDANCE models when the models were evaluated only at sites with detectable *Vibrio* counts. The fact that HYBRID outperformed ABUNDANCE, when evaluated at sites where *Vibrio* was present, is surprising, since the ABUNDANCE models benefited from data on presence versus absence. Examination of partial dependence plots indicated differences in model performance are a product of differences between the HYBRID and ABUNDANCE models at moderate values of temperature and salinity. The differences were relatively small, however, and no major difference was noted in structure between RF ABUNDANCE model and RF component of the HYBRID model. We conclude that the enhanced performance of HYBRID, relative to ABUNDANCE, most probably derives from the fact that models trained for a broader range of conditions—as was the case for HYBRID, since errors in the GAM LIKELIHOOD model led to a more diverse training set for the RF component of the HYBRID—tend to be more generalizable than models trained under more narrow conditions, even when these narrow conditions capture the range of the specific response variable of interest (Nateghi and Guikema, 2013). HYBRID performed at least as well as ABUNDANCE, which indicates that the HYBRID approach allows for modeling both presence and abundance without loss of skill relative to an abundance model supplied with perfect prior information on presence versus absence.

When both the HYBRID and ABUNDANCE models were evaluated for all sites, the predictive accuracy of the HYBRID was better than that of the ABUNDANCE model, though the difference was not statistically significant for *V. parahaemolyticus.* Similar to model behavior observed for the ABUNDANCE models for both species, overall error reduction using the HYBRID modeling approach showed the two-step approach tends to over predict counts at low *Vibrio* concentrations. Furthermore, when evaluating prediction performance of each model relative to the mean model, a statistically significant improvement over the mean value of the validation dataset in all models was noted, except for the *V. parahaemolyticus* HYBRID model. It is important to emphasize that when using the complete original dataset for validation (HYBRID), zero-inflation and a lower overall mean model value must be considered. In future model evaluation using zero-inflated datasets, alternative methods of mean model comparisons should be employed.

The empirical models presented here offer a novel approach for estimating *Vibrio* spp. concentration in the upper Chesapeake Bay. We note that the study was limited by the small number of samples available to train and evaluate the models. First, the field data used in this study was limited to the oligohaline (0 ~ 6 ‰) and mesohaline (6 ~ 18 ‰) regions of the upper Chesapeake Bay. Because our models were trained using data for fresh and brackish water, extrapolation, of the models to saline regions may result in greater error and thus, decreased accuracy of prediction for *Vibrio* spp. near the mouth of the Chesapeake Bay. Specific attention to this discrepancy will be required if the models developed in this study are to be applied to coastal regions. Therefore, data from more saline waters will be needed to train the model. Work in progress covers whole Bay hindcast predictions using temperature and salinity from satellite sensors (Urquhart et al., 2012) to understand long-term trends of *Vibrio* spp. likelihood and abundance throughout the Bay. Water samples were collected only in the upper Chesapeake at a limited number of stations over a two-year period. A longer and more intense sampling record would be valuable to produce more robust models with improved predictive capability.

Since the risk of vibriosis is directly related to *Vibrio* spp. pathogenicity and abundance, the primary motivation for this study was in the absence of models available to estimate *Vibrio* spp. concentration in coastal waters. Therefore, an objective of this study was to extend available modeling capabilities to provide concentration estimates of *Vibrio* spp. bacteria in the upper Chesapeake Bay. While the statistical models presented here demonstrate significant skill in estimating Vibrio abundance, the errors associated with the estimates are not insignificant, particularly at high bacterial concentrations. These error rates could, presumably, be lowered with more spatially and temporally distributed training data. Just as importantly, more information is needed on the quantitative relationships between abundance, pathogenicity, and human infection. From an applications perspective, improved abundance estimates are valuable only insomuch as they translate into improved assessment of public health risk.

