

**APPLICATION OF SATELLITE REMOTE SENSING TO WATER QUALITY AND  
PATHOGENIC BACTERIA PREDICTION IN THE CHESAPEAKE BAY**

by

Nicole M. DeLuca

A dissertation submitted to Johns Hopkins University in conformity with  
the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October, 2019

© 2019 Nicole M. DeLuca

All Rights Reserved

## ABSTRACT

The Chesapeake Bay is home to an extensive *in situ* sampling campaign that has provided water quality measurements over multiple decades, aiding in the detection and regulation of environmental conditions that affect aquatic life, public health, and local economies. However, the current bi-monthly sampling can lack the temporal and spatial coverage needed for monitoring and modeling dynamic estuarine systems. While the time and cost of obtaining additional *in situ* samples can exceed available resources, satellite remote sensing has the potential to provide this higher temporal and spatial resolution data. The objective of this dissertation is to investigate the use of satellite remote sensing in the Chesapeake Bay for both water quality monitoring and the prediction of a naturally-occurring pathogenic bacterium, *Vibrio parahaemolyticus*, that is a leading cause of food-borne illness. The dissertation does this by exploring the use of multispectral information to improve satellite-derived total suspended solids concentration and the potential for remotely sensed water quality products to predict *V. parahaemolyticus* in the Chesapeake Bay. In addition, the dissertation uses the application of remote sensing for *V. parahaemolyticus* prediction as a case study to present a prospective tool for communicating predictive model uncertainty to environmental management decision-makers and end-users. The work in this dissertation provides insights and recommendations that can aid in future development of operational models for water quality parameters or bacterial pathogens that incorporate remotely sensed data. As the effects of poor water quality are better understood and the incidence of *Vibrio*

illness increases, improved operational models and uncertainty communication will become progressively important for protecting public and ecosystem health.

**Advisor:**

Dr. Benjamin F. Zaitchik, *Department of Earth and Planetary Sciences*

**Thesis Reader:**

Dr. Maya L. Gomes, *Department of Earth and Planetary Sciences*

## **ACKNOWLEDGMENTS**

I would first like to thank my advisory committee: Ben Zaitchik, Anand Gnanadesikan, and Darryn Waugh. Ben “adopted” me as his student mid-way through my PhD, and I truly appreciate that he never gave any less of himself as an advisor to me than he did to his other students. I quickly discovered Ben’s knack and passion for interdisciplinary collaboration, and I hope to also emulate that in my career. Ben has truly helped me to become a better scientist in a short amount of time, and I thank him for always listening to and steering me towards my career goals. I would also like to thank Anand Gnanadesikan for being a kind and supportive mentor over the past five years and providing me with many opportunities to engage in my new field of research. I thank Darryn Waugh for his wisdom and humor that was needed at many points during my PhD, and I thank him for advocating for my best interests as a graduate student. I would also like to thank Sarah Horst for her advice and support. Lastly, I thank Naomi Levin for seeing the potential in me and admitting me into this graduate program. I truly appreciate the amazing start she gave me as a PhD student and all of the scientific training she instilled that has allowed me to succeed.

I am very thankful for the friendships I have made with current and past EPS students over the years who I look forward to remaining friends and colleagues with for many years to come. I am especially thankful to Asha Jordan and Chris Holder, who have kept me laughing and filled with hot chocolate, milkshakes, and cookies throughout my PhD. I thank all of the current and past members of the Hydroclimate Group for their support and endless

willingness to share their expertise with me. I also would like to thank Ben Davis for his friendship, support, and collaboration that made this dissertation possible.

I thank my friends at Tranquility Manor Farms, who became my Maryland family and helped to give me a home away from home. To my horse, Epic, and my dog, Finn, I thank you both for giving me reasons to smile and relax, especially on the tougher days.

I thank my incredibly supportive parents and family for always believing in me and doing everything in their power to make my dreams and goals attainable. To my father, I am thankful for the example you gave me of determination, bravery, and willingness to take risks that has allowed me to persevere, especially over the past five years. To my mother, I am thankful for your encouragement, friendship, and the confidence you have in me that motivates me to overcome the challenges in both my education and personal life.

I am forever grateful for my husband, Zack, who has kept me grounded and laughing every single day. I thank him for his unconditional love, his understanding, his patience, and his endless support of my dreams and goals.

Finally, I am thankful for the financial support I have received through the Department of Earth and Planetary Sciences, the NSF IGERT Water, Climate, and Health fellowship, and the National Institutes of Allergy and Infectious Diseases.

## TABLE OF CONTENTS

ABSTRACT.....	ii
ACKNOWLEDGEMENTS.....	iv
LIST OF TABLES .....	x
LIST OF FIGURES .....	xii
1. CHAPTER 1: INTRODUCTION .....	1
1.1 Chesapeake Bay .....	1
1.2 Satellite Remote Sensing in the Chesapeake Bay .....	2
1.3 Satellite-Derived Total Suspended Solids .....	4
1.4 Satellite-Derived <i>Vibrio parahaemolyticus</i> Concentration.....	5
1.5 Dissertation Outline .....	6
2. CHAPTER 2: CAN MULTISPECTRAL INFORMATION IMPROVE REMOTELY SENSED ESTIMATES OF TOTAL SUSPENDED SOLIDS? A STATISTICAL STUDY IN CHESAPEAKE BAY .....	8
2.1 Introduction .....	9
2.2 Materials and Methods .....	12
2.2.1 Data Description .....	12
2.2.1.1 Satellite Data Processing .....	12
2.2.1.2 In Situ Measurements .....	12
2.2.1.3 Satellite-In Situ Matchups .....	14
2.2.2 Methods .....	15

2.2.2.1	Statistical Methodology .....	15
2.2.2.2	Statistical and Machine Learning Models .....	16
2.2.2.3	Geographic and Temporal Cross-Validation .....	18
2.3	Results and Discussion .....	19
2.3.1	Model Comparison .....	19
2.3.2	Comparison with Single-Band TSS Algorithm .....	21
2.3.3	Daily Satellite, In Situ Mapped Comparisons .....	26
2.3.4	Geographic and Temporal Cross-Validation .....	27
2.4	Conclusions .....	29
3.	CHAPTER 3: EVALUATION OF REMOTELY SENSED PREDICTION AND FORECAST MODELS FOR <i>VIBRIO PARAHAEMOLYTICUS</i> IN CHESAPEAKE BAY .....	33
3.1	Introduction .....	34
3.2	Data Description .....	38
3.2.1	In Situ Measurements .....	38
3.2.2	Satellite Data Processing and Algorithms .....	39
3.2.3	Satellite-In Situ Matches .....	40
3.3	Methods .....	42
3.3.1	Statistical Methodology .....	42
3.3.2	Classification Methods for Bacterial Presence and Absence Prediction.....	46

3.3.3	Regression Methods for Bacterial Abundance Prediction. ....	48
3.3.4	Statistical Null Models .....	49
3.3.5	Spatial Comparison .....	50
3.4	Results and Discussion .....	50
3.4.1	Presence, Absence Predictions for Same Week .....	51
3.4.2	Abundance Predictions for Same Week .....	55
3.4.3	Spatial Comparison .....	57
3.4.4	Presence, Absence for 1-Week Forecast .....	60
3.4.5	Abundance for 1-Week Forecast .....	63
3.5	Conclusion .....	63
4.	CHAPTER 4: PROVIDING PREDICTION UNCERTAINTY ON ENVIRONMENTAL MODELS USING QUANTILE REGRESSION FORESTS: A CASE STUDY FOR <i>VIBRIO PARAHAEMOLYTICUS</i> IN CHESAPEAKE BAY .....	73
4.1	Introduction .....	74
4.2	Data Description .....	78
4.2.1	In Situ <i>Vibrio parahaemolyticus</i> Measurements .....	78
4.2.2	Satellite Data Processing .....	79
4.2.3	Shellfish Harvesting Areas .....	80
4.3	Methods .....	80
4.3.1	Modeling Methodology .....	80



4.3.2	Quantile Regression Forest .....	81
4.3.3	Model Evaluation .....	82
4.4	Results and Discussion .....	83
4.4.1	Model Performance .....	83
4.4.2	Mean Versus Quantile Predictions .....	84
4.4.3	Scenarios Based on Level of Risk .....	87
4.4.4	Scenarios Varying Abundance Threshold and Level of Risk .....	93
4.5	Conclusions .....	95
5.	CHAPTER 5: CONCLUSIONS .....	98
5.1	Future Directions .....	101
	REFERENCES .....	104
	AUTHOR'S CURRICULUM VITAE .....	119

## LIST OF TABLES

<b>Table 2.1</b> Summary of variables used in statistical and machine learning model development and single-band algorithm calculation for this study.....	13
<b>Table 2.2</b> Holdout mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) for all models evaluated in the study for the 20% holdout validation dataset.....	20
<b>Table 2.3</b> Mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) for the pruned Random Forest (RF) model, Random Forest model using only 645-nm (RF(645)), the O-2012 algorithm as published, and the O-2012 algorithm fitted for our dataset (O-2012(fit)) for the 20% holdout validation dataset. ....	23
<b>Table 2.4</b> Mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) for the pruned RF model and O-2012 algorithm for holdout validation dataset above and below the 80th percentile.....	25
<b>Table 2.5</b> Generalizability cross-validation results. Mean absolute error (MAE) and mean square error (MSE) for the pruned RF model and O-2012 algorithm. Naming scheme for cross-validation is as follows: “East for West” indicates model trained on East dataset and validated West dataset.....	30
<b>Table 3.1</b> Description of environmental and remote sensing reflectance input structures used in this study.....	45

<b>Table 3.2</b> Summary of remote sensing products and <i>in situ</i> <i>Vibrio parahaemolyticus</i> abundance used in this study.....	54
<b>Table S3.1</b> Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for Same Week holdout classification predictions using remotely sensed inputs.....	70
<b>Table S3.2</b> Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for Same Week holdout classification predictions using <i>in situ</i> measured inputs.....	70
<b>Table S3.3</b> Weighted mean absolute error (WMAE) for Same Week holdout abundance predictions using remotely sensed (RS) and <i>in situ</i> inputs.....	71
<b>Table S3.4</b> Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for 1-Week Forecast holdout classification predictions using remotely sensed inputs.....	71
<b>Table S3.5</b> Weighted mean absolute error (WMAE) for 1-Week Forecast holdout abundance predictions using remotely sensed inputs.....	72

## LIST OF FIGURES

<b>Figure 2.1</b> Map of Chesapeake Bay estuary showing the 86 Chesapeake Bay Program measurement stations used in the satellite- <i>in situ</i> matchup dataset in Chapter 2.....	14
<b>Figure 2.2</b> Partial dependence plots for the 11 MODIS bands used as predictors in the Random Forest model before pruning.....	22
<b>Figure 2.3</b> Log-log plots showing one-to-one regressions of CBP <i>in situ</i> measured versus satellite-derived TSS from the pruned Random Forest and the O-2012 algorithm.....	24
<b>Figure 2.4</b> Mapped comparisons of daily remotely sensed TSS (mg/L) derived from O-2012 and RF model for 2017 dates.....	28
<b>Figure S2.1</b> Histogram of sampling times for all Chesapeake Bay program TSS data at or above 1 meter depth for years 2003-2016.....	32
<b>Figure 3.1</b> Map of Chesapeake Bay showing locations of satellite- <i>in situ</i> matches used for <i>Vibrio parahaemolyticus</i> modeling.....	43
<b>Figure 3.2</b> Pearson correlations between daily MODIS-derived sea surface temperature, sea surface salinity, total suspended solids, chlorophyll-a and corresponding <i>in situ</i> measurements.....	52
<b>Figure 3.3</b> Boxplots showing accuracy, sensitivity, and specificity for Same Week holdout classification predictions.....	56
<b>Figure 3.4</b> Boxplots showing weighted mean absolute error (WMAE) for Same Week holdout presence-abundance predictions.....	59

<b>Figure 3.5</b> Spatial comparison of Random Forest abundance predictions on MODIS 8-day composite from 20 July 2013 to 27 July 2013 using (A) sea surface temperature and sea surface salinity and (B) sea surface temperature, sea surface salinity, total suspended solids, and chlorophyll-a as remotely sensed inputs.....	62
<b>Figure 3.6</b> Boxplots showing accuracy, sensitivity, and specificity for 1-Week Forecast holdout classification predictions.....	64
<b>Figure 3.7</b> Boxplots showing weighted mean absolute error (WMAE) for 1-Week Forecast holdout abundance predictions.....	65
<b>Figure 4.1</b> Schematic of possible output scenarios from a Quantile Regression Forest.	77
<b>Figure 4.2</b> Time series of predicted <i>V. parahaemolyticus</i> abundance (GE/ml) at the Chester River, Fishing Bay, and the James River over 30 weeks in 2010 showing mean RF prediction values, 50 <sup>th</sup> quantile (median) QRF prediction values, and 75 <sup>th</sup> quantile prediction values.....	85
<b>Figure 4.3</b> Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 40% level of acceptable risk.....	87
<b>Figure 4.4</b> Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 50% level of acceptable risk.....	88
<b>Figure 4.5</b> Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 60% level of acceptable risk.....	90
<b>Figure 4.6</b> Frequency of shellfish harvesting area closures at Chester River, Fishing Bay, and James River affected by a range of risk levels from 0% to 100%.....	91

**Figure 4.7** Contours showing the proportion of Chesapeake Bay shellfish harvesting areas affected by a range of risk levels over the 30-week period in 2010..... 92

**Figure 4.8** Contours showing the frequency of shellfish harvesting areas at Chester River, Fishing Bay, and James River affected by a range of risk levels from 0% to 100% and a range of abundance thresholds from 0.5 GE/ml to 7 GE/ml over the 30-week period in 2010..... 94

## 1. CHAPTER 1: INTRODUCTION

### 1.1 Chesapeake Bay

The Chesapeake Bay is a large estuary in the eastern United States that stretches over 300 km from Havre de Grace, Maryland to Virginia Beach, Virginia, where it connects to the Atlantic Ocean. The Chesapeake Bay watershed covers 165,800 square kilometers of land located in six states, with over 150 major rivers and tributaries that feed freshwater into the estuary (Kemp et al., 2005). Three regions are typically used to describe the Bay based on its north-south salinity gradient: the oligohaline region in the Upper Bay (0-6), the mesohaline region in the Mid-Bay (6-18), and the polyhaline region in the Lower Bay (18-30) (Baird & Ulanowicz, 1989). The Chesapeake Bay has been used as a test bed for numerous studies relating to water quality because of an extensive and continuous *in situ* sampling network that spans multiple decades.

Increasing sediment and nutrient loads in tributary runoff have significantly hindered the ecological health and productivity of the Chesapeake Bay since the onset of agriculture in the region (Brush, 1989). The cooperation of the six watershed states, some of which have no physical boundary with the Chesapeake Bay, in addressing the estuary's health has had a contentious history ending with legislative action under the Clean Water Act in the form of Total Maximum Daily Load (TMDL) targets (Clean Water Act, 2017). The TMDL sets limits on the load of pollutants, specifically nitrogen, phosphorus, and sediment, in rivers and streams that eventually flow into the

Chesapeake Bay. These pollutants contribute to poor water quality, algae blooms, and turbid waters, which affect light penetration for submerged grasses, dissolved oxygen levels for aquatic life, and lower the appeal of recreational waters.

*In situ* sampling is an essential tool for monitoring water quality, ecological processes, and environmental health, but it can lack the temporal and spatial coverage needed in dynamic estuarine systems and environmental management applications. While the time and cost of obtaining additional *in situ* samples can exceed available resources, satellite remote sensing has the potential to provide this higher temporal and spatial resolution data in coastal water bodies.

## **1.2 Satellite Remote Sensing in the Chesapeake Bay**

Satellite remote sensing in Chesapeake Bay waters is mainly achieved using techniques from ocean color remote sensing. Ocean color satellite sensors measure the upwelling visible radiation at the top of the atmosphere that emanates from water bodies. Visible light is the only electromagnetic radiation that can penetrate through a significant portion of the water column, and the absorption and scattering of this radiation is dependent on the constituents in the water column (Brown et al., 2007). Therefore, the radiation scattered out of the water and towards the satellite sensor can be used to infer water column properties. The upwelling water-leaving radiance only accounts for about 1% of the downwelling irradiance, so ocean color sensors measure in narrow, discrete wavelengths that can be used to both remove atmospheric affects



and distinguish between water characteristics (Brown et al., 2007).

Ocean color remote sensing techniques were originally developed for open ocean waters, where the main determinant of water color is phytoplankton concentration. However, the Chesapeake Bay contains more optically complex waters in which water color also depends on dissolved organic material, suspended sediment, detrital matter, and concentrated phytoplankton blooms (Morel & Prieur, 1977). In these waters, changes to the atmospheric correction and other data processing techniques can be necessary to retain pixels that would otherwise be masked out for open ocean studies because of their proximity to land or brightness (Bailey et al., 2010; Aurin et al., 2013).

Ocean pixels in a satellite sensor swath are distinguished from land pixels by the lack of upwelling near-infrared (NIR) radiance, due to water's strong absorption of this wavelength. However, turbid coastal waters can have near-infrared water-leaving radiance signals from the interaction of downwelling irradiance with suspended sediment particles. The use of shortwave infrared (SWIR) wavelengths, which are absorbed even in turbid waters, has instead been proposed for use in coastal areas and the Chesapeake Bay (Werdell et al., 2010).

Many coastal waters and estuaries have complicated geometries, and near-shore data retrieval can be as important as the data from open water regions for many applications such as oyster bed management, recreational water use, and seagrass bed habitats for aquatic life. While ocean color remote sensing for the open ocean

removes data near land due to the contamination from bright land pixels, coastal ocean color remote sensing studies attempt to keep as many of these pixels as possible by reducing the buffer around land pixels that is typically masked (Ondrusek et al., 2012; Aurin et al., 2013). This allows for satellite data retrieval not only in the Chesapeake Bay's Mainstem, but also in its tributaries and coastlines.

### **1.3 Satellite-Derived Total Suspended Solids**

Total suspended solids (TSS) is an important water quality parameter to monitor in estuarine systems due to its ecological, economic, and human health impacts. Suspended matter in the water column limits light penetration and radiation availability for phytoplankton and submerged aquatic vegetation growth and has been shown to have biological importance for predicting water-borne bacteria like *Vibrio parahaemolyticus* that pose a growing food safety concern (Caburlotto et al., 2010; Johnson et al., 2012; Davis et al., 2017). Increased sedimentation rates can also reduce the extent of benthic organism habits (Boyd, 2015).

It can be impractical to monitor TSS *in situ* at a resolution sufficient for studying sediment plume dynamics, forecasting algal blooms, or modeling pathogenic bacteria due to the costs and time involved necessary for *in situ* sampling campaigns. Therefore, remotely sensed TSS estimates can be a useful alternative. Previous studies in the Chesapeake Bay used the observed correlation between TSS and reflectances in the red and NIR wavelengths to derive TSS concentrations from ocean color sensors (Stumpf, 1988; Hu et al., 2004; Ondrusek et al., 2012; Shen et al., 2014; Hasan & Benninger, 2017). Fewer studies have explored the relationships between TSS

concentration wavelengths outside of the red and NIR ranges for TSS algorithm development in estuarine environments (Qiu, 2013; Sokoletsky et al., 2016).

#### **1.4 Satellite-Derived *Vibrio parahaemolyticus* Concentration**

*Vibrio parahaemolyticus* is a naturally occurring, gram-negative bacterium found globally in brackish waters like those in the Chesapeake Bay and is considered one of the most common causes of seafood-borne illnesses in the United States (Colwell et al., 1977; Scallan et al., 2011). Over the past few decades, the rate of *V. parahaemolyticus* illnesses in the United States has been gradually increasing, with particularly large upticks recorded in 2013 and 2018 (Newton et al., 2012; U. S. Centers for Disease Control and Prevention, 2019).

Studies have attribute this increase to rising sea surface temperatures and prolonged warm seasons in the mid-latitudes, as well of the arrival of new strains of the bacterium in these waters (Martinez-Urtaza et al., 2013; Baker-Austin et al., 2013; Baker-Austin et al., 2017). Historically, pre-harvest water temperature and post-harvest air temperature have been the main environmental determinants used to estimate *V. parahaemolyticus* presence and abundance in shellfish and surface water. However, studies using *in situ* sampling campaigns have shown that additional environmental and water quality parameters are useful in predicting where and when high concentrations of *V. parahaemolyticus* occur (DePaola et al., 2003; Zimmerman et al., 2007; Patra et al., 2009; Caburlotto et al., 2010; Johnson et al., 2010; Johnson et

al., 2012; Turner et al., 2014; Davis et al., 2017).

Satellite remote sensing has been suggested as a valuable tool for *V. parahaemolyticus* prediction and forecasting (Johnson, 2015). Previous studies have developed models to predict *V. parahaemolyticus* in surface waters and shellfish using only remotely sensed sea surface temperature (SST) (Phillips et al., 2007; Konrad et al., 2017). While some hydrodynamic models have begun to also include sea surface salinity (SSS) into *Vibrio* prediction models, there is a lack of remote sensing-based modeling that incorporates water quality products in addition to SST (National Centers for Coastal Ocean Science, 2017).

## 1.5 Dissertation Outline

The first chapter of this dissertation provides an introduction to satellite remote sensing in the Chesapeake Bay and how it can be used for two contemporary issue-related applications: (1) total suspended solids retrieval and (2) prediction of *Vibrio parahaemolyticus*. The second chapter presents the evaluation and improvement of satellite retrieval algorithms for total suspended solids (TSS) concentration using additional spectral information, combined with advanced statistical and machine learning models. The third chapter evaluates prediction and forecasting models for *V. parahaemolyticus* concentration in Chesapeake Bay surface waters using satellite-derived water quality variables that have not yet been utilized in remote sensing-based models. The fourth chapter presents Quantile Regression

Forests as a tool for providing decision-makers with prediction uncertainty, using remotely sensed *V. parahaemolyticus* abundance models in the Chesapeake Bay as a case study. Finally, the fifth chapter concludes the dissertation work. The dissertation contributes methodology and insights for the improvement of coastal satellite remote sensing applications in the Chesapeake Bay that include environmental health for aquatic life, food safety, and science communication.

## 2. CHAPTER 2: CAN MULTISPECTRAL INFORMATION IMPROVE REMOTELY SENSED ESTIMATES OF TOTAL SUSPENDED SOLIDS? A STATISTICAL STUDY IN CHESAPEAKE BAY<sup>1</sup>

### ABSTRACT

Total suspended solids (TSS) is an important environmental parameter to monitor in the Chesapeake Bay due to its effects on submerged aquatic vegetation, pathogen abundance, and habitat damage for other aquatic life. Chesapeake Bay is home to an extensive and continuous network of *in situ* water quality monitoring stations that include TSS measurements. Satellite remote sensing can address the limited spatial and temporal extent of *in situ* sampling and has proven to be a valuable tool for monitoring water quality in estuarine systems. Most algorithms that derive TSS concentration in estuarine environments from satellite ocean color sensors utilize only the red and near-infrared bands due to the observed correlation with TSS concentration. In this study, we investigate whether utilizing additional wavelengths from the Moderate Resolution Imaging Spectroradiometer (MODIS) as inputs to various statistical and machine learning models can improve satellite-derived TSS estimates in the Chesapeake Bay. After optimizing the best performing multispectral model, a Random Forest regression, we compare its results to those from a widely used single-band algorithm for the Chesapeake Bay. We find that the Random Forest model modestly outperforms the single-band algorithm on a holdout cross-validation dataset and offers particular advantages under high TSS conditions. We also find that both methods are similarly generalizable throughout various partitions of space and time. The multispectral Random Forest model is, however, more data intensive than the

---

<sup>1</sup>DeLuca, N. M., Zaitchik, B. F., Curriero, F. C. (2018). Can Multispectral Information Improve Remotely Sensed Estimates of Total Suspended Solids? A Statistical Study in Chesapeake Bay. *Remote Sensing*, 10, 1393.

single band algorithm, so the objectives of the application will ultimately determine which method is more appropriate.

## 2.1 Introduction

Total suspended solids (TSS) is an important parameter to monitor in estuarine systems due to its ecological, economic, and human health impacts. Suspended solids in a water column reduce light and other radiation availability for phytoplankton and submerged aquatic vegetation (SAV) growth, while increased sedimentation rates reduce benthic organism habitability. It has also been shown that the volume of suspended particles in water may have biological importance for predicting incidence and abundance of pathogenic bacteria like *Vibrio parahaemolyticus* (Caburlotto et al., 2010; Johnson et al., 2012; Davis et al., 2017). Forms of suspended particles that contribute to TSS concentrations include sediment, detrital matter, and microorganisms. The type of sediment and organics in a water body can vary widely by region because they are introduced into the water column through processes that include watershed inputs, resuspension of bottom sediments, and ecological productivity.

It can be impractical to monitor TSS *in situ* at a resolution sufficient for studying sediment plume dynamics, monitoring and forecasting algal blooms, or modeling pathogenic bacteria, due to both the cost and time involved in *in situ* sampling and to the sometimes prohibitive logistics of installing dense monitoring networks. Remotely sensed TSS estimates can be a useful alternative. Many satellite-derived and *in situ*-measured radiance algorithms for estuarine TSS retrieval have been developed based on the observed correlation between TSS and reflectances in the red and near-infrared (NIR) wavelengths (Stumpf, 1988; Hu et al., 2004; Wang et al., 2010; Chen et al., 2011; Zhao et al., 2011; Ondrusek et al., 2012; Doxaran et al., 2012; Chen et al., 2013; Feng et

al., 2014; Shen et al., 2014; Dogliotti et al., 2015; Han et al., 2016; Hasan & Benninger, 2017). Use of these bands is a particular advantage for studies that use the Moderate Resolution Imaging Spectroradiometer (MODIS), due to the higher spatial resolution of those bands on MODIS. This can be useful in geographically complex estuarine environments with small-scale dynamics and regions of interest near the shoreline. Many of these TSS algorithms are tuned to a specific site or region, since differences between water bodies in inorganic and organic particle types and sizes change the inherent optical properties (IOPs) of the water (Stumpf & Pennock, 1989; Tzortziou et al., 2006).

Some TSS algorithms employ a switching function to swap between the red and NIR bands in accordance with some radiance threshold in order to account for a larger range of TSS values (Dogliotti et al., 2015; Han et al., 2016). Other ocean color remote sensing studies have found that two-band differencing approaches are effective algorithms for similar applications (Mitchell et al., 2017). Fewer studies have explored relationships between TSS concentration and optical properties in wavelengths outside of the red and NIR ranges for TSS algorithm development in estuarine environments (Qiu, 2013; Sokoletsky et al., 2016). This has historically been due to both the limits of spatial resolution in other visible bands and the non-negligible scattering from phytoplankton pigments and water at wavelengths less than 580-nm (Smith & Baker, 1978; Bukata et al., 1995). However, the use of shorter wavelengths may be beneficial for estimation of low TSS concentrations due to the higher signal-to-noise ratio that offers higher sensitivity to suspended particles than longer wavelengths provide.

Chesapeake Bay in the eastern United States has been used as a test bed for numerous studies relating to water quality because of an extensive and continuous *in situ* sampling network that spans multiple decades. It is the largest estuary in the United States, with a 165,800 square kilometer (64,000 square mile) watershed including six



states (Kemp et al., 2005). Over 150 major rivers and tributaries feed freshwater from the watershed into the estuary, the largest of which is the Susquehanna River in the northernmost reaches. Increasing sediment and nutrient loads in tributary runoff have significantly hindered the ecological health and productivity of the Chesapeake Bay since the onset of agriculture in the region (Brush, 1989). More recently, continued degradation of the estuary's health has prompted legislative action under the Clean Water Act in the form of Total Maximum Daily Load (TMDL) targets (Clean Water Act, 2017). Remote sensing can aid *in situ* monitoring efforts to follow improvements or declines in Chesapeake Bay ecosystem health influenced by sediment loads.

This study uses the Chesapeake Bay as a test case to investigate whether remotely sensed TSS estimates can be improved by utilizing additional MODIS Aqua bands alongside the commonly used red and NIR bands. Eight statistical and machine learning models, using 11 of MODIS Aqua's visible and NIR bands as predictor variables, are evaluated for their prediction performance using Chesapeake Bay Program (CBP) *in situ* TSS measurements over years 2003 to 2016. TSS predictions from the best performing model are then compared to those from a single-band algorithm that is widely used for remotely sensed TSS retrieval in the Chesapeake Bay. Utilization of additional sensor bands can potentially improve satellite TSS retrievals in optically complex estuarine systems like the Chesapeake Bay. At the same time, highly multispectral sensors are expensive to build and to fly relative to sensors with fewer bands, so it is useful to understand if and how the addition of spectral bands contributes to environmental monitoring of TSS and other parameters.

## 2.2 Materials and Methods

### 2.2.1 Data Description

#### 2.2.1.1 Satellite data processing

Daily Level 1A MODIS Aqua (R2014.0.2) ocean color products for the Chesapeake Bay were downloaded from NASA's ocean color archive (<http://ocean.color.gsfc.nasa.gov/>) for years 2003 through 2016 for statistical and machine learning model analyses, and 2017 for mapped comparisons (NASA Ocean Color Web, 2017). All images were batch processed from Level 1 to Level 2 with NASA's SeaDAS 7.4 software using the standard iterative NIR atmospheric correction at 1-km resolution (Bailey et al., 2010). Instead of the default cloud mask we used the 2130 nm band for cloud detection with a threshold albedo of 0.018 (Aurin et al., 2013). We also turned off the Level 2 high light mask so that pixels in which the NIR bands saturate or nearly saturate, which is common for regions of high turbidity, would be included in the output (Ondrusek et al., 2012; Aurin et al., 2013). The Level 2 stray light mask removes pixels adjacent to bright objects such as clouds and land due to possible light contamination. However, the default setting in SeaDAS may be too conservative for an estuarine system like the Chesapeake Bay. To avoid removing crucial areas around clouds or near shorelines, we decreased the SeaDAS stray light mask to a 3 by 3 array (Aurin et al., 2013). This procedure produced all of the Level 2 MODIS data, including remote sensing reflectance ( $R_{rs}$ ) and normalized water leaving radiance ( $nL_w$ ) products, used in the subsequent analyses (Table 2.1).

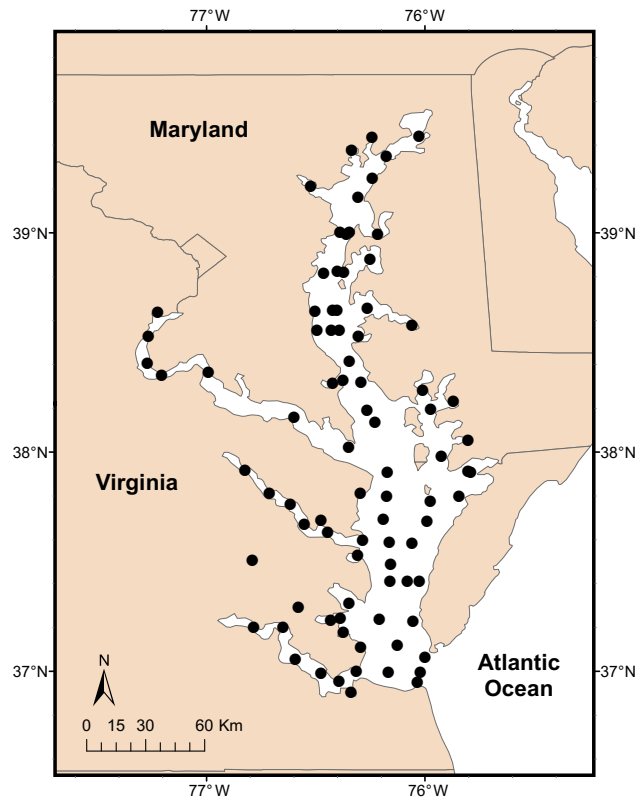
#### 2.2.1.2 *In situ* measurements

*In situ* TSS measurements used in this study for years 2003 to 2016, and subsequently 2017 for mapped comparisons, were downloaded from the Chesapeake

**Table 2.1.** Summary of variables used in statistical and machine learning model development and single-band algorithm calculation for this study.

<b>Variable</b>	<b>Mean</b>	<b>Standard Deviation</b>	<b>Maximum</b>	<b>Minimum</b>
<i>In situ</i> TSS (mg/L)	8.50	5.41	47.00	2.18
nL <sub>w</sub> (645) (μW/cm <sup>2</sup> /nm/sr)	0.7127	0.5240	3.4386	0.0683
Rrs_412 (sr <sup>-1</sup> )	0.0021	0.0021	0.0239	$8.71 \times 10^{-10}$
Rrs_443 (sr <sup>-1</sup> )	0.0030	0.0019	0.0223	0.0002
Rrs_469 (sr <sup>-1</sup> )	0.0039	0.0020	0.0215	0.0002
Rrs_488 (sr <sup>-1</sup> )	0.0044	0.0021	0.0192	0.0004
Rrs_531 (sr <sup>-1</sup> )	0.0068	0.0029	0.0189	0.0013
Rrs_547 (sr <sup>-1</sup> )	0.0075	0.0032	0.0212	0.0017
Rrs_555 (sr <sup>-1</sup> )	0.0074	0.0031	0.0209	0.0016
Rrs_645 (sr <sup>-1</sup> )	0.0045	0.0033	0.0217	0.0004
Rrs_667 (sr <sup>-1</sup> )	0.0036	0.0030	0.0205	0.0002
Rrs_678 (sr <sup>-1</sup> )	0.0036	0.0029	0.0205	0.0001
Rrs_859 (sr <sup>-1</sup> )	0.0009	0.0009	0.0079	$2.00 \times 10^{-6}$

Bay Program's online water quality database (<http://datahub.chesapeakebay.net>) (Chesapeake Bay Program, 2017). This online database is a collection of bi-monthly water quality measurements taken by state and federal agencies at predetermined stations throughout the Chesapeake Bay (Figure 2.1). *In situ* TSS measurements in this database are determined according to the procedure from U.S. EPA method 160.2. Only the measurements taken within the top 1 m of the water column were used for these analyses (Ondrusek et al., 2012; Urquhart et al., 2012).



**Figure 2.1.** Map of Chesapeake Bay estuary showing the 86 Chesapeake Bay Program measurement stations (black dots) used in the satellite-*in situ* matchup dataset in this study.

#### 2.2.1.3 Satellite-*in situ* matchups

*In situ* measurements were matched with same-day processed MODIS pixels within 250 meters of the *in situ* station's coordinates using a geographical distance matrix in the R package “fields” of the R Statistical Computing Environment (Nychka et al., 2015; R Core Team, 2013). Over the 16-year period, we created a dataset of 1360 satellite-*in situ* matchups found on 490 unique days. Statistics for this matchup dataset are found in Table 2.1. The 490 days include representative data from all seasons and years, giving us a comprehensive dataset from which to conduct our analyses. There is thought to be no more than a few hours between satellite overpass and *in situ* sampling times for most matchups because Chesapeake Bay Program

sampling for 2003–2016 occurs most frequently during late-morning (Figure S2.1) while MODIS overpasses occur in the early afternoon. TSS can change within a few hours under storm conditions and in smaller tributaries, but the temporal proximity of the matchups is adequate under most weather conditions and away from the shore. It is also consistent with the matchup criteria used in previous studies (Ondrusek et al., 2012). Our satellite-*in situ* dataset also includes eighty-six unique matchup locations that cover a large portion of the Chesapeake Bay spatially, including both the Mainstem and several large tributaries (Figure 2.1).

Where one or more MODIS bands are saturated in our matchup dataset due to high surface reflectance or incompatibilities with the atmospheric correction method, missing values are present in the vectors of predictor variables. Because it is necessary to omit missing values in order to run many of our modeling approaches, all satellite-*in situ* matchups where these missing values occur were excluded from the dataset. This trimming of missing predictor values resulted in 36% of the raw satellite-*in situ* matchup dataset ( $n = 2111$ ) being excluded from the final matchup dataset ( $n = 1360$ ) used for analyses. After this incomplete data was removed (including *in situ* TSS values up to 137.0 mg/L), the range of *in situ* measured TSS used in our analyses was limited to 2.18 to 47.0 mg/L (Table 2.1).

## 2.2.2 Methods

### 2.2.2.1 Statistical Methodology

Following a procedure outlined by Urquhart et al. (2012), eight statistical and machine learning models were evaluated for empirically estimating total suspended solids in the Chesapeake Bay using satellite-derived ocean color products. The 11 MODIS remote-sensing reflectances ( $R_{rs}$ ) used as predictor variables in these models

are summarized in Table 2.1. We randomly split our dataset of 1360 satellite-*in situ* matchups into 80% training data and 20% holdout data for cross-validation. We evaluated the predictive performance of our models using three metrics—mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). The absolute error gives equal weight to all errors while the two square error metrics give more weight to larger errors or outlier points, with RMSE having the same units as MAE for better comparison. All statistical and machine learning modeling and subsequent computations were done in R Statistical Computing Environment software (version 3.3.2) (R Core Team, 2013).

#### 2.2.2.2 Statistical and Machine Learning Models

Eight statistical and machine learning models were chosen for their ability to regress non-parametric, high dimensional data with a continuous response variable. The remote-sensing reflectances ( $R_{rs}$ ) from 11 MODIS-Aqua bands used for ocean color and land remote sensing are used as the predictor variables in these models, while TSS concentration is the response variable.