In summary, we have presented several empirical algorithms for estimating the likelihood of *Vibrio* occurrence as well as abundance (cells 10 ml^{-1}) in Chesapeake Bay surface water. To estimate the probability of *Vibrio* spp. being detected in Bay water, we tested several binary classification methods. To model *Vibrio* spp. abundance, several regression methods were applied to samples positive for *Vibrio* spp. A two-step hybrid approach using GAM for classification and RF for regression was employed to estimate abundance of *Vibrio* spp. in the absence of bacteriological data. For LIKELIHOOD models, GAM demonstrated a greater accuracy and improved positive rate than GLM and RF models. ABUNDANCE models, GLM and GAM exhibited higher prediction accuracy when counts of *Vibrio* spp. were low. However, RF exhibited lower overall mean absolute error. HYBRID performed better than ABUNDANCE at sites where *Vibrio* presence had been confirmed by bacteriological methods, and predicted abundance as well or better than ABUNDANCE even when evaluated for sites both with and without *Vibrio* spp. confirmed to be present. Thus, HYBRID modeling offers the potential to predict both presence and abundance of *Vibrio* bacteria in Chesapeake Bay surface water. Since presence and abundance of *Vibrio* spp. are relevant to the risk of infection, this capability offers meaningful improvement over existing monitoring and prediction systems.

Acknowledgments. This study was funded in part by the Johns Hopkins University Department of Earth and Planetary Sciences and by a grant from the NASA Headquarters Applied Sciences Division. The authors would like to thank the Maryland Department of Natural Resources, Chesapeake Bay Program, and NASA GEO-CAPE field campaigns for allowing sample collection upon routine Bay cruises as well as Valerie Cohen and Merina Elahi (MPRI, UMCP) for wet lab assistance. The authors gratefully acknowledge partial support at the Maryland Pathogen Research Institute at the University of Maryland, College Park, from National Institutes of Health Grant No. 2RO1A1039129-11A2 and National Science Foundation Grant No. 0813066.

References

Banakar, V., Constantin de Magny, G., Jacobs, J., Murtugudde, R., Huq, A., Wood, R.J., and Colwell, R.R. (2011). Temporal and spatial variability in the distribution of Vibrio vulnificus in the Chesapeake Bay: A hindcast study. *Ecohealth,* 8(4), 456-67. http://dx.doi.org/10.1007/s10393-011-0736-4

Bauer, A., and Rorvik, L.M. (2007). A novel multiplex PCR for the identification of Vibrio parahaemolyticus, Vibrio cholerae and Vibrio vulnificus. *Lett. Appl. Microbiol.,* 45(4), 371-375. http://dx.doi.org/10.1111/j.1472-765X.2007.02195.x

Breiman, L. (2001). Random forests. *Mach. Learning,* 45(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

Cameron, A.C., and Trivedi, P.K. (2013). Regression analysis of count data, *Econometric Society Monograph,* Cambridge University Press, UK. http://dx.doi.org/10.1017/CBO9781139013567

Centers for Disease Control and Prevention (2013). Foodborne Diseases Active Surveillance Network (FoodNet). Accessed: November 10, 2013. http://www.cdc.gov/foodnet/data/trends/trends-2012.html

Colwell, R.R., Kaper, J., and Joesph, S.W. (1977). Vibrio cholerae, Vibrio parahaemolyticus, and other Vibrios: Occurrence and distribution in Chesapeake Bay. *Science,* 198(4315), 394-396. http://dx.doi.org/10.1126/science.198.4315.394-a

Constantin de Magny, G., Long, W., Brown, C., Hood, R., Huq, A., Murtugudde, R., and Colwell, R. (2009). Predicting the distribution of Vibrio spp. in the Chesapeake Bay: A Vibrio cholerae case study. *Ecohealth,* 6(3), 378-389. http://dx.doi.org/10.1007/s10393-009-0273-6

DePaola, A., Motes, M.L., Cook, D.W., Veazey, J., Garthright, W.E., and Blodgett, R. (1997). Evaluation of an alkaline phosphatase-labeled DNA probe for enumeration of Vibrio vulnificus in Gulf Coast oysters. *J. Microbiol. Methods,* 29(2), 115-120. http://dx.doi.org/10.1016/S0167-7012(97)00030-4

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models,* SAGE Publications, Inc.