We included three types of decision tree regression models: Classification and regression tree (CART) (Breiman et al., 1984), Bayesian additive regression tree (BART) (Chipman et al., 2010), and Random Forest (RF) (Breiman, 2001). These tree-based models follow rules at data-defined nodes in the covariates to predict output of the response variable. The RF model uses a large number of trees, thereby creating a “forest” and determines output based on the most commonly used decision path for the dataset.

A generalized linear model (GLM) modifies the standard ordinary least squares linear regression model by adding a link function that can account for non-

normal distributions of response data (Nelder & Wedderburn, 1972). We use a logarithmic link function for our GLM based on the log-normal distribution of TSS measurements in our dataset. Our generalized additive model (GAM) (Hastie & Tibshirani, 1986) works similarly to the GLM, but accounts for possible non-linear effects of the predictors on the response by first using a smoothing function on the predictor variables. The Multivariate Adaptive Regression Spline (MARS) (Friedman, 1991) model is also similar to a GLM, but the number and type of link functions are automatically determined for the given dataset.

Neural networks (NN) take their name from their resemblance to the interactions of neurons and synapses in the brain. A layer of input data consisting of predictor variable data are sent to a predetermined number of hidden layers through connections, for which weights are calculated. When the sum of a connection's weights reaches a given threshold, the connection "fires" like a synapse and continues to the output data layer (Lee et al., 1992).

Support Vector Machines (SVM) find a regression fit by using a scaled-linear contribution from data whose residuals fall outside of a user-specified threshold, while data whose residuals are within the threshold do not contribute to the fit (James et al., 2013).

All 11 predictor variables (MODIS-Aqua bands) are retained in the training phase of model development except for in two models. The CART model only finds 645-nm, 667-nm, and 412-nm to be significant predictor variables for tree splitting, while the MARS model excludes the 443-nm, 488-nm, 547-nm and 667-nm predictors. Forcing these models to use all 11 predictors presents the risk of overfitting, which would decrease their performance on the holdout cross-validation dataset.

Prediction errors from the eight models (models fit on the training set and predictions generated for the hold out sample) were compared to each other and to a mean statistical null model, which was a model containing only the mean of the TSS response variable for all predictions. P-values were calculated for each pair-wise comparison for MAE and MSE and corrected for multiple comparisons using the Holm method to test for statistical significance (Holm, 1979).

TSS predictions from the best performing model were then compared to those from the optically based, single-band algorithm presented in Ondrusek et al. (2012) (referred to as O-2012 henceforth) for the same satellite-*in situ* matchups. The O-2012 algorithm uses MODIS normalized water leaving radiance ( $nL_w$ ) at 645-nm and is commonly used for estimating TSS in the Chesapeake Bay. It was derived using *in situ*-measured optical data from only the mid-Chesapeake Bay in 2008 within a few hours of satellite overpass. The authors fit the following 3<sup>rd</sup> order polynomial function to their data to relate *in situ*  $nL_w$  (645) to 35 *in situ* total suspended matter samples ranging from 4.50 to 14.92 mg/L:

$$TSM(mg\ L^{-1}) = 3.8813(nL_w(645))^3 - 13.822(nL_w(645))^2 + 19.61(nL_w(645)) \quad (1)$$

The algorithm was validated using 270 matchups between MODIS pixels and Chesapeake Bay Program *in situ* TSS data from 2009 with concentrations up to 100 mg/L. The authors found a mean percent difference of -4.2% and a mean absolute percent difference of 36%.

### 2.2.2.3 Geographic and Temporal Cross-Validation

A common criticism of empirical remote sensing models is that they are not generalizable throughout space and time when they are trained on a comprehensive



dataset. In order to test the ability of the model to deal with variable sediment distributions between the Mainstem of the Chesapeake Bay and its tributaries as well as missing data due to cloud cover, we trained our highest performing model on one region or time period and tested its performance on another region or time period. To do this, we divided our satellite-*in situ* matchup dataset into various training datasets based on latitude, longitude, region and time of year. Dataset partitions include East and West based on a selected longitude, North and South based on a selected latitude, Mainstem and tributaries of the estuary, and the high and low flow discharge seasons. We compared the MAE, MSE, and RMSE of our best model on the corresponding holdout datasets to that of O-2012 for the same holdout datasets to test our model's ability to predict TSS in the Chesapeake Bay as successfully as a single-band, optically-based algorithm.

## **2.3 Results and Discussion**

### *2.3.1 Model Comparison*

All eight statistical and machine learning models in this study outperformed the mean statistical null model in MAE and RMSE (Table 2.2). They were also statistically significant in MAE compared to the mean statistical null model based on a Holm multiple comparisons correction. This suggests that our eight models provide more information than assuming the mean of the dataset.

Of the eight models tested in this study, the Random Forest had the lowest MAE of 2.42, followed by the SVM with MAE of 2.57, the GAM with MAE of 2.64, and the GLM with MAE of 2.69 (Table 2.2). The RF model also had the best square error predictive accuracy with RMSE of 4.36, followed by the GAM with RMSE of 4.53, the BART with RMSE of 4.66, and MARS with RMSE of 4.69 (Table 2.2).

In order to better understand how the RF model predicts TSS in our holdout dataset, we looked at each predictor’s contribution to the model through partial dependence plots (Friedman, 2001). Partial dependence plots are valuable tools for interpreting predictive machine learning models like a Random Forest in which the relationships between predictors and response are not intuitive. They allow visualization of the amount of change in predicted response a given predictor variable produces when all other predictors in the model are averaged.

**Table 2.2.** Holdout mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) for all models evaluated in the study for the 20% holdout validation dataset. All models significantly outperform the mean statistical null model at  $p < 0.05$  in both MAE and MSE, adjusted for multiple comparisons using the Holm correction.

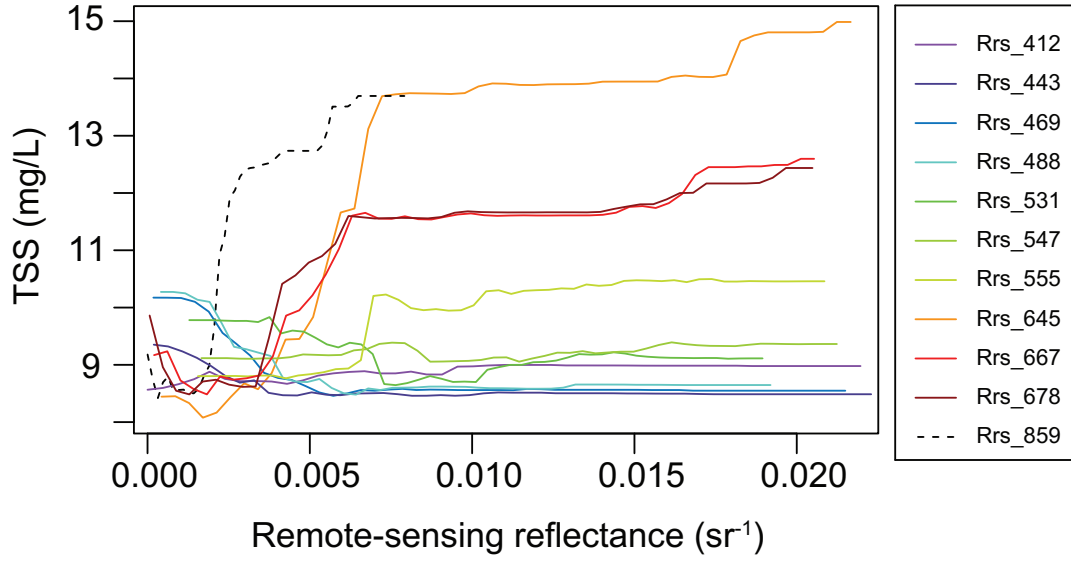
	<b>RF</b>	<b>GAM</b>	<b>GLM</b>	<b>NN</b>	<b>MARS</b>	<b>CART</b>	<b>BART</b>	<b>SVM</b>	<b>Mean</b>
MAE	2.42	2.64	2.69	2.76	2.74	2.87	2.73	2.57	4.01
MSE	19.04	20.49	22.43	24.32	21.98	23.22	21.73	22.19	35.36
RMSE	4.36	4.53	4.74	4.93	4.69	4.82	4.66	4.71	5.95

In order to better understand how the RF model predicts TSS in our holdout dataset, we looked at each predictor’s contribution to the model through partial dependence plots (Friedman, 2001). Partial dependence plots are valuable tools for interpreting predictive machine learning models like a Random Forest in which the relationships between predictors and response are not intuitive. They allow visualization of the amount of change in predicted response a given predictor variable produces when all other predictors in the model are averaged.

The partial dependence plots for all 11 MODIS bands in the RF model are shown in Figure 2.2. The red (645-nm, 667-nm, and 678-nm) and NIR (859-nm) bands play the largest role in predicting higher TSS values, which reflects the use of the red and NIR bands used in many TSS algorithms' development (Stumpf, 1988; Hu et al., 2004; Wang et al., 2010; Chen et al., 2011; Zhao et al., 2011; Ondrusek et al., 2012; Doxaran et al., 2012; Chen et al., 2013; Feng et al., 2014; Shen et al., 2014; Dogliotti et al., 2015; Han et al., 2016; Hasan & Benninger, 2017). Several of the blue (496-nm, 448-nm and 443-nm) and green (531-nm and 555-nm) bands appear to be important to predicting lower ranges of TSS. These results suggest that while most algorithms use the “red-ness” of sediment to remotely sense TSS concentrations, it could also be beneficial to use the “blue-ness” of clearer waters in developing algorithms to estimate TSS via satellite. The partial dependence plots also showed that the 412-nm, 443-nm, and 547-nm bands contribute very little to the RF model, shown by the lack of variation in TSS prediction over the respective ranges of remote-sensing reflectances (Figure 2.2). Pruning the RF model by excluding these three MODIS bands decreased the MAE and RMSE (2.38 and 4.30, respectively). The RF model referred to in the study henceforth is the pruned RF model that includes only 8 MODIS bands.

### *2.3.2 Comparison with Single-Band TSS Algorithm*

We compared our top performing model, a Random Forest that was pruned to minimize error in MAE, MSE and RMSE, to the O-2012 for the holdout validation dataset. The O-2012 algorithm was applied with its published coefficients, as that is the version commonly used in Chesapeake Bay applications. We also generated a version of the same polynomial form but with updated coefficients and without the y-intercept being forced through zero (referred to as O-2012(fit)), fit using the training and



**Figure 2.2.** Partial dependence plots for the 11 MODIS bands used as predictors in the Random Forest model before pruning.

evaluation datasets used for our other models:

$$\text{TSM}(\text{mg L}^{-1}) = 19.222(\text{nL}_w(645))^3 - 10.293(\text{nL}_w(645))^2 + 115.539(\text{nL}_w(645)) + 8.467 \quad (2)$$

This customized polynomial model showed an improvement over the standard O-2012 algorithm in our holdout validation test, which may be due to the larger range of TSS concentrations used to train our polynomial regression. However, the RF model using multiple bands still had better performance than the newly fitted O-2012(fit) using a single band (Table 2.3). Since O-2012 and O-2012(fit) show nearly identical performance and have similar behavior across geography and TSS concentrations, we focus on the original, published O-2012 algorithm through the rest of our analysis.

The RF model produces significantly lower errors from the O-2012 algorithm in both MAE and MSE (Table 2.3). The RF model also introduces less bias (mean error of  $-0.04$ ) in the holdout cross-validation than O-2012 (mean error of  $-0.54$ ). The RF model

**Table 2.3.** Mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) for the pruned Random Forest (RF) model, Random Forest model using only 645-nm (RF(645)), the O-2012 algorithm as published, and the O-2012 algorithm fitted for our dataset (O-2012(fit)) for the 20% holdout validation dataset. Asterisk (\*) indicates where RF model is statistically different than O-2012 at  $p < 0.05$ .

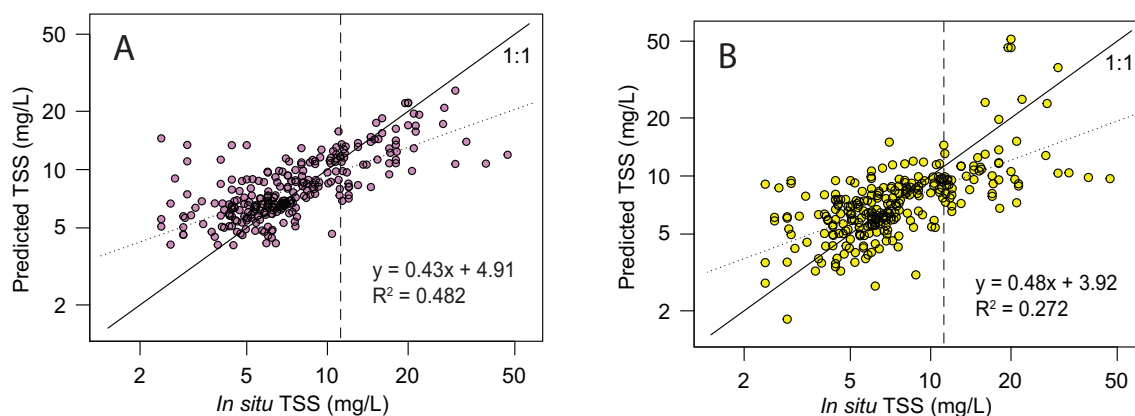
Model	MAE	MSE	RMSE
RF	2.38*	18.46*	4.30
RF(645)	2.76	20.81	4.56
O-2012	2.97	31.44	5.61
O-2012(fit)	2.71	21.77	4.67

introduces less bias (mean error of  $-0.04$ ) in the holdout cross-validation than O-2012 (mean error of  $-0.54$ ). The RF model's MAE accounts for 5.3% of the range in TSS values used in this study (2.4 to 47.0 mg/L), while O-2012's MAE accounts for 6.6%. The accuracy necessary for each user's application may indicate whether either, or both, methods have acceptable performance for TSS estimation.

In order to assess the effect of the algorithmic approach (machine learning versus polynomial regression) we also compared a Random Forest model using only the same remote sensing reflectance in 645-nm as in the O-2012 algorithm. The single-band RF model, referred to as RF(645), performed better in MAE (2.76 vs 2.97) and RMSE (4.57 vs. 5.61) than O-2012 and better in RMSE (4.56 vs. 4.67) than O-2012(fit) for the holdout cross-validation. This indicates that the application of machine learning alone provides some advantage over the polynomial method. The use of multiple bands provides further improvement (Table 2.3).

The one-to-one regressions for *in situ* measured TSS to predicted TSS for the pruned RF model and O-2012 algorithm show that higher TSS values are not predicted as well as lower TSS values for either prediction method (Figure 2.3). The pruned RF model consistently underestimates the higher range of TSS values, which may be due to the log-normal distribution of the response variable in our training dataset. More data in a higher TSS range in the training dataset may allow the model to better predict higher TSS values, but could come at the cost of not predicting the more frequently measured lower range as accurately. There is a wide spread in predictions for these high TSS events using O-2012.

To better investigate this lack of prediction skill in both the RF model and O-2012 algorithm at higher ranges of TSS, we compared the error for both models for holdout data above the 80th percentile (TSS > 11.3 mg/L, n = 55) (Figure 2.3). On high



**Figure 2.3.** Log-log plots showing one-to-one regressions of CBP *in situ* measured versus satellite-derived TSS from (A) the pruned Random Forest model and (B) the O-2012 algorithm. Solid black line is 1:1 line, dotted line is the linear regression and dashed line shows the cutoff for higher TSS range analyses.

TSS values, the RF model had a lower MAE of 5.16 than O-2012, with MAE of 8.33 (Table 2.4). The RF model also had a lower RMSE than O-2012, 8.24 vs. 11.71, respectively. The RF model's MAE and MSE error metrics are significantly lower than the O-2012 algorithm's error metrics to the 95% confidence interval (Table 2.4). The RF model and O-2012 have similar bias (mean error of 4.50 and 4.51, respectively) for the high TSS range predictions. These results suggest that the RF model is able to better predict TSS values farther from the mean than O-2012, although both methods produce larger errors for higher TSS ranges.

It is also important to compare how the two models perform below the 80th percentile (TSS < 11.3 mg/L, n = 217) because it represents more commonly measured TSS values in the Chesapeake Bay. We find that the RF model and O-2012 perform very similarly for MAE (Table 2.4). The O-2012 algorithm performs better than the RF model for MSE, 4.67 and 5.92 respectively (Table 2.4). However, neither MAE nor MSE from the two methods are significantly different at the 0.05 level for this lower TSS holdout validation data (Table 2.4).

**Table 2.4.** Mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) for the pruned RF model and O-2012 algorithm for holdout validation dataset above and below the 80th percentile. Asterisk (\*) indicates where RF model is statistically different O-2012 at  $p < 0.05$ .

	Above 80 <sup>th</sup> Percentile		Below 80 <sup>th</sup> Percentile	
	RF	O-2012	RF	O-2012
MAE	5.16*	8.33	1.67	1.62
MSE	67.97*	137.07	5.92	4.67
RMSE	8.24	11.71	2.43	2.16

It is also important to compare how the two models perform below the 80th percentile ( $\text{TSS} < 11.3 \text{ mg/L}$ ,  $n = 217$ ) because it represents more commonly measured TSS values in the Chesapeake Bay. We find that the RF model and O-2012 perform very similarly for MAE (Table 2.4). The O-2012 algorithm performs better than the RF model for MSE, 4.67 and 5.92 respectively (Table 2.4). However, neither MAE nor MSE from the two methods are significantly different at the 0.05 level for this lower TSS holdout validation data (Table 2.4).

### 2.3.3 Daily Satellite, *In Situ* Mapped Comparisons

In order to compare the performance of the RF and O-2012 methods in space, we predicted remotely sensed TSS using each method on data that was not included in either the training dataset or holdout validation dataset. Figure 2.4 shows three MODIS images from 2017 dates with good spatial coverage of Chesapeake Bay from three seasons: (1) March 22, 2017 in the spring, (2) June 28, 2017 in the summer, and (3) October 22, 2017 in the fall. *In situ* measurements from the same, previous, or subsequent day (in order to increase the number of comparison points) are also shown for comparison between remotely sensed and *in situ* measured TSS. Satellite-*in situ* matchups were determined by the same method outlined in Section 2.3.

Overall, Figure 2.4 shows that both the RF model and O-2012 capture the general structure of TSS across the Bay in these randomly selected matchups. However, relative performance in MAE, MSE, and RMSE depends on the date chosen. In general, O-2012 offers more consistent performance in the lower Mainstem for the three dates shown. In the upper Mainstem and tributaries, the RF model generally performs better than O-2012 for the summer and fall dates, but not for the spring date. The differences in performance

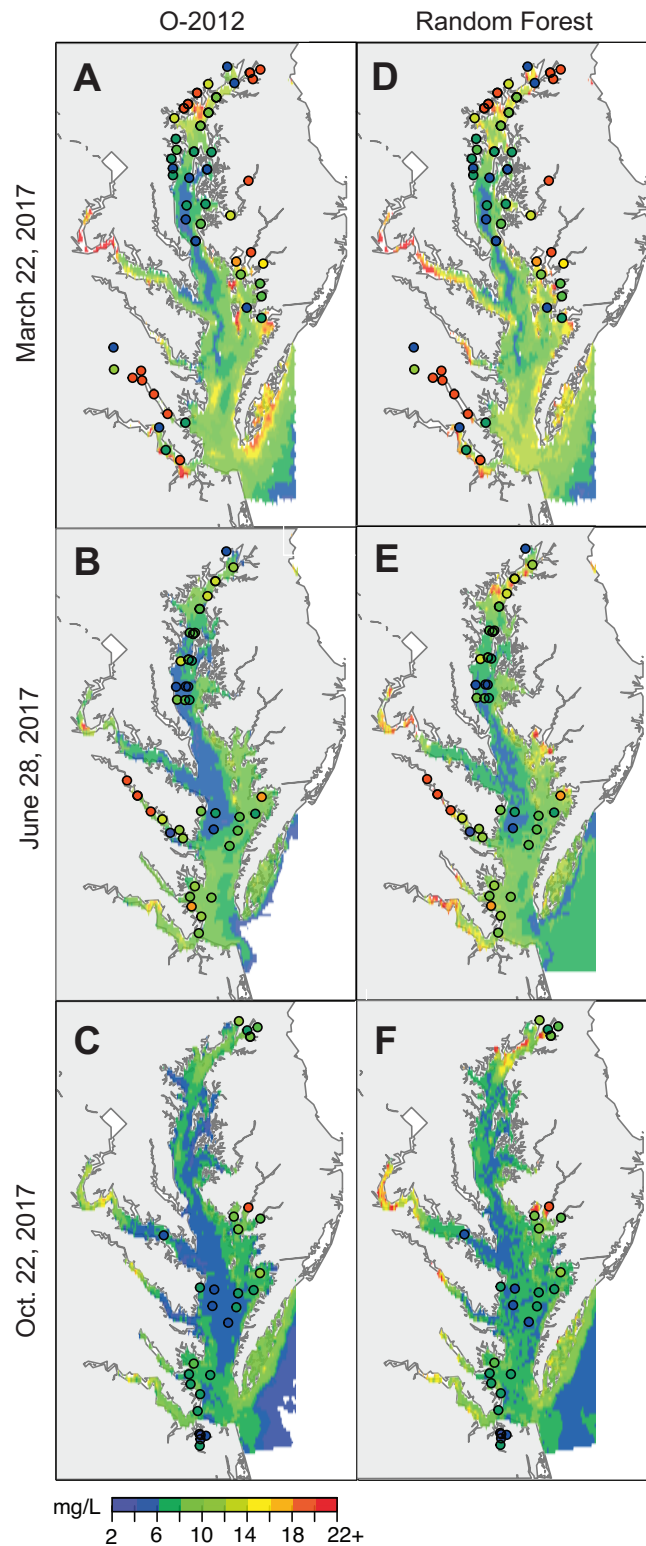


between the two methods are not statistically significant in MAE or MSE for any of the regions (upper Mainstem, lower Mainstem, and tributaries) on any of the dates.

#### *2.3.4 Geographic and Temporal Cross-Validation*

In order to test whether our top performing model was generalizable in predicting TSS throughout the Chesapeake Bay, we tested the performance of the RF model for various training and holdout datasets. We then compared the generalizability of the RF model to that of O-2012 for the same holdout datasets using MAE and MSE metrics. The naming convention works similarly to the following example: “East for West” indicates that the model is trained on a dataset encompassing stations only on the Eastern side of the Chesapeake Bay and the holdout validation dataset is comprised of the remaining stations on the Western side.

The results of the geographical and temporal cross-validation are presented in Table 2.5. Neither the RF model nor the O-2012 algorithm consistently outperforms the other in terms of MAE across the different validation scenarios. Based on MSE however, which gives more weight to larger errors, the RF model performs better than O-2012 in all but one scenario, West for East, and many of the differences between methods are statistically significant. Both methods were more accurate when predicting the East, South, and Mainstem sub-datasets than they were in other regions, and both also had higher skill for holdout predictions in the low discharge season. Both methods are generally similar overall in the accuracy of their predictions across various spatial and temporal partitions of the dataset, even though some differences have statistical significance. Again, the application may determine what level of accuracy or precision is needed in TSS estimations.



**Figure 2.4.** Mapped comparisons of daily remotely sensed TSS (mg/L) derived from O-2012 (A–C) and RF model (D–F) for 2017 dates not included in model training or holdout datasets. *In situ* measurement values shown in color-filled black circles.

## 2.4 Conclusions

This study investigates whether additional information can be gained from statistical and machine learning models that utilize multispectral MODIS information when predicting TSS in estuarine systems, using the Chesapeake Bay as a case study. Eight models using 11 MODIS bands, as well as a single-band algorithm, were evaluated for their performance on predicting TSS measurements taken *in situ* over a 14-year time period throughout the estuary. The Random Forest model performed best out of the eight models and the single-band algorithm on the holdout validation dataset. It also outperformed the single-band algorithm on the top 20th percentile of test data, but did not perform better on the lower 80th percentile. We found that both methods of TSS prediction were generalizable throughout space and time, with relative performance dependent on the error metric used for comparison.

The results of this study suggest that the single-band O-2012 algorithm is a valuable tool for estimating TSS via remote sensing in the Chesapeake Bay for general environmental applications. However, for applications where more accuracy or precision in waters with higher TSS concentrations may be needed, such as when modeling *Vibrio* bacteria (Davis et al., 2017), a statistical or machine learning model that utilizes additional MODIS bands could prove valuable. In estuarine systems, additional MODIS bands could provide information about suspended particles other than sediment that contribute to a TSS measurement, such as phytoplankton and detrital matter. Determining the supplementary bands to be added to statistical or machine learning models could vary from estuary to estuary based on the watershed characteristics, biology, and corresponding optical properties of each system. A recent study by Hasan et al. (2017) found that it is beneficial to determine the relationship between reflectance and TSS concentration separately for the Mainstem and tributaries of the Chesapeake Bay,

**Table 2.5.** Generalizability cross-validation results. Mean absolute error (MAE) and mean square error (MSE) for the pruned RF model and O-2012 algorithm. Naming scheme for cross-validation is as follows: “East for West” indicates model trained on East dataset and validated West dataset. Asterisk (\*) indicates O-2012 is statistically different than RF model at  $p < 0.05$ . The number of data points used in the training dataset for each scenario is indicated by  $n$ .

	Model	East for West (n = 604)	West for East (n = 756)	North for South (n = 581)	South for North (n = 779)	Mainstem for Tributary (n = 878)	Tributary for Mainstem (n = 482)	High for Low (n = 556)	Low for High (n = 804)
<b>MAE</b>	RF	3.32*	2.41*	2.70	3.66*	3.65 *	2.58*	2.62	2.90
	O-2012	3.50	2.08	2.68	3.12	3.89	2.31	2.75	3.04
<b>MSE</b>	RF	26.95*	14.37	21.71	25.37	33.24 *	15.23	17.00*	21.50 *
	O-2012	35.26	11.94	24.31	25.70	40.34	16.43	22.82	27.91

which may account for variability in the optical properties between the two estuarine features. This poorly characterized variability between regions of the estuary may be why a decision tree-based model like a Random Forest better predicts the high TSS values typically found in tributaries.

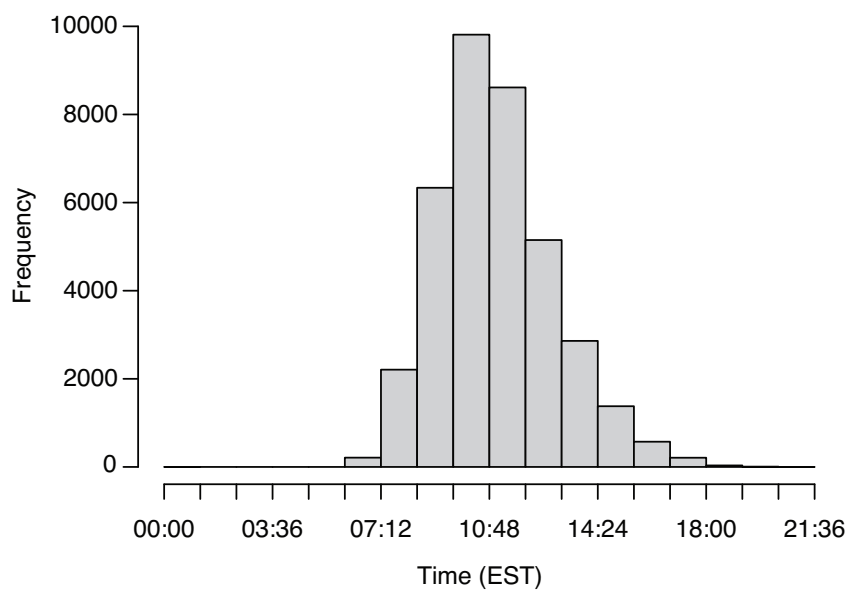
However, it is important to recognize the utility of a single-band, optically based algorithm like O-2012. These algorithms are generally easier to use for conventional applications of remotely sensed TSS estimates, and their physical basis is more straightforward (Chen et al., 2013). In a comprehensive quantitative review of published

TSS algorithms, Dorji et al. (2016) found that optically based algorithms generally perform better than empirical algorithms in waters with unknown composition.

Another noteworthy ability of the O-2012 algorithm and other red and NIR algorithms is that of having higher spatial resolution. While red and NIR algorithms can provide 250-meter resolution, using other MODIS wavelengths limits spatial resolution to 1-km. This could be a significant disadvantage to using other MODIS bands in a model when near-shore TSS estimates are important. However, recently launched ocean color sensors like the Ocean and Land Colour Instrument (OLCI) provide improved spatial resolution at 300-meter resolution bands outside of the red and NIR wavelengths. Therefore, this MODIS study could provide a foundation for multispectral TSS retrieval in higher spatial resolution using newer sensors.

Overall, our results suggest that the application of multispectral data using statistical and machine learning methods to estimate TSS in Chesapeake Bay may offer an increase in skill over standard single-band approaches. It may be particularly useful in regions that often experience higher TSS levels than the lower Mainstem region. However, the existing single-band O-2012 algorithm does perform reasonably well when compared to our best-performing multispectral RF model in Bay-wide evaluations, performing particularly well in lower TSS regions. The relative performance of the two models appears to be sensitive to sub-region, date, and the error metric being evaluated. For this reason, we conclude that the use of additional MODIS bands outside of the red and NIR wavelengths can be utilized to aid in satellite-derived TSS retrieval in estuaries like the Chesapeake Bay, but that the decision on whether adopt such an approach will depend on the objectives of the application.

**Figure S2.1** Histogram of sampling times for all Chesapeake Bay program TSS data at or above 1 meter depth for years 2003-2016. Times shown are Eastern Standard Time.



### 3. CHAPTER 3: EVALUATION OF REMOTELY SENSED PREDICTION AND FORECAST MODELS FOR *VIBRIO PARAHAEMOLYTICUS* IN CHESAPEAKE BAY<sup>2</sup>

#### ABSTRACT

Over the last decade, an increase of gastrointestinal illness due to *Vibrio parahaemolyticus* in the consumption of raw shellfish has been reported in multiple regions around the United States. Studies mainly attribute this increase to rising sea surface temperatures and prolonged warm seasons in the mid-latitudes. Historically, temperature has been the main environmental determinant used to predict *V. parahaemolyticus* concentrations in shellfish and surface water. However, studies using *in situ* sampling campaigns have shown that additional water quality parameters can be useful in predicting the bacterium. While the time and cost of obtaining *in situ* samples throughout the Chesapeake Bay at regular time intervals can exceed available resources, satellite remote sensing has the potential to provide predictions at higher temporal and spatial resolutions. This study uses satellite ocean color remote sensing and sea surface temperature (SST) from the Moderate Resolution Imaging Spectroradiometer (MODIS) to investigate the utility of remotely sensed information for *Vibrio parahaemolyticus* predictions in the Chesapeake Bay and whether additional remotely sensed information can improve predictions over conventional SST-based models. We find that the addition of remotely sensed salinity, total suspended solids, and chlorophyll-a improved presence and abundance predictions compared to SST-only models. Remote sensing reflectances also showed promise for *V. parahaemolyticus*

---

<sup>2</sup> DeLuca, N. M., Zaitchik, B.F., Guikema, S. D., Jacobs, J. M., Davis, B. J. K., Curriero, F. C. (In review) Evaluation of Remotely Sensed Prediction and Forecast Models for *Vibrio Parahaemolyticus* in Chesapeake Bay. *Remote Sensing of Environment*.

prediction, which could bypass the intermediary step of deriving water quality products from reflectances. Remotely sensed information from one week prior to the *in situ* *V. parahaemolyticus* measurements was also evaluated for its ability to forecast the bacterium, which could provide lead time for management decisions. The forecasts using additional ocean color products and remote sensing reflectances showed improvement over SST-based forecasts. The results of this study suggest that remote sensing could be a valuable tool to aid in higher resolution *V. parahaemolyticus* predictions and forecasts in the Chesapeake Bay, particularly when multiple water quality predictors are employed.

### **3.1 Introduction**

The Chesapeake Bay is the largest estuary in the United States, spanning over 300 km from its northern reach in Havre de Grace, Maryland to its southernmost point in Virginia Beach, Virginia. The estuary provides habitats that foster a prosperous seafood industry, with over 500 million pounds harvested every year (Chesapeake Bay Program, 2018A). The oyster industry in particular has made a large contribution to the region's economy and culture over the last century (Chesapeake Bay Program, 2018B). Along with the recent growth of the oyster aquaculture industry, an increase in the number of filter-feeding bivalves has the potential to improve water quality. Influence on water quality is an important consideration for a region with historically problematic sediment and nutrient pollution, which prompted legislation under the Clean Water Act (USEPA, 2011). However, the bivalves' filtering mechanism also permits the accumulation of microorganisms in concentrations up to 100 times that which is found in the surrounding waters (Froelich et al., 2017).



*Vibrio parahaemolyticus* is a naturally occurring, gram-negative bacterium found globally in brackish waters like those in the Chesapeake Bay (Colwell et al., 1977). There are both pathogenic and non-pathogenic strains of the bacterium, yet it is considered one of the most common causes of seafood-borne illnesses in the United States (Scallan et al., 2011). Infection typically occurs through the consumption of contaminated raw shellfish, causing gastroenteritis and more rarely septicemia.

The U.S. Centers for Disease Control reported that *V. parahaemolyticus* incidence increased 4-fold from 1996 to 2010 based on FoodNet surveillance (Newton et al., 2012). While the trend has continued to increase since 2010, U.S. water bodies experienced a particularly large spike in illnesses related to *V. parahaemolyticus* in both 2013 and 2018 with 147 and 187 cases reported, respectively (U. S. Centers for Disease Control and Prevention, 2019). Studies have attributed this positive trend to warming water temperatures, particularly in higher latitude water bodies, and the arrival of new strains of the bacterium in such waters (Martinez-Urtaza et al., 2013; Baker-Austin et al., 2013; Baker-Austin et al., 2017).

The increasing trend in incidence is expected to continue under the influence of anthropogenic climate change, and the spatial and temporal range in which *Vibrio* spp. related illnesses are found is anticipated to expand (Martinez-Urtaza et al., 2010; Jacobs et al., 2015; Muhling et al., 2017). Optimal conditions for the bacteria were previously found mainly in the Mid-Bay, Upper-Bay and western tributaries with concentrations peaking in July and August (Banaker et al., 2011). With Chesapeake Bay water temperatures projected to increase up to 5°C through the end of the 21<sup>st</sup> Century and extreme precipitation events projected to increase, these regions are likely to expand or shift and the timing of peak

concentrations is likely to extend into additional summer months (Jacobs et al., 2010; Banaker et al., 2011). Therefore, the ability to predict concentrated regions of *V. parahaemolyticus* is important for environmental managers and shellfish harvesters in order to reduce the risk of *Vibrio*-related illness (Davis et al., 2019). The recent surge in illnesses has also prompted interest in early warning systems, in which forecasted predictions would provide decision-makers and harvesters time to prepare for the implementation of preventative and control measures (Konrad et al., 2017).

While water temperature has been shown to drive most *V. parahaemolyticus* variability (Kaneko & Colwell, 1973; Caburlotto et al., 2010; Johnson et al., 2010; Johnson et al., 2012), recent work has improved the characterization of other environmental determinants of the bacterium in estuarine waters. Salinity is a frequently studied variable in relation to *V. parahaemolyticus* concentration in estuarine environments, but there is little consensus in the results. Studies have found both positive and negative associations with salinity and *V. parahaemolyticus*, while others have shown that there is an optimal range of salinity in which the bacteria thrive (DePaola et al., 2003; Zimmerman et al., 2007; Patra et al., 2009; Johnson et al., 2010; Davis et al., 2017). These differences are thought to be due to variations in the range of salinity investigated, where low salinity waters limit *V. parahaemolyticus* growth and higher salinity waters are typically outside of the temperature range of optimal growth (Davis et al., 2017).

*Vibrio* species are commonly associated with plankton density because the plankton provide the nutrient-rich surfaces to which the bacteria attach (Turner et al., 2014). Because chlorophyll pigments can be an indicator of plankton density, such pigment measures could be useful for *V. parahaemolyticus* prediction, particularly following algal blooms (Greenfield

et al., 2017). Algal blooms can also be an indicator of nutrient runoff, which introduces dissolved organic matter that has been found to proliferate *V. parahaemolyticus* (Thickman & Gobler, 2017). A lagged association of bacterium abundance with chlorophyll was previously suggested for further research (Davis et al., 2017).

The tendency for *V. parahaemolyticus* to occur in an attached state rather than free-floating also suggests that the volume of suspended matter in water could be an environmental determinant for the bacterium (Caburlotto et al., 2010; Johnson et al., 2012; Davis et al., 2017). It has been proposed that *Vibrio* spp. resides in estuarine sediments, and so the indication of resuspended sediments could also be a useful predictor for bacterial presence and abundance (Johnson et al., 2010; Davis et al., 2017).

The time and cost of obtaining *in situ* samples at sufficient spatial and temporal resolution for Chesapeake Bay-wide *V. parahaemolyticus* modeling can be prohibitive, and the laboratory measurements needed for some determinants make much of the data unavailable in real-time. Satellite remote sensing could be a valuable tool to complement *in situ* sampling and hydrological modeling for real-time *V. parahaemolyticus* detection and early warning forecasts (Johnson, 2015; Konrad et al., 2017). Previous remote sensing-based studies have developed linear regression models to predict *V. parahaemolyticus* concentration in both oysters and surface waters using only sea surface temperature (SST) (Phillips et al., 2007; Konrad et al., 2017). While some hydrodynamic models have begun to also include sea surface salinity (SSS) into their *Vibrio* prediction models, fewer remote sensing-based models incorporate the additional environmental predictors that have shown importance in the *in situ*-based studies (Tyberghein et al., 2012; National Centers for Coastal Ocean Science, 2017). Although apparent optical properties of waters, like remote sensing

reflectances ( $R_{rs}$ ), are usually converted into water quality products using bio-optical or empirical algorithms, models that use the remote sensing reflectances themselves as predictors could bypass this intermediary step.