Guikema, S.D., and Quiring, S.M. (2012). Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. *Reliab. Eng. Syst. Saf.,* 99, 178-182. http://dx.doi.org/10.1016/j.ress.2011.10.012

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29-36. http://dx.doi.org/10.1148/radiology.143.1.7063747

Hastie, T., and Pregibon, D. (1992). *Generalized Linear Models in: Statistical Models in S,* Chapman & Hall/CRC, London, UK.

Hastie, T., and Tibshirani, R. (1986). Generalized additive models. *Stat. Sci.,* 1(0), 297-310. http://dx.doi.org/10.1214/ss/1177013604

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall/CRC, London, UK.

Hoffman, M.J., Miyoshi, T., Haine, T.W.N., Ide, K., Brown, C.W., and Murtugudde, R. (2012). An advanced data assimilation system for the Chesapeake Bay: Performance evaluation. *J. Atmos. Ocean. Technol.,* 29(10), 1542-1557. http://dx.doi.org/10.1175/JTECH-D-11-00126.1

Hoge, C.W., Watsky, D., Peeler, R.N., Libonati, J.P., Israel, E., and Morris, J.G. (1989). Epidemiology and spectrum of Vibrio infections in a Chesapeake Bay community. *J. Infect. Dis.,* 160(6), 985-993. http://dx.doi.org/10.1093/infdis/160.6.985

Howard, R.J., and Bennett, N.T. (1993). Infections caused by halophilic marine Vibrio bacteria. *Ann. Surg.,* 217(5), 525-531. http://dx.doi.org/10.1097/00000658-199305010-00013

Jacobs, J., Rhodes, M., Brown, C., Hood, R., Leight, A., Long, W., and Wood, R. (2010). Predicting the distribution of Vibrio vulnificus in Chesapeake Bay. *NOAA Tech. Mem. NOS NCCOS,* 112(12).

Kaneko, T., and Colwell, R. (1973). Ecology of Vibrio parahaemolyticus in Chesapeake Bay. *J. Bacteriol.,* 113(1), 24-32.

Kaneko, T., and Colwell, R. (1974). Incidence of Vibrio parahaemolyticus in Chesapeake Bay. *Appl. Microbiol.,* 30(2), 251-257.

Kaper, J.B., Remmers, E.F., Lockman, H., and Colwell, R.R. (1981). Distribution of Vibrio parahaemolyticus in Chesapeake Bay during the summer season. *Estuaries,* 4(4), 321-327. http://dx.doi.org/10.2307/1352156

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News,* 2, 18-22.

Lipp, E., Rodriguez-Palacios, C., and Rose, J. (2001). Occurrence and distribution of the human pathogen Vibrio vulnificus in a subtropical Gulf of Mexico estuary. *Hydrobiologia,* 460(1-3), 165-173. http://dx.doi.org/10.1023/A:1013127517860

Louis, V.R., Russek-Cohen, E., Choopun, N., Rivera, I.N.G., Gangle, B., Jiang, S.C., Rubin, A., Patz, J.A., Huq, A., and Colwell, R.R. (2003). Predictability of Vibrio cholerae in Chesapeake Bay. *Appl. Environ. Microbiol.,* 69(5), 2773-2785. http://dx.doi.org/10.1128/AEM.69.5.2773-2785.2003

Maryland Department of Health and Mental Hygiene (2013). Maryland Department of Health and Mental Hygiene, Cases of Selcted Notifiable Conditions Reported in Maryland. Accessed: April 6, 2013. http://phpa.dhmh.maryland.gov/SitePages/disease-conditions-count-rates.aspx/

McCarthy, S.A., DePaola, A., Cook, D.W., Kaysner, C.A., and Hill, W.E. (1999). Evaluation of alkaline phosphatase - and digoxigenin-labelled probes for detection of the thermolabile hemolysin (tlh) gene of Vibrio parahaemolyticus. *Lett. Appl. Microbiol.,* 28(1), 66-70. http://dx.doi.org/10.1046/j.1365-2672.1999.00467.x

Motes, M.L., DePaola, A., Cook, D.W., Veazey, J.E., Hunscuker, J.C., Garthright, W.E., Blodgett, R.J., and Chirtel, S.J. (1998). Influence of water temperature and salinity on Vibrio vulnificus in Northern Gulf and Atlantic Coast oysters (Crassostrea virginica). *Appl. Environ. Microbiol.,* 64(4), 1459-1465.