This study uses an extensive *in situ* sampling campaign conducted across Chesapeake Bay to evaluate remote sensing-based models for *V. parahaemolyticus* presence and abundance. Accurate pathogen predictions are important for the Chesapeake Bay, where a rising incidence rate is anticipated as water temperatures increase over the next several decades. Satellite data from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS) aboard the Aqua spacecraft (2002 –) are used to investigate the following questions: 1) Do environmental parameters beyond SST improve remotely sensed predictions of *V. parahaemolyticus*? 2) Can *V. parahaemolyticus* be predicted using raw remote sensing reflectances instead of converting them into environmental determinants? 3) Do remote sensing-based *V. parahaemolyticus* prediction models perform as well as models that use *in situ* environmental measurements? 4) Can we forecast *V. parahaemolyticus* presence and abundance one week in advance using remote sensing-based models?

## 3.2 Data Description

### 3.2.1 *In situ* measurements

Methodology for the water sampling campaign and *V. parahaemolyticus* analysis used in this study is described in detail in Jacobs et al. (2014) and Davis et al. (2017). A brief summary of these methods as they pertain to this study is described here. A total of 1,523 surface water (0.5 m depth) samples were collected for *V. parahaemolyticus* determination at 148 monitoring stations located throughout the Mainstem and tributaries

of the Chesapeake Bay, which coincide with the decades-long record of Chesapeake Bay Program sampling stations, during three representative seasons (April - spring, July - summer, and October - autumn) in years 2007 to 2010. Multiple water quality measurements, following Chesapeake Bay Program protocols, were taken *in situ* alongside the collection of each water sample (U.S. Environmental Protection Agency, 1996). Procedures for *in situ* water temperature, chlorophyll-a, and total suspended solids measurements used in this study can be found in Jacobs et al. (2014). Abundance of total *V. parahaemolyticus* in units of genomic equivalents of colony-forming units (CFU) per milliliter (GE/ml) was determined for each water sample via quantitative polymerase chain reaction (qPCR). A large proportion of water samples (80.4%) contained abundances below the limit of detection and for this study was treated as an absence of *V. parahaemolyticus*. This dataset is unique because it provides comprehensive coverage of *V. parahaemolyticus* measurements throughout Chesapeake Bay along with the corresponding *in situ* water quality measurements for each sampling location and date.

### 3.2.2 Satellite data processing and algorithms

Uncalibrated Level 1A MODIS Aqua (R2018) products for years 2007 through 2010 were downloaded from NASA's ocean color archive (<http://ocean.color.gsfc.nasa.gov/>) and extracted for the Chesapeake Bay region. Images were batch processed to calibrated and binned Level 3 (weekly composite files) for both ocean color and sea surface temperature (SST) products using NASA's SeaDAS 7.5 software. The standard iterative NIR atmospheric correction was chosen for ocean color processing in this study because of the potential signal-to-noise limitation of MODIS' SWIR bands

addressed in Bailey et al. (2010). Changes to the default l2gen processing settings for ocean color products include 1) the 2130-nm band used for cloud detection with threshold albedo 0.018, 2) the high light mask turned off, and 3) the reduction of the stray light mask to a 3x3 array (Aurin et al. 2013). These changes were shown to increase the number of valid pixels retrieved in coastal waters, specifically in the Chesapeake Bay, where the default masking of pixels due to the brightness of turbid waters and adjacency to land can be too conservative (Aurin et al., 2013). The sea surface temperature product was processed using the standard SST masks for invalid pixels provided in SeaDAS.

This study uses 9 products at 1-km resolution derived from the MODIS processing procedure described above. Ocean color products include chlorophyll-a, total suspended solids (TSS), salinity, and four remote-sensing reflectances –  $R_{rs}(412\text{-nm})$ ,  $R_{rs}(555\text{-nm})$ ,  $R_{rs}(678\text{-nm})$ , and  $R_{rs}(859\text{-nm})$ . These four particular remote sensing reflectance wavelengths were chosen using a variance inflation function to reduce the correlation between them in order to avoid overfitting in the modeling efforts (James et al., 2014). The algorithms used to derive chlorophyll-a (using a regionally tuned OC3 algorithm for the Chesapeake Bay), TSS, and sea surface salinity are described in Werdell et al. (2007), DeLuca et al. (2018), and Urquhart et al. (2012), respectively. The sea surface temperature product is obtained by transforming MODIS thermal infrared radiances into degrees Celsius ( $^{\circ}\text{C}$ ) using the Planck function.

### 3.2.3 Satellite-in situ matches

*In situ* measurements were matched to MODIS pixels within 1-km distance using a geographical distance matrix in the R package “fields” (Nychka et al., 2016, R Core

team, 2013). Where multiple MODIS pixels were paired to one *in situ* measurement, the mean value of the surrounding pixels was used. Records were dropped when no pixel was recorded within 1-km of the *in situ* sampling location due to cloud cover or the limit of spatial resolution of this sensor in tributary and coastal regions. Weekly Level 3 MODIS data that were collected during the same 8-day week as the *in situ* measurement was taken were used to create a “Same Week” matchup dataset, while weekly Level 3 MODIS data collected the 8-day week prior to the *in situ* measurement were used to create the “1-Week Forecast” matchup dataset. The binned 8-day composites were used as a Same Week dataset in place of daily Level 2 data in order to increase the number of satellite-*in situ* matchups for a robust statistical analysis. This was particularly necessary for matchups during summer months when *V. parahaemolyticus* concentrations increase and cloud cover can severely limit MODIS data retrieval over the Bay (Urquhart et al., 2013).

All sampling months and years are represented in the final Same Week (n = 572) and 1-Week Forecast (n = 605) datasets. The satellite-*in situ* match-ups are located at 84 stations throughout the Mainstem of the Chesapeake Bay and within several large tributaries (Figure 3.1). Match-ups where one or more MODIS band was saturated were excluded from the final datasets in order to omit missing values before running the models.

*Vibrio parahaemolyticus* is heavily zero-inflated (with non-detects being the majority class) in the final Same Week and 1-Week Forecast datasets, with detected presence comprising only 13% (n = 77) of the Same Week dataset and only 16% (n = 98) of the 1-Week Forecast dataset. There were 22 (31) detected *V. parahaemolyticus*

samples from 2007, 13 (13) from 2008, 6 (10) from 2009, and 36 (44) from 2010 in the Same Week (1-Week Forecast) dataset. The number of match-ups and the percentage of presence in both final datasets are influenced by the coverage of optically valid MODIS pixels in the weekly composites and how that coverage corresponds to *in situ* measurements taken in the 8-day periods.

### 3.3 Methods

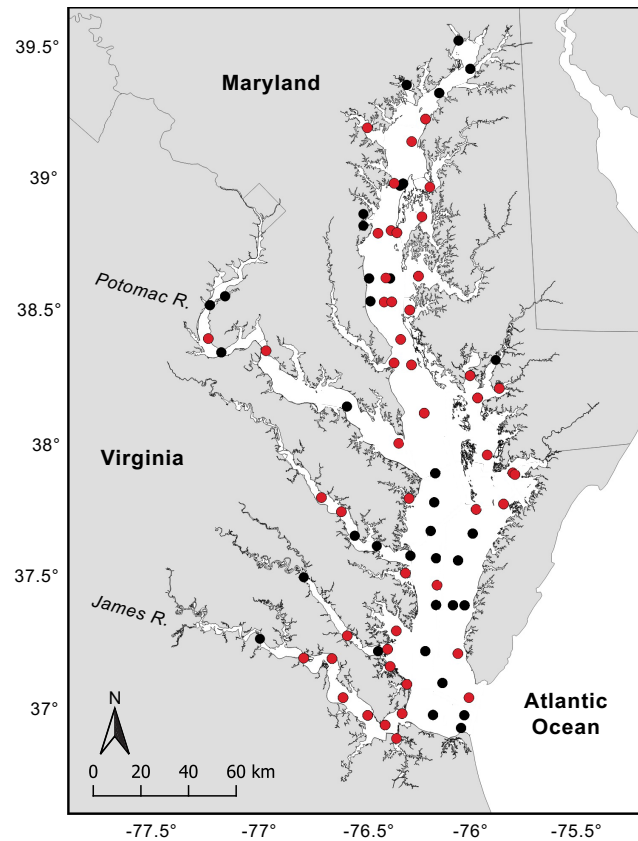
#### 3.3.1 Statistical methodology

In order to investigate our four objectives, seven input variable structures with various combinations of environmental determinants (denoted as “EC” model names) and remote sensing reflectances (denoted as “Rrs” model names) as predictor variables were used in the following analyses (Table 3.1). Several modeling approaches were applied to each of the input variable structures – two statistical models, a Generalized Linear Model (GLM) and a Generalized Additive Model (GAM), and two more complex machine learning models, a Random Forest (RF) and a Support Vector Machine (SVM) (Nelder & Wedderburn, 1972; Hastie & Tibshirani, 1986; Breiman 2001; James et al. 2013). All were chosen for their ability to exploit multi-dimensional explanatory variables and perform regression on both binary and continuous response variables. We chose to test our objectives using four different models in order to avoid potential bias in the results introduced by one specific model type.

A GLM is a modification of an ordinary least squares (OLS) linear regression model in which the sum of squared residuals is minimized (Hastie & Tibshirani, 1986). The GLM includes a link function that describes how the expected response relates to the



linear predictors, thereby accounting for non-normal distributions of response data (Nelder & Wedderburn, 1972). However, GLMs are only able to incorporate a linear



**Figure 3.1.** Map of Chesapeake Bay showing locations of satellite-*in situ* match-ups used in this study. Red circles indicate match-ups where *Vibrio parahaemolyticus* was detected, while black circles indicate match-ups where the bacterium was not detected.

combination of predictor variables. A GAM is then an extension of the GLM, but is able to account for potential non-linear effects of predictor variables on the response variable through use of a non-parametric smoothing function (Hastie & Tibshirani, 1986).

A Random Forest is tree-based machine learning model where decision trees are constructed to follow rules at data-defined nodes in the predictor variables in order to predict the response variable. The RF is an extension of the classic classification and regression tree (CART), where multiple trees are constructed using bagging for the training data and predictor variables used in each tree (Breiman, 2001). This ensemble approach capitalizes on the majority or average output from the “forest” of decision trees in order to increase the accuracy of predictions and avoid overfitting. However, the relationships between predictors and the response in a RF model are not easily interpreted like a GLM or GAM, and the model is not able to predict a response outside of the range of the training dataset.

A Support Vector Machine (SVM) is a machine learning model that performs classification by optimizing a non-linear hyperplane and margin between two classes of training data in multi-dimensional space (Smola & Schölkopf, 2004). Test data is then classified based on which side of the hyperplane it falls. For regression, the SVM minimizes a loss function using only data whose residuals fall outside of a user-specified threshold similar to the margin used in classification. Data whose residuals fall within the threshold do not contribute to the loss function fit (James et al., 2013). The response from a SVM is less susceptible to the influence of outliers, but finding appropriate tuning parameters is not trivial and large datasets can be computationally demanding.

Due to the strong zero-inflation of our datasets (with non-detects), we predicted *V. parahaemolyticus* using both classification (detected presence versus absence) and regression (bacterial abundance) methods. To avoid bias toward the majority class during classification predictions, the training dataset was rebalanced so that the minority class

(presence) was oversampled and majority class (absence) was undersampled. The performance of each variation of model was evaluated over 100 iterations of a Monte Carlo holdout scheme, where the dataset was randomly split into variations of 80% training data and 20% holdout validation data. The evaluation metrics presented in this study are the mean value of that metric over the 100 iterations of holdout validation data. The same methodology was used for both the Same Week prediction models and the 1-Week Forecast models. All modeling and statistical computations were performed in the R Statistical Computing Environment (version 3.3.2) (R Core Team, 2013).

**Table 3.1.** Description of environmental and remote sensing reflectance input structures used in this study.<sup>a</sup>

Structure Name	Description
EC 1	SST
EC 2	SST + SSS
EC 3	SST + SSS + CHLA
EC 4	SST + SSS + TSS
EC 5	SST + SSS + TSS + CHLA
Rrs 1	$R_{rs}(412) + R_{rs}(555) + R_{rs}(678) + R_{rs}(859)$
Rrs 2	$R_{rs}(412) + R_{rs}(555) + R_{rs}(678) + R_{rs}(859) + SST$

<sup>a</sup>SST, sea surface temperature; SSS, sea surface salinity; TSS, total suspended solids; CHLA, chlorophyll-a;  $R_{rs}$ , remote sensing reflectance.

### 3.3.2 Classification methods for bacterial presence and absence prediction

The first phase in the modeling effort of the zero-inflated Same Week and 1-Week Forecast datasets was to use binary classification models to predict detected presence or absence of *Vibrio parahaemolyticus* in Chesapeake Bay. To avoid bias toward the majority class in our predictions, in this study the absence of *V. parahaemolyticus*, the training data in each of the Monte Carlo iterations was rebalanced using the “ROSE” package in the R Statistical Computing Environment (Lunardon et al., 2014; R Core team, 2013). The chosen rebalancing technique both oversampled the minority class to 40% and undersampled the majority class to 60% of the training dataset.

Each of the four statistical and machine learning models were run using seven variations of remote sensing-based EC and Rrs models (Table 3.1). The best threshold for classification into present or absent from predicted probability was determined iteratively by each model run, since there is no ecological basis to the default threshold of 0.5. Using the Same Week dataset, *in situ* measurements were input into the EC models in place of remote sensing data in order to compare the utility of remote sensing-based to *in situ*-based models.

Binary classification predictions from these models were evaluated using three metrics, which are calculated from the features of a confusion matrix – accuracy (ACC), sensitivity (SE), and specificity (SP). Accuracy is the measure of a model’s ability to correctly distinguish between positive and negative, or presence and absence, *V. parahaemolyticus* samples. While accuracy is a widely used metric for classification models, we also consider the two additional metrics due to the potential for accuracy to misrepresent performance when a strong bias toward the majority class is predicted.

Sensitivity, also referred to as the true positive rate, is a measure of a model's ability to correctly predict *V. parahaemolyticus* presence (detects). Specificity, or true negative rate, is a measure of its ability to correctly predict absence of *V. parahaemolyticus* (non-detects). While we consider all three metrics equally within this study for the application to *V. parahaemolyticus* prediction, previous work has suggested other evaluation criteria (MacKenzie et al., 2002; Lobo et al., 2007).

The calculations of classification metrics used in this study to describe the holdout validation results are shown below:

$$ACC = (TP + TN) / (P + N) \quad (1)$$

where TP is the number of true positives, or correctly predicted presence samples, TN is the number of correctly predicted absence samples, and the sum of both is divided by the total number of presence samples (P) and absence samples (N).

$$SE = TP / (TP + FN) \quad (2)$$

where FN is the number of false negatives, or presence samples that are misclassified as absence.

$$SP = TN / (FP + TN) \quad (3)$$

where FP is the number of false positives, or absence samples that are misclassified as presence.

We visualize the distribution of these classification metrics over the 100 iterations of randomly sampled training and holdout data using boxplots. A boxplot is constructed using quartiles, where the length of the box is bounded by the first and third quartiles (25<sup>th</sup> and 75<sup>th</sup> percentiles) and the median value is indicated by a horizontal line within the box. The “whiskers” extending from the box show the variability of data outside of

the first and third quartile range, while outliers are represented by open circles above or below these whiskers.

### 3.3.3 Regression methods for bacterial abundance prediction

To predict and forecast the abundance of *V. parahaemolyticus*, we employed the same statistical and machine learning models described above for continuous regression analyses. Abundance data was transformed using a  $\log(x + 1)$  transformation in order to avoid negative values and preserve the natural limit of zero in the analyses.

Following the precedent of Urquhart et al. (2015) and offering realistic operational potential, the abundance predictions used a combination of the above classification procedure and continuous regression. The classification model first separated the holdout validation data into present or absent. Four modeling approaches (GLM, GAM, RF, SVM) capable of regression were then trained on the presence-only data within the same training dataset as was used for the classification model. Finally, the models predicted abundance for all holdout data that was classified as present for *V. parahaemolyticus*. This combination of procedures was done for each of the seven EC and Rrs models, using same EC or Rrs input variable structures for both the classification and regression steps.

The regression models' performance on the holdout validation data was evaluated using a weighted mean absolute error (WMAE), introduced for zero-inflated datasets in Shashanni et al. (2016). The WMAE takes into consideration true positives and false positives from the prior classification stage when evaluating the success of the regression holdout predictions. In this study, we weighted false positives in the mean absolute error

calculations more heavily (by a factor of 10) than true positives in the classified presence data. Various weights were evaluated in the previous study, but ultimately it was found to be an arbitrary choice based on the application (Shashanni et al., 2016). The WMAE metric produces large values compared to the range of the dataset and are not presented in the original units of the data. It is used in this study only as a tool to compare performance between models.

In this work, we chose to weight by a factor of 10, versus 1 or 100, in order to represent an intermediate effect on the regression model results. This weighting could represent a scenario in which there is an economic cost to incorrectly predicting harmful concentrations of *V. parahaemolyticus* in a shellfish harvesting area due to an unnecessary closure. Currently there are no *V. parahaemolyticus* prediction or forecast-based closures of shellfish harvesting areas, so the weighting method used to evaluate abundance predictions here is conservative.

#### 3.3.4 Statistical null models

Statistical null models provide a baseline to which the classification and regression models in this study can be compared. The null classification models used in holdout validation were structured to imitate the ratio of presence to absence in the Same Week and 1-Week Forecast datasets, composed of randomly distributed 13% presence and 87% absence and randomly distributed 16% presence and 84% absence, respectively. For regression, the statistical null model used the above classification distribution of randomly assigned presence and absence and then assumed the mean abundance of the

datasets for presence predictions ( $1.13 \log(\text{GE/ml} + 1)$  for Same Week,  $1.12 \log(\text{GE/ml} + 1)$  for 1-Week Forecast).

### 3.3.5 Spatial comparison

To provide an example of the spatial variation in predictions between models using only SST versus those using additional environmental predictors, we compared maps of predicted abundances from two environmental input structures. We applied these models to a MODIS 8-day composite from dates 20-27 July 2013 that was not included in the previous model comparisons. Abundance predictions were generated through the classification-regression combination method using RF-EC1, with only SST as its input, and RF-EC5, which uses all of the remotely sensed environmental information available.

Our *in situ* dataset does not extend into this time period for a quantitative “ground truthing” comparison between models. However, we analyzed the similarities and differences between the models’ output qualitatively and in reference to a reported year-long shellfish harvesting area closure by the Virginia Department of Health on 12 July 2013 that was enforced after three confirmed *V. parahaemolyticus* illness cases (Virginia Department of Health, 2013). The MODIS 8-day composite used here occurred after the infected shellfish were harvested, but it provides nearly full coverage of the Bay that is less limited by cloud cover than preceding 8-day composites from that summer.

## 3.4 Results and Discussion

The relationships between weekly MODIS-derived water quality variables and corresponding *in situ* measurements in the Same Week dataset are found in Figure 3.2.



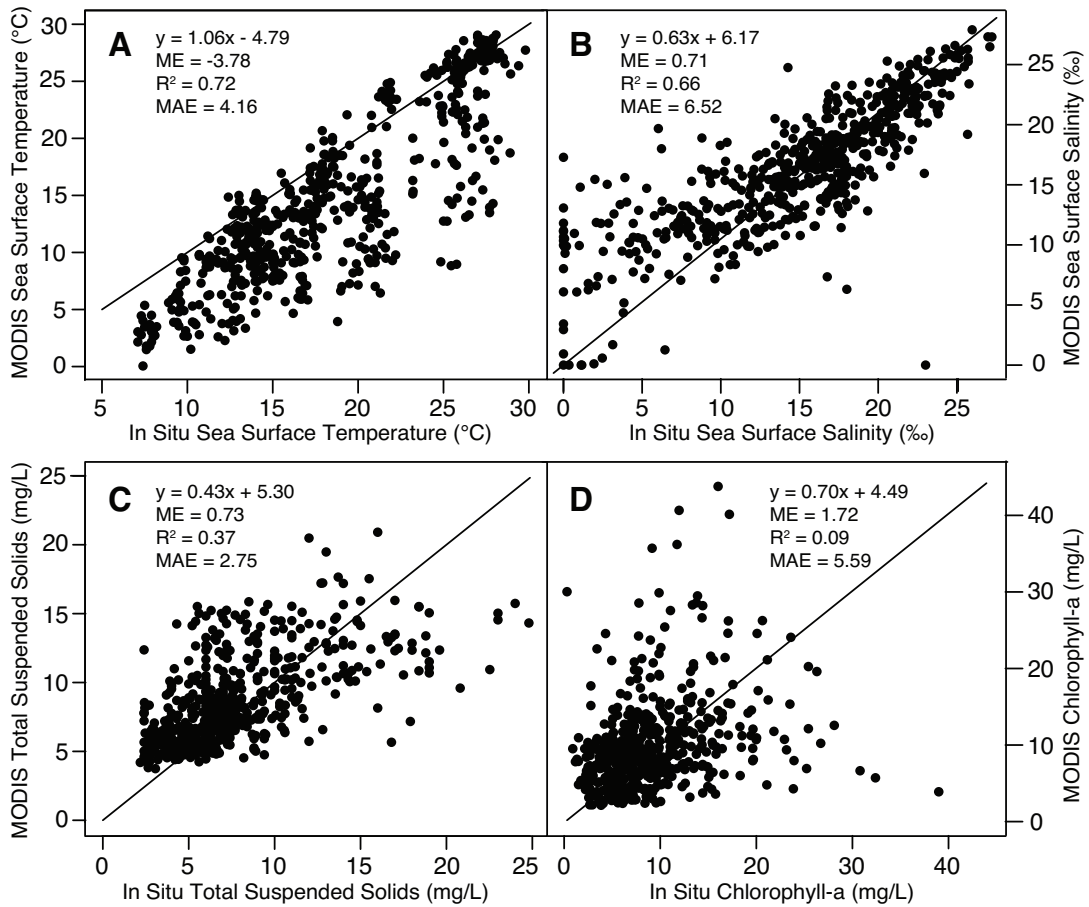
Linear regression equations, explained variance ( $R^2$ ), bias (mean error - ME), and mean absolute error (MAE) are reported. While SST and SSS showed strong correlation ( $R^2 > 0.6$ ), MODIS-derived TSS and chlorophyll-a showed weak correlation between the satellite estimates and *in situ* measured values (Figure 3.2). The limited ability to distinguish between suspended solids, dissolved organic material, and chlorophyll-a using ocean color remote sensing in optically complex Case II waters like Chesapeake Bay has been well documented (Gons, 1999; Dall’Olmo et al., 2005; Schalles, 2006; Gitelson et al., 2007; Werdell et al., 2009). Because MODIS 8-day composites are used here to increase satellite coverage in cloudy summer months, the satellite-*in situ* relationships are likely weakened further due to the dynamic nature of estuarine systems on sub-week timescales. However, satellite TSS and chlorophyll-a retrievals showed a low bias (0.73 and 1.72, respectively) while SST retrievals showed a larger bias towards lower values than *in situ* measurements (-3.78). Summary statistics for MODIS and *V. parahaemolyticus* data in the satellite-*in situ* matchup datasets are found in Table 3.2.

#### 3.4.1 Presence, absence predictions for Same Week

All variations of the classification models used here performed better than the statistical null model (ACC = 76.73%, SE = 13.48%, SP = 86.88%) when taking all three metrics into consideration (Figure 3.3). While the accuracy of the null model appears to perform well, the true classification performance on the zero-inflated holdout data is revealed by the low sensitivity metric. Low sensitivity and high specificity indicates that the null model predictions rarely capture true positive samples. This result suggests that the

models in this study were better able to classify true presence and absence than a model that randomly guesses.

For all four modeling approaches, the accuracy of the holdout predictions increased with the addition of remotely sensed environmental predictor variables (EC1



**Figure 3.2.** One-to-one comparison between Same Week satellite-*in situ* matchups for (A) sea surface temperature, (B) sea surface salinity, (C) total suspended solids, (D) chlorophyll-a and corresponding *in situ* measurements. Bias (ME) and mean absolute error (MAE) are reported along with linear regression equations.

versus EC4, EC5) (Figure 3.3). The accuracy of all model types' EC4 and EC5 variations were statistically different ( $p < 0.05$ ) than the respective EC1 variation that uses only SST (Table S3.1). The SVM-Rrs and RF-Rrs2 models also produced accuracies that were higher and statistically significant than their SST-only models (Table S3.1).

Models with the highest accuracies were SVM-EC4 (70.74%) and GLM-EC4 (70.63%) and SVM-Rrs2 (69.03%) (Table S3.1). The addition of chlorophyll-a in EC5 did not improve the accuracy over EC4 in any of the models (Figure 3.3, Table S3.1). The Rrs models generally underperformed compared to the best EC models except for SVM-Rrs2, but they provided higher accuracies than EC1 and EC2 in all cases except RF-Rrs1 (Figure 3.3, Table S3.1). *In situ* models showed similar performance to their corresponding remote sensing models in most cases, with GLM-EC4 being a notable exception where the remote sensing model outperformed the *in situ* model by a wider margin (Figure 3.3, Table S3.2).

The sensitivity metric helps to further differentiate the performance between the classification models by indicating the ability of the models to correctly predict presence. The GLM-EC and SVM-EC models showed decreased sensitivity with the addition of remotely sensed variables (EC1 versus EC4, EC5), while the GAM-EC and RF-EC models showed higher sensitivities than their SST-only models (Figure 3.3).

The highest sensitivities are achieved by GLM-EC1 (82.13%) and RF-Rrs1 (79.02%) and the lowest sensitivities were produced by SVM-EC4 (57.97%) and RF-EC1 (61.22%) (Table S3.1). The GAM-Rrs and RF-Rrs models perform better than their best respective EC models, while the GLM-Rrs and SVM-Rrs models did not perform as well as their EC models (Figure 3.3). The addition of SST in Rrs2 improved the sensitivity

over Rrs1 in all models except RF-Rrs (Table S3.1). The *in situ* models showed high sensitivities, outperforming their remote sensing counterparts in all cases (Figure 3.3, Table S3.2).

**Table 3.2.** Summary of remote sensing products and *in situ* *Vibrio parahaemolyticus* measurements used in this study.

Variable	<u>Same Week</u> <sup>a</sup>				<u>1-Week Forecast</u> <sup>b</sup>			
	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum
<i>Vibrio parahaemolyticus</i> abundance (GE/mL) <sup>c</sup>	2.80	3.06	0.40	17.71	2.64	2.75	0.40	17.71
Sea Surface Temperature (°C)	14.49	7.36	0.05	29.06	16.99	7.92	1.35	29.95
Sea Surface Salinity (‰)	15.51	4.94	0.00	26.49	15.19	4.94	0.00	28.70
Total Suspended Solids (mg/L)	8.71	3.59	3.80	31.97	8.97	3.71	4.15	26.67
Chlorophyll-a (mg/L)	10.67	12.29	2.16	180.33	12.56	24.76	0.10	472.19
Rrs_412-nm (sr <sup>-1</sup> ) <sup>d</sup>	1.71E-03	1.19E-03	0.00E+00	7.17E-03	1.78E-03	1.34E-03	1.00E-05	9.96E-03
Rrs_555-nm (sr <sup>-1</sup> )	7.50E-03	3.25E-03	2.02E-03	2.23E-02	7.50E-03	3.30E-03	1.65E-03	2.30E-02
Rrs_678-nm (sr <sup>-1</sup> )	3.83E-03	3.59E-03	8.00E-05	3.08E-02	3.77E-03	3.04E-03	5.00E-05	1.96E-02
Rrs_859-nm (sr <sup>-1</sup> )	7.83E-04	7.97E-04	0.00E+00	5.90E-03	7.74E-04	7.84E-04	1.50E-05	4.90E-03

<sup>a</sup>n = 572

<sup>b</sup>n = 605

<sup>c</sup>Calculated using *in situ* samples where Vp was detected, Daily (n =77 ) 1 Week Forecast (n = 98)

<sup>d</sup>Remote sensing reflectance

The specificity metric also helps to further diagnose the performance of the classification models by indicating the ability of the models to correctly predict absence. The performance in specificity generally increased with the addition of remotely sensed environmental variables (EC1 versus EC4, EC5) for each model type (Figure 3.3). Models with the highest specificities were SVM-EC4 (72.64%) and GLM-EC4 (71.14%)

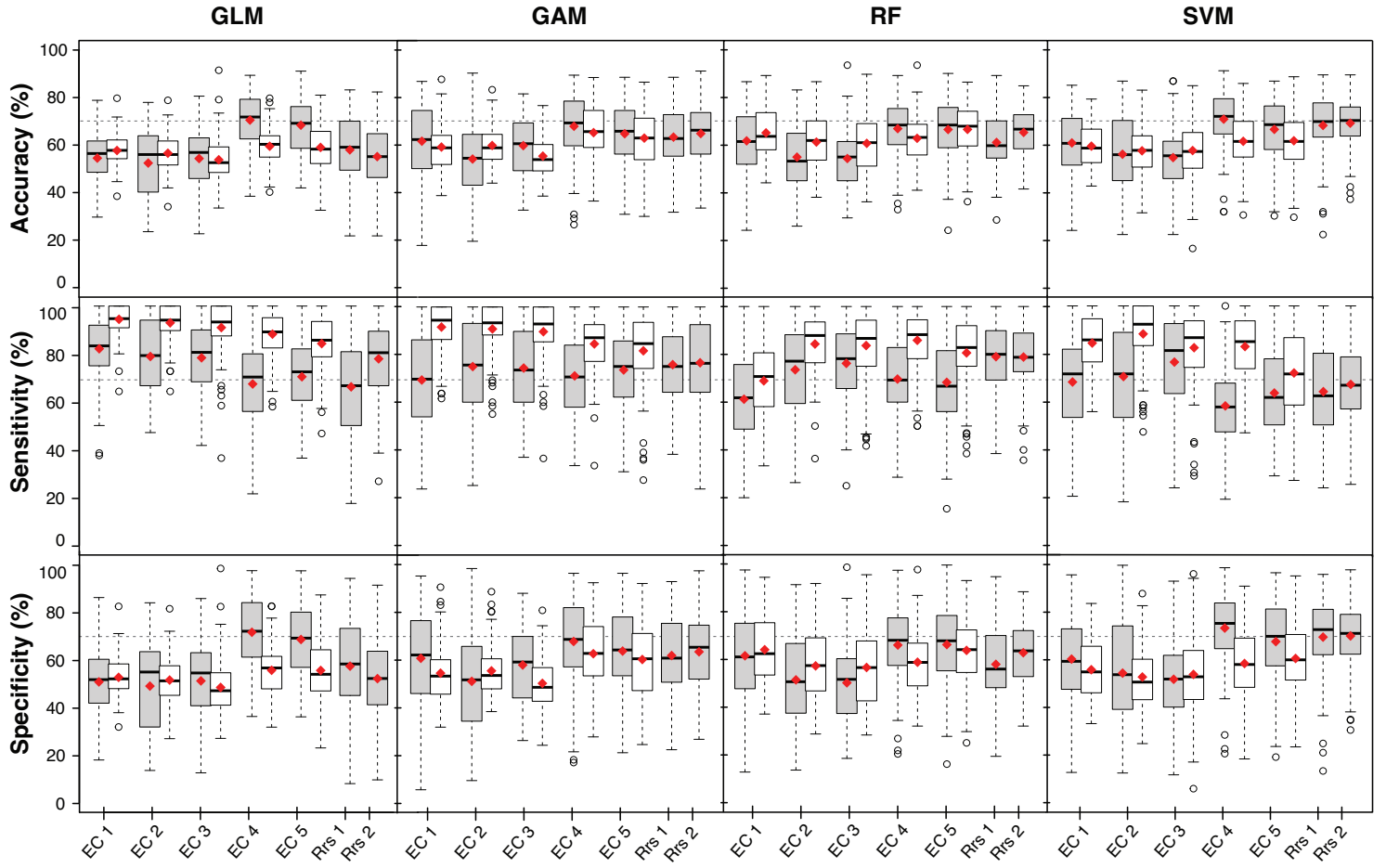
(Table S3.1). The Rrs models showed lower specificities compared to the best EC models, but were higher than EC1 and EC2 in all cases except RF-Rrs1 (Figure 3.3, Table S3.1). *In situ* models had similar or lower specificities than their corresponding remote sensing models for all model types (Figure 3.3, Table S3.2).

Based on all three classification metrics reported, we determined that the best performing Same Week classification models to predict *V. parahaemolyticus* presence and absence were GLM-EC4, GLM-EC5, and SVM-Rrs2 (Table S3.1). However, no model type performed better than the others when considering performance over all remotely sensed input variations.

#### 3.4.2 Abundance predictions for Same Week

Abundance models in this study used a combination of classification and continuous regression due to the strong zero-inflation in the datasets. Here we used RF classification for all models, while varying only the model type for abundance predictions. The same input structures (EC1, EC2, Rrs1, etc.) were used for both the RF classification and regression steps.

All variations of the classification-regression combination models were on average able to predict *V. parahaemolyticus* abundance better than the mean statistical null model (WMAE = 10.02) (Figure 3.4). This indicates that the models can better predict abundance than a model that assumes the mean detected abundance at sites with randomly predicted presence of *V. parahaemolyticus*.



**Figure 3.3.** Boxplots showing accuracy, sensitivity, and specificity for Same Week holdout classification predictions. Remote sensing-based models shown in gray and corresponding *in situ* models shown in white. Red diamonds indicate mean value. Dotted line shown only for ease of visual comparisons between models.

Models with more remotely sensed environmental information (EC4 and EC5) produced lower WMAEs than models with less environmental information (EC1 and EC2) for all model types except SVM, where SVM-EC4 produced lower WMAE than SVM-EC1 (Figure 3.4). Models with the lowest WMAE were SVM-EC4 (8.36) and

SVM-Rrs1 (8.43) (Table S3.3). GLM, GAM, and RF reflectance (Rrs) models exhibited higher WMAEs than the best EC models for each model type, but performed better than EC1 and EC2 in all cases (Table S3.3). Only SVM-Rrs1 produced lower WMAE than the SST-only SVM-EC1 model (Table S3.1). Many of the SVM models outperformed the other model types when considering all remotely sensed input structures. Unlike presence and absence prediction accuracy, some *in situ* regression models performed considerably better than their equivalent remote sensing-based models (Figure 3.4).

Improvements that remote sensing offers over *in situ*-based models could be due to the nature of integrating information over a larger area. However, models using *in situ* data are able to incorporate environmental parameters that are not currently available from remote sensing (e.g. dissolved organic nitrogen). Compared to previous *in situ* modeling studies for *V. parahaemolyticus* in Chesapeake Bay, our remote sensing models performed similar in accuracy and sensitivity on holdout validation data (Urquhart et al., 2015; Davis et al., 2019). However, the differences in size and distribution between data used in previous work and the dataset used here may not offer a fair comparison between studies.

### 3.4.3 Spatial comparison

To provide an example of the spatial variation in predictions, we compared maps of predicted abundances from two different environmental input models (Figure 3.5). Abundance predictions were generated through the classification-regression combination method using RF-EC1, with only SST as its input, and RF-EC5, which uses all four remotely sensed environmental input variables. The weekly composite used corresponds

to a reported year-long shellfish harvesting area closure by the Virginia Department of Health on 12 July 2013 due to confirmed *V. parahaemolyticus* illness cases (Virginia Department of Health, 2013).

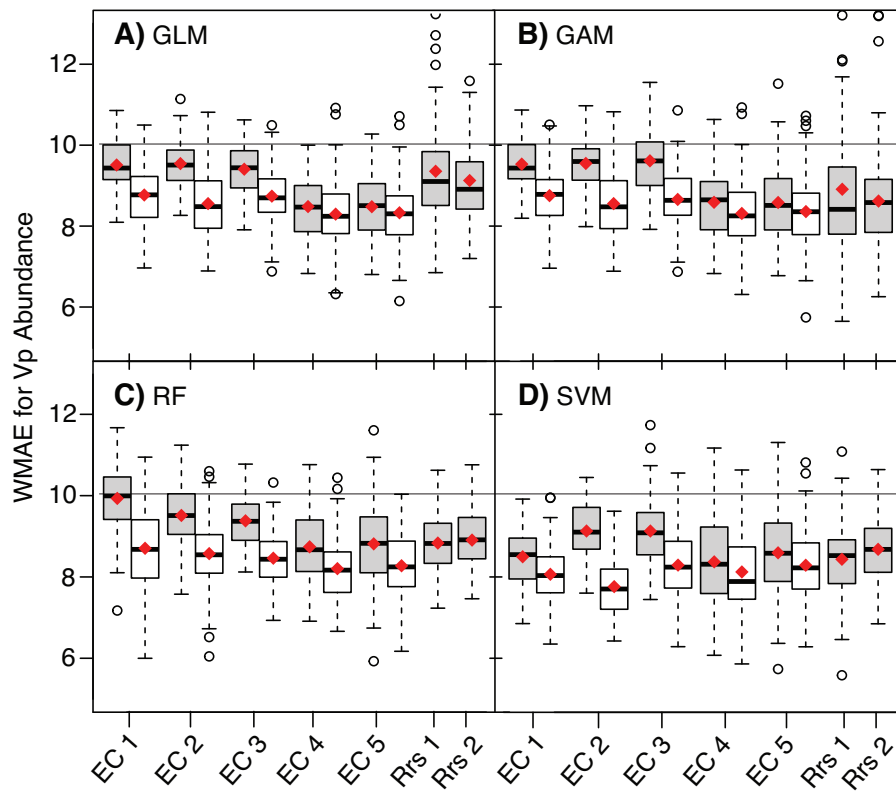
Both models predicted absence or low abundance of *V. parahaemolyticus* in the Upper Chesapeake Bay region and on the eastern coast of the mid-Bay region. Spatial differences in predictions between the models are apparent in many areas. In the James River and Potomac River, the SST-only model predicted absence or low abundance while the EC5 model predicted much higher abundances of the bacterium (Figure 3.5). The SST-only model predicted considerable *V. parahaemolyticus* abundance in the Mainstem of the Lower Bay and north of the Potomac River, but the EC5 model predicted that these regions were predominantly clear of detectable *V. parahaemolyticus*.