Nateghi, R., and Guikema, S.D. (2013). Estimating power distribution system outages during tropical cyclones in the Gulf Region of the U.S. with reduced complexity models. *Risk Anal.* (under review).

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Stat. Soc. Ser. A. (Stat. Soc.),* 135(3), 370-384. http://dx.doi.org/10.2307/2344614

Parveen, S., DaSilva, L., DePaola, A., Bowers, J., White, C., Munasinghe, K.A., Brohawn, K., Mudoh, M., and Tamplin, M. (2013). Development and validation of a predictive model for the growth of Vibrio parahaemolyticus in post-harvest shellstock oysters. *Int. J. Food Microbiol.,* 161(1), 1-6. http://dx.doi.org/10.1016/j.ijfoodmicro.2012.11.010

Parveen, S., Hettiarachchi, K.A., Bowers, J.C., Jones, J.L., Tamplin, M.L., McKay, R., Beatty, W., Brohawn, K., DaSilva, L.V., and DePaola, A. (2008). Seasonal distribution of total and pathogenic Vibrio parahaemolyticus in Chesapeake Bay oysters and waters. *Int. J. Food Microbiol.,* 128(2), 354-361. http://dx.doi.org/10.1016/j.ijfoodmicro.2008.09.019

Strom, M.S., and Paranjpye, R.N. (2000). Epidemiology and pathogenesis of Vibrio vulnificus. *Microb. Infect.,* 2(2), 177-188. http://dx.doi.org/10.1016/S1286-4579(00)00270-7

Urquhart, E.A., Hoffman, M.J., Murphy, R.R., and Zaitchik, B.F. (2013). Geospatial interpolation of MODIS-derived salinity and temperature in the Chesapeake Bay. *Remote Sens. Environ.,* 135(0), 167-177. http://dx.doi.org/10.1016/j.rse.2013.03.034

Urquhart, E.A., Zaitchik, B.F., Hoffman, M.J., Guikema, S.D., and Geiger, E.F. (2012). Remotely sensed estimates of surface salinity in the Chesapeake Bay: A statistical approach. *Remote Sens. Environ.,* 123(0), 522-531. http://dx.doi.org/10.1016/j.rse.2012.04.008

Urquhart, E.A., Zaitchik, B.F., Waugh, D.W., Guikema, S.D., Del Castillo, C.E. (2014). Uncertainty in Modelling Predictions of *Vibrio Vulnificus* Response to Climate Variability and Change: A Chesapeake Bay Case Study. PloS ONE 9(5), e98256. doi: 10.1371/journal.pone.0098256

Virginia Department of Health (2013). Virginia Department of Health, Virginia Reportable Disease Surveillance Data. Accessed: May 5, 2013. http://www.vdh.virginia.gov/Epidemiology/Surveillance/SurveillanceData/

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC, London, UK.

Wright, A.C., Hill, R.T., Johnson, J.A., Roghman, M.C., Colwell, R.R., and Morris, J.G. (1996). Distribution of Vibrio vulnificus in the Chesapeake Bay. *Appl. Environ. Microbiol.,* 62(2), 717-724.

Xu, J., Long, W., Wiggert, J., Lanerolle, L.J., Brown, C., Murtugudde, R., and Hood, R. (2012). Climate forcing and salinity variability in Chesapeake Bay, USA. *Estuaries Coasts,* 35(1), 237-261. http://dx.doi.org/10.1007/s12237-011-9423-5

Yamazaki, K., and Nwadiuto, E. (2012). Environmental predictors of pathogenic Vibrios in South Florida coastal waters. *Open Epidemiol.,* 5(0), 1-4. http://dx.doi.org/10.2174/1874297101205010001