When considering the area around Fisherman's Island that would have been closed to shellfish harvesting during this time, both models predicted there to be small areas of higher bacterium concentrations. However, the SST-only model predicted this higher region to occur to the east of the Island while the EC5 model predicted it to be more westward (Figures 3.5C, 3.5D). It should be noted that bottom reflectance from shallow waters like those found in this area have been shown to affect ocean color retrievals and may contribute to particularly inaccurate TSS and chlorophyll-a retrievals (Carder et al., 2005).

A key difference between the models shown in the mapped predictions is the ability of each to distinguish between *V. parahaemolyticus* presence and absence. Presence predictions, where the abundance shown is greater than zero, are widespread in the Mid and Lower Bay for the SST-only model (Figure 3.5A). However, the EC5 model



incorporating other water quality predictors distinguishes areas of presence in the Mid Bay and near shore Lower Bay regions from areas of absence in the Mainstem of the Bay (Figure 3.5B). This is consistent with our results in Section 4.1 where specificity, or the ability to correctly predict absence, is increased with the addition of SSS, TSS, and chlorophyll-a. While we do not have *in situ* data from this 8-day period to validate the model results, they agree with the smaller percentage of detects found in the Lower Mainstem in a previous study (Davis et al., 2017).



**Figure 3.4.** Boxplots showing weighted mean absolute error (WMAE) for Same Week holdout presence-abundance predictions. Remote sensing-based models shown in gray and corresponding *in situ* models shown in white. Red diamonds indicate mean value. Mean statistical null model ( $WMAE = 10.02$ ) shown as solid black line.

#### 3.4.4 Presence, absence for 1-Week Forecast

Results shown in the previous sections represented zero-lag predictions. Now we consider the ability of the remote sensing-based models to forecast *V. parahaemolyticus* one week in advance, which could provide shellfish managers and harvesters additional time to prepare preventative and control measures (Konrad et al., 2017). Here we assess a temporal lag of one week, but other sub-monthly lags could be investigated for future operational forecasts (Davis et al., 2019). Accurate forecasting is particularly important given the time that it takes to generate and disseminate predictions, which can be up to several days on account of satellite data latency and communication requirements.

Classification models for forecasting *V. parahaemolyticus* in Chesapeake Bay one week in advance performed better than the statistical null model (ACC = 73.84%, SE = 17.01%, SP = 84.56%) when taking all three evaluation metrics into consideration. Similar to the Same Week case, the low sensitivity and high specificity indicates that the null model's predictions capture very few true positive samples.

All four modeling approaches displayed increased accuracies on predicted presence and absence with an increase of remotely sensed environmental information (EC1 to EC5) (Figure 3.6). The accuracy of all model types' EC4 and EC5 variations were statistically different ( $p < 0.05$ ) than the respective EC1 variation that uses only SST (Table S3.4). Forecast models producing the highest accuracies were SVM-EC5 (67.80%), SVM-EC4 (67.57%), and GAM-EC4 (65.94%) (Table S3.4).

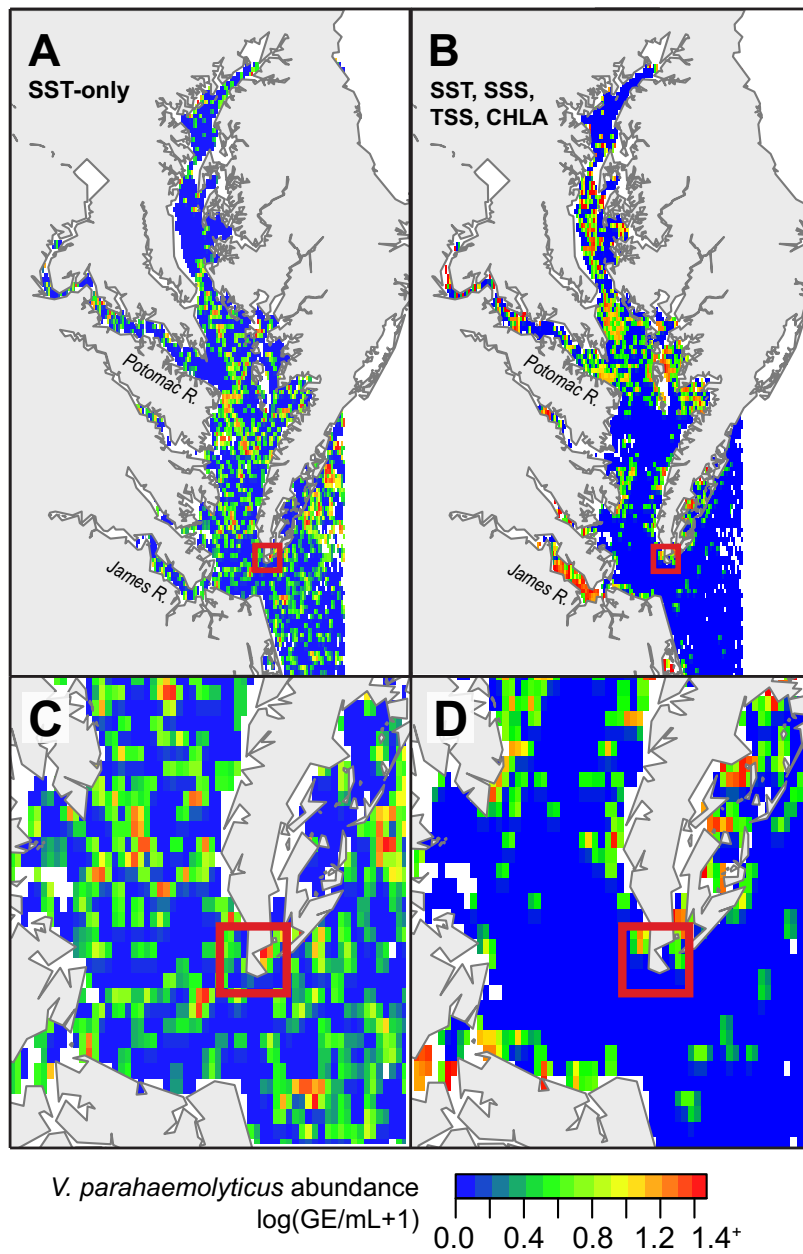
While reflectance (Rrs) forecast models did not perform as well the best environmental models in terms of accuracy, all but RF-Rrs1 outperformed the SST-only model for each model type. Both Rrs models for GAM and SVM, as well as GLM-Rrs2,

produced accuracies that were also statistically significant from the SST-only models (Table S3.4).

The GLM-EC, GAM-EC, and SVM-EC models showed decreased sensitivity with the addition of remotely sensed variables (EC1 versus EC5), while the RF-EC models showed increasing sensitivity with the addition of inputs (Figure 3.6). The GLM-Rrs, GAM-Rrs, and SVM-Rrs models produced lower sensitivities than their respective EC models, while RF-Rrs2 exhibited higher sensitivity than all of its EC models (Figure 3.6). The addition of SST in Rrs2 improves the sensitivity over Rrs1 in all models except SVM-Rrs (Table S3.4).

The specificity increased with the addition of remotely sensed environmental variables (EC1 versus EC4, EC5) for each model type (Figure 3.6). The Rrs models generally showed lower specificities compared to the best EC models, but were higher than EC1 and EC2 in all cases except RF-Rrs1 and RF-Rrs2 (Figure 3.6, Table S3.4). The addition of SST in Rrs2 increased the specificity over Rrs1 for all model types (Table S3.4).

Based on all three classification metrics reported, the best performing Same Week classification model to forecast *V. parahaemolyticus* presence and absence was determined to be SVM-EC4 (Table S3.4). No model type outperformed the others when considering performance over all remotely sensed input variations. In contrast to the Same Week model results, the addition of remotely sensed chlorophyll-a in EC5 improved the performance over EC4 for GLM, RF, and SVM models (Figure 3.6).



**Figure 3.5.** Spatial comparison of Random Forest abundance predictions on MODIS 8-day composite from 20 July 2013 to 27 July 2013 using (A) sea surface temperature and (B) sea surface temperature, sea surface salinity, total suspended solids, and chlorophyll-a as remotely sensed inputs. Red boxes show shellfish harvesting area near Fisherman's Island, with (C) and (D) showing zoomed views of (A) and (B), respectively.

### 3.4.5 Abundance for 1-Week Forecast

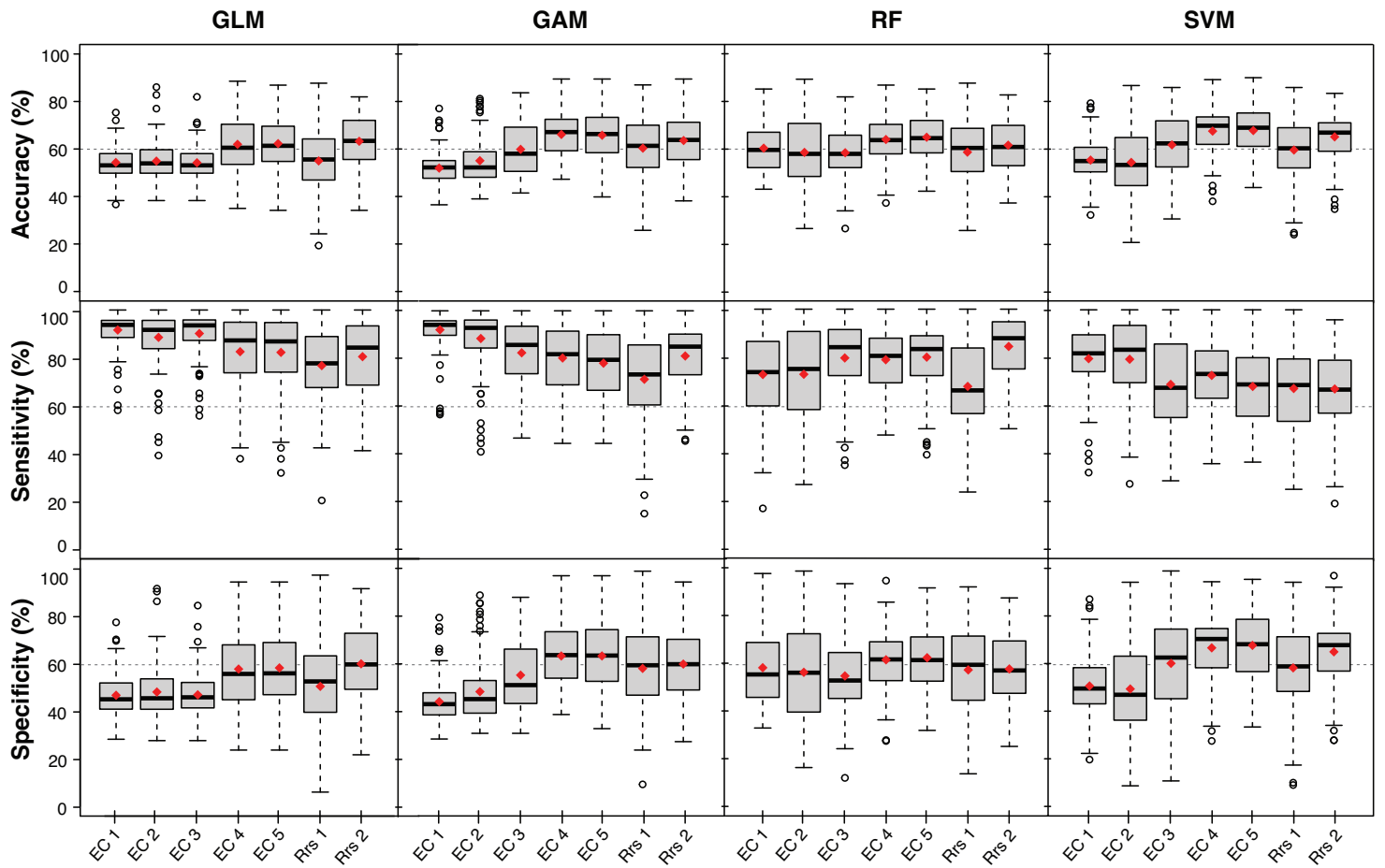
We again used a combination of classification and continuous regression due to zero-inflation in the matchup dataset. Random forest classification was again used for all models while the model type was varied only for regression. The same input structures (EC1, EC2, Rrs1, etc.) were used for both the classification and regression models.

All classification-regression combination models were on average able to forecast *V. parahaemolyticus* abundance better than the mean statistical null model (WMAE = 9.04) (Figure 3.7, Table S3.5). More remotely sensed environmental information (EC3 to EC5) improved the abundance forecasts over forecast from less environmental information (EC1, EC2) for GLM, GAM, and RF (Figure 3.7, Table S3.5). For SVM, only the EC4 and EC5 forecasts improved over EC1 and EC2. The lowest WMAEs were achieved by RF-EC5 (7.56) and SVM-EC4 (7.69) (Table S3.5).

Forecasts using the remote sensing reflectance input structure Rrs2 performed similar to or better than EC models for GLM and GAM (Figure 3.7). This performance compared to EC models was not shown for RF and SVM. However, SVM-Rrs1 and SVM-Rrs2 produced lower WMAE than SVM-EC1 and RF-Rrs2 achieved lower WMAEs than RF-EC1 and RF-EC2 (Figure 3.7).

## 3.5 Conclusion

This study evaluates prediction and forecast models for *V. parahaemolyticus* in Chesapeake Bay using satellite-derived products. It investigates whether additional remotely sensed environmental information improves predictions and forecasts, and if untransformed remote sensing reflectances could be used as predictors in place of environmental products.

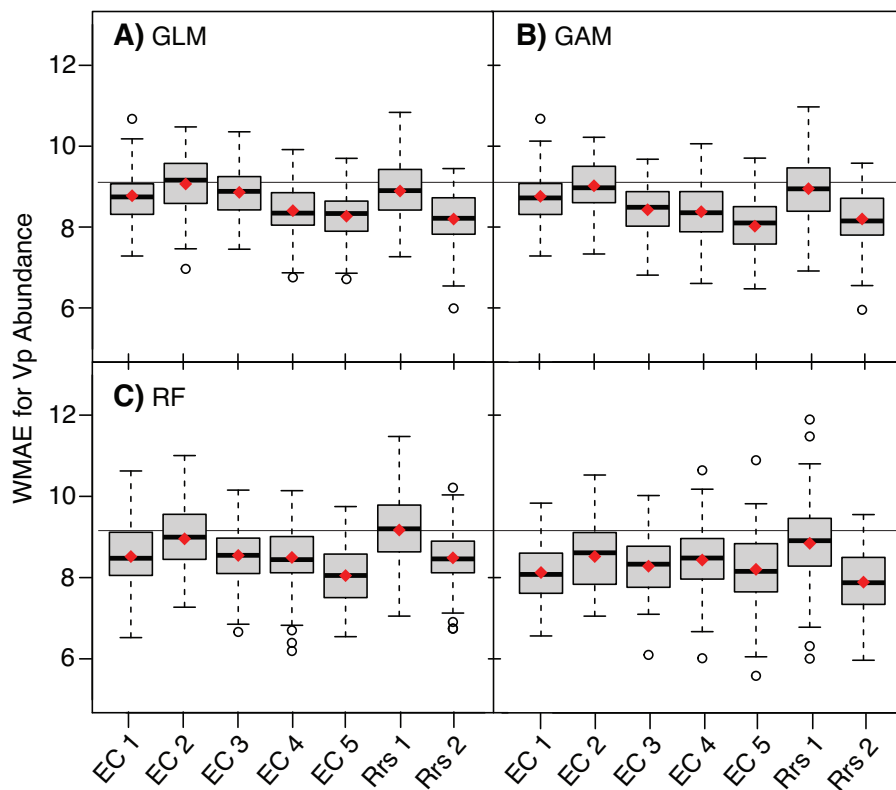


**Figure 3.6.** Boxplots showing accuracy, sensitivity, and specificity for 1-Week Forecast holdout classification predictions. Red diamonds indicate mean value. Dotted line shown only for ease of visual comparisons between models.

Finally, the study considers whether remotely sensed predictions are comparable to predictions from *in situ* measurements.

Ocean color products, including environmental variables and remote sensing reflectances, and SST derived from MODIS were used as inputs to four statistical and machine learning modeling approaches to predict presence and abundance of the bacterium

over a 4-year period. MODIS was chosen for this work because of its overlap with the *in situ* dataset and its vast literature of established ocean color algorithms. However, we acknowledge that these results serve as a proof of concept and additional testing is required when applying these methods to newer sensors with ocean color capacity like NOAA Visible Infrared Imaging Radiometer Suite (VIIRS) sensor aboard the Suomi National Polar-Orbiting Partnership spacecraft (2011 –) or the EUMETSAT Ocean and Land Colour Instrument (OLCI) aboard the Sentinel-3 spacecraft.



**Figure 3.7.** Boxplots showing weighted mean absolute error (WMAE) for 1-Week Forecast holdout abundance predictions. Red diamonds indicate mean value. Mean statistical null model (WMAE = 9.29) shown as solid black line.

More environmental information generally improved the accuracy of predicting presence and improved abundance predictions in the same week. The same was found while forecasting *V. parahaemolyticus* presence and abundance, which was also shown in a recent study examining monthly temporal lags using *in situ* data in Chesapeake Bay (Davis et al., 2019). Although an association with salinity has been commonly observed (DePaola et al., 2003; Zimmerman et al., 2007; Patra et al., 2009; Johnson et al., 2010; Davis et al., 2017), the results of this study showed little improvement with the addition of SSS over models using only SST. However, the salinity estimates from hydrodynamic models have been shown to be more accurate than satellite-derived estimates (Vogel & Brown, 2016). Therefore, a mixed-media approach could improve these models and should be investigated in the future.

The addition of chlorophyll-a to EC models improved the performance of classification and abundance predictions for the 1-Week forecast, while they did not improve predictions for the Same Week models. This could be due to the association of *V. parahaemolyticus* and chitin-rich zooplankton, where chlorophyll-a indicates the presence of the primary food source of phytoplankton.

Additional environmental information improved accuracy over SST models primarily because it reduced false positives (i.e., better prediction of absence), evident where accuracy results mirror the specificity results when comparing from EC1 to EC5. Our results suggest that including additional information could decrease the likelihood of false positive predictions thereby decreasing the likelihood of a shellfish harvesting area being unnecessarily affected by regulatory action. However, false positives may not be detrimental



from the perspective of protecting public health where a cautionary approach could be preferred.

Models using reflectances as inputs did not show accuracies as high as models using environmental inputs for Same Week presence predictions. However, the reflectance models did show potential for forecasting presence one week in advance. The improvement of Rrs models over SST-only models indicates that these products alone have the potential to improve *V. parahaemolyticus* models in Chesapeake Bay. As new ocean color sensors become operational, the ability to use remote sensing reflectances instead of algorithm-derived environmental products could streamline modeling efforts by eliminating the need to first develop algorithms for each environmental parameter needed. Remote sensing reflectance models may outperform models using environmental parameters because they capture ecological processes not represented in the chosen parameters. Their use could bypass the cascading effect on model accuracy first introduced in the derivation of ocean color products from imperfect algorithms.

The remote sensing-based models in this study generally performed similarly to their equivalent *in situ* models for presence and absence prediction accuracy. Some models using more environmental information than only SST showed better accuracy than their *in situ* models by a relatively large margin. However, the remote sensing-based models were not always able to predict *V. parahaemolyticus* abundance as well as the *in situ* models.

Although more complex machine learning models (RF and SVM) at times performed better than simpler models (GLM and GAM), the relationships between *V. parahaemolyticus* and the predictor variables in machine learning-type models are not always obvious. For this reason, some users prefer to implement models like GLMs where the relationships between

variables are apparent and can be verified with previous knowledge of *Vibrio* ecology. However, a larger satellite-*in situ* matchup dataset may contain more complex relationships between variables that cannot be captured by simpler models. Machine learning models may even be able to capture relationships between *V. parahaemolyticus* and environmental variables that are not yet known or understood. An ensemble prediction from both simple and complex models could also improve future predictions.

Larger datasets with additional measurements of detected *V. parahaemolyticus* matched to satellite data are needed to robustly determine the preferred model for operational predictions and early warning systems. While cloud cover and coarse spatial resolution can limit the coverage of remote sensing datasets in coastal waters like the Chesapeake Bay, geostatistical interpolations (i.e. kriging) could be utilized for future models to fill gaps in space and time where satellite data is missing. Previous attempts to kriging with MODIS data in the Chesapeake Bay have been effective (Urquhart et al., 2013) and recent advances in appropriately applying kriging to complex bodies of water could be useful in developing complete forecast maps based solely on remotely sensed data (Davis & Curriero, 2019).

The perspective from both shellfish harvesters and public health professional should be considered in future research leading towards an appropriate operational model for predicting or forecasting *V. parahaemolyticus* in water bodies. For example, there could be an advantage from the public health perspective in choosing a model that produces high sensitivity with low specificity, maximizing correctly predicted positives, if an overestimation of risk is acceptable. In contrast, a more conservative approach where the model maximizes correctly predicted absences could be preferred. In future work leading

towards an operational product, these varying perspectives such as these should be considered.

Overall the results of this study show that satellite remote sensing can be a valuable tool for *V. parahaemolyticus* monitoring and forecasting in Chesapeake Bay surface waters, particularly when products in addition to SST are used. Utilizing higher spatial and spectral resolution sensors could further enhance the value of remote sensing data by providing more coverage of near-shore regions and additional spectral information that could capture additional environmental processes or associations. The results and considerations presented in this study could expand the use and effectiveness of future satellite-informed *V. parahaemolyticus* predictions and forecasts in estuarine and coastal waters.

**Table S3.1** Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for Same Week holdout classification predictions using remotely sensed inputs.

		<b>EC 1</b>	<b>EC 2</b>	<b>EC 3</b>	<b>EC 4</b>	<b>EC 5</b>	<b>Rrs 1</b>	<b>Rrs 2</b>
<b>GLM</b>	ACC	54.63	52.56	54.46	70.63*	68.49*	58.11	55.27
	SE	82.13	78.89	78.36	67.40	70.41	66.19	77.92
	SP	50.31	48.54	50.72	71.14	68.11	56.90	51.68
<b>GAM</b>	ACC	61.58	54.04*	59.71	67.89*	64.75*	63.31	64.82
	SE	69.32	74.97	74.34	71.05	73.53	75.77	76.56
	SP	60.31	50.63	57.50	67.39	63.34	61.47	63.04
<b>RF</b>	ACC	61.70	54.79*	54.18*	66.88*	66.52*	61.03	65.27*
	SE	61.22	73.54	76.21	69.64	68.27	79.02	78.86
	SP	61.87	51.74	50.50	66.37	66.56	58.28	63.15
<b>SVM</b>	ACC	60.69	55.90*	54.48*	70.74*	66.41*	68.09*	69.03*
	SE	68.02	70.31	76.39	57.97	63.40	63.97	67.02
	SP	59.63	53.75	51.17	72.64	66.97	68.86	69.34

\*indicates accuracy of input structure is statistically different from SST-only model

**Table S3.2** Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for Same Week holdout classification predictions using *in situ* measured inputs.

		<b>EC 1</b>	<b>EC 2</b>	<b>EC 3</b>	<b>EC 4</b>	<b>EC 5</b>
<b>GLM</b>	ACC	57.84	56.70	53.94	59.63	59.13
	SE	94.45	92.99	91.00	88.26	84.29
	SP	52.22	51.08	47.98	55.18	55.12
<b>GAM</b>	ACC	59.11	59.77	55.28	65.20	62.89
	SE	91.55	90.77	89.62	84.48	81.58
	SP	54.07	55.04	49.78	62.24	59.80
<b>RF</b>	ACC	65.11	61.18	60.63	62.78	66.55
	SE	68.91	84.31	83.66	85.85	80.60
	SP	64.41	57.61	56.95	59.15	64.13
<b>SVM</b>	ACC	59.33	57.36	57.46	61.37	61.60
	SE	84.39	88.24	82.44	82.93	71.82
	SP	55.19	52.13	53.21	57.79	59.89

**Table S3.3** Weighted mean absolute error (WMAE) for Same Week holdout abundance predictions using remotely sensed (RS) and *in situ* inputs.

		EC 1	EC 2	EC 3	EC 4	EC 5	Rrs 1	Rrs 2
<b>GLM</b>	RS	9.51	9.55	9.41	8.48*	8.48*	9.35	9.13*
	<i>In situ</i>	8.77	8.56	8.74	8.30	8.33	-	-
<b>GAM</b>	RS	9.53	9.54	9.61	8.59*	8.58*	8.91*	8.62*
	<i>In situ</i>	8.75	8.55	8.66	8.31	8.36	-	-
<b>RF</b>	RS	9.93	9.51*	9.38*	8.74*	8.81*	8.83*	8.91*
	<i>In situ</i>	8.71	8.57	8.46	8.20	8.28	-	-
<b>SVM</b>	RS	8.49	9.13*	9.13*	8.36	8.59*	8.43	8.67
	<i>In situ</i>	8.06	7.75	8.28	8.11	8.28	-	-

\*indicates accuracy of input structure is statistically different from SST-only model

**Table S3.4** Percent accuracy (ACC), sensitivity (SE), and specificity (SP) for 1-Week Forecast holdout classification predictions using remotely sensed inputs.

		EC 1	EC 2	EC 3	EC 4	EC 5	Rrs 1	Rrs 2
<b>GLM</b>	ACC	54.11	54.69	54.00	61.78*	62.08*	54.76	63.06*
	SE	91.60	88.52	90.14	82.49	82.2	76.65	80.41
	SP	46.76	48.13	46.95	57.76	58.24	50.49	59.95
<b>GAM</b>	ACC	51.79	54.90*	59.68*	65.94*	65.61*	60.18*	63.39*
	SE	92.05	88.40	82.41	80.23	78.06	71.31	81.07
	SP	44.03	48.24	55.21	63.29	63.30	58.04	59.88
<b>RF</b>	ACC	60.25	58.46	58.32	63.89*	64.84*	58.57	61.51
	SE	72.67	72.78	79.60	78.86	79.94	67.70	84.41
	SP	57.81	55.89	54.38	61.12	61.98	56.90	57.30
<b>SVM</b>	ACC	55.36	54.39	61.78*	67.57*	67.80*	59.60*	65.15*
	SE	79.60	79.35	68.84	72.60	67.99	67.14	66.93
	SP	50.58	49.33	60.10	66.62	67.68	58.21	64.96

\*indicates accuracy of input structure is statistically different from SST-only model

**Table S3.5** Weighted mean absolute error (WMAE) for 1-Week Forecast holdout abundance predictions using remotely sensed inputs.

	<b>EC 1</b>	<b>EC 2</b>	<b>EC 3</b>	<b>EC 4</b>	<b>EC 5</b>	<b>Rrs 1</b>	<b>Rrs 2</b>
<b>GLM</b>	8.76	9.05*	8.85	8.39*	8.26*	8.88	8.19*
<b>GAM</b>	8.75	9.01*	8.42*	8.37*	8.01*	8.93	8.19*
<b>RF</b>	8.53	8.97*	8.56	8.51	8.06*	9.19*	8.50
<b>SVM</b>	8.13	8.52*	8.28	8.44*	8.21	8.85*	7.89*

\*indicates accuracy of input structure is statistically different from SST-only model

#### **4. CHAPTER 4: PROVIDING PREDICTION UNCERTAINTY ON ENVIRONMENTAL MODELS USING QUANTILE REGRESSION FORESTS: A CASE STUDY FOR *VIBRIO PARAHAEMOLYTICUS* IN CHESAPEAKE BAY**

##### **ABSTRACT**

With the improvement in predictions and forecasts from increasingly complex environmental models, Quantile Regression Forests (QRF) have the potential to communicate operationally relevant predictions and associated uncertainty to decision-makers and end-users. In this study we apply the QRF method to a case study in the Chesapeake Bay, where the prediction of *Vibrio parahaemolyticus* bacteria abundance in shellfish harvesting waters aims to decrease the incidence of shellfish-borne illness. Comparison of the mean, median (50<sup>th</sup> quantile), and 95<sup>th</sup> quantile predictions showed that the mean predicted bacterial abundance value was generally higher than the median predicted value, indicating a right-skewed prediction distribution that generally reflects bacterial abundance observations. This could be a particularly important difference for decision-makers during summer months, when the *V. parahaemolyticus* abundance and associated illness cases increase. The application of QRF to a user-defined threshold and acceptable level of risk of the prediction exceeding that threshold showed differences in the spatial extent and temporal frequency of affected harvesting sites when both decisions were varied. While the results of this study are not intended to suggest thresholds or acceptable risk levels for this case study or other environmental health applications, it provides an example for how QRF results could be presented in real-world operations to inspire better-informed decisions.

## 4.1 Introduction

Machine learning and other complex modeling techniques have proven to be effective for predicting and forecasting various types of environmental responses (Kanevski et al., 2004; Urquhart et al., 2012; Crisci et al., 2012; Dou & Yang, 2018; Shen et al., 2019). Many of these models are designed to provide insight for a specific purpose or problem, from informing decision-makers on a local scale to influencing the outcome of global negotiations (Beven 2007, p. 27). However, it can be difficult to present the results from these complex models to the end-users and decision-makers for whom they were intended. A challenge for scientists has been effective communication of the uncertainty associated with the output, which is important when considering environmental management action (Busch et al., 2015; Uusitalo et al., 2015). Communicating uncertainty to decision-makers properly and successfully allows them to evaluate the potential economic or public health consequences of their decision, making their evaluations better informed.

An example of a public health issue in which decision-makers have begun to rely on complex environmental modeling output is the prevention of illness associated with *Vibrio parahaemolyticus*. *Vibrio parahaemolyticus* is a naturally occurring bacterium in brackish waters that is considered one of the most common causes of seafood-borne illnesses in the United States (Colwell et al., 1977; Scallan et al., 2011). Infection most commonly occurs through the consumption of contaminated raw shellfish, with gastroenteritis the main symptom of illness. The U.S. Centers for Disease Control reported that the rate of *Vibrio spp.* illnesses has been increasing over the past three decades, with particularly large numbers of incidence reported in 2013 and 2018 (U.S. Centers for Disease Control and Prevention, 2019). Prediction and forecasting of the bacterium's abundance in shellfish harvesting waters



could help to reduce the public's exposure to harmful concentrations of the bacteria by informing decision-makers to close a particular harvesting site or to impose strict post-harvesting measures to reduce bacterial growth. Increasingly complex statistical and machine learning models have improved *V. parahaemolyticus* prediction (Davis et al, 2017; Davis et al., 2019; Chapter 3). While these models have begun to incorporate a variety of environmental predictor variables shown to be important for the bacterium's prediction, there remains a need to effectively translate model output as valuable information for both the shellfish harvesters and regulatory decision-makers. Providing a measure of uncertainty with model predictions can aid these users in evaluating the implications of over-prediction or under-prediction from both the public health and harvesting industry perspectives.

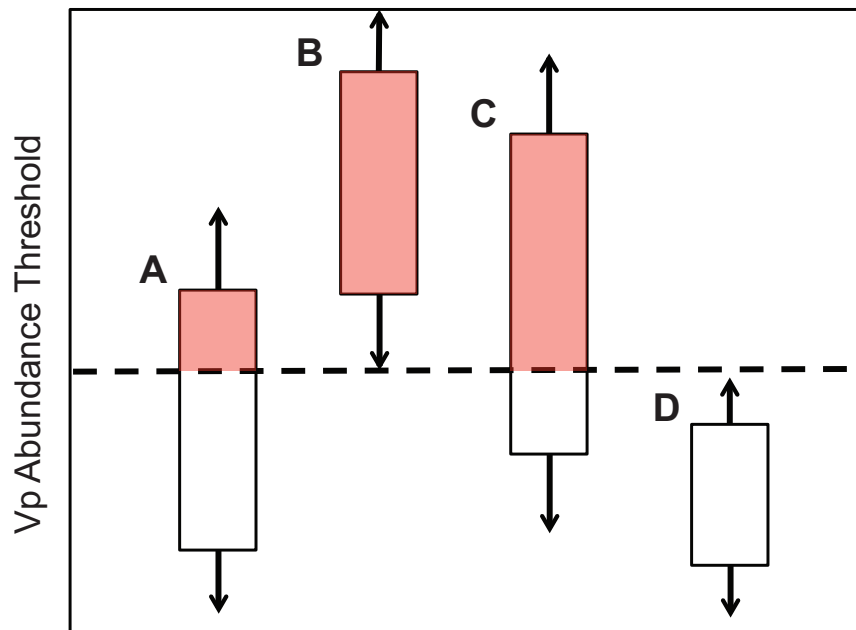
This study uses *Vibrio parahaemolyticus* prediction in the Chesapeake Bay as a case study to examine the potential of Quantile Regression Forests (QRF) for communicating scientific model output to users and decision-makers. Quantile Regression Forest (QRF) described in Meinshausen (2006) has been shown to be useful for various environmental applications (Vaysse & Lagacherie, 2017; Bhuiyan et al., 2018; Bogner et al., 2019). QRFs are an extension of an ensemble decision tree-based model known as a Random Forest (RF) that can be used for both classification and regression (Breiman, 2001). A QRF differs from a RF regression model because it provides a prediction interval that represents the full conditional distribution of response values from the ensemble, whereas a RF provides only one prediction value representing the mean of response values. This can be particularly useful when the mean prediction differs from the median prediction due to some skew in the prediction distribution. While predictions can be made from the median (50<sup>th</sup> quantile) values, they can also be made using a user-defined quantile value. For example, the 5<sup>th</sup>

quantile prediction may be used in applications where there is a need for predictions to be conservative. In this case, the predicted output would be lower than most of predicted outcomes. In contrast, a more radical prediction could be made at the 95<sup>th</sup> quantile, where output would be higher than most the distribution of predicted outcomes.

The uncertainty provided by a QRF differs from the uncertainty that can be provided by other regression models because it is determined for each individual prediction point, known as a prediction interval. In contrast, the uncertainty for a typical regression model is defined by a confidence interval, which is a distribution of estimates of the true population mean of the predictions. This can be useful for applications where individual predictions and their uncertainty need to be communicated to individual users or decision-makers, such as providing predictions for specific shellfish harvesting beds.

The prediction intervals can also be useful for applications that need to communicate the risk of a prediction exceeding a threshold. An example of this applied to shellfish harvesting area closures is shown in Figure 4.1. In the figure, some threshold of *V. parahaemolyticus* abundance is determined (dashed black line). Prediction interval distributions are represented by boxplots, where the 25<sup>th</sup> and 75<sup>th</sup> quartiles define the lower and upper limits of the box and the tails (arrows) represent the range of the distribution. While a RF model would deliver the mean value prediction for a given harvesting area, the QRF model can now provide the percent of the distribution that is above the abundance threshold (red shaded box). Figure 4.1A shows a scenario where 30% of the prediction distribution is above the threshold, while Figure 4.1C show a scenario where 75% of the predictions are above the threshold. It is also a plausible scenario for 100% of the distribution (Figure 4.1B) or 0% (Figure 4.1D) to be above the abundance threshold. A higher percentage of the distribution being above the

threshold indicates a higher risk for that shellfish bed. A risk-based threshold can then be used to determine harvesting area closures. For example, a decision-maker could set a percent level of 50% where a harvesting area is closed if 50% of the predicted distribution is above the bacterial abundance threshold.



**Figure 4.1** Schematic of possible output scenarios (A-D) from a Quantile Regression Forest (QRF), showing how the prediction interval (boxplots) can be used to communicate the risk (red shaded area) of the prediction exceeding a user-defined *Vibrio parahaemolyticus* (Vp) threshold (dashed black line).

The Chesapeake Bay, a large estuary in the eastern United States, is an ideal test bed for modeling studies due to its extensive history of water quality and aquatic biome monitoring. Chesapeake Bay water quality variables used in this study to model *V. parahaemolyticus* are

derived from satellite remote sensing, which is able to provide continuous data coverage throughout our 30-week period in 2010. Scenarios that could be applicable to a decision-maker responsible for the closure of a shellfish harvesting area are investigated by varying the acceptable level of uncertainty in predictions or a bacterial abundance threshold. While it is still unclear whether the prediction of *V. parahaemolyticus* in surface waters used here has a direct relationship with the concentration of the bacterium in shellfish (Nilsson et al., 2019), and therefore whether it is the most appropriate environmental indicator to apply for operational harvesting decisions, these case study results have the potential to influence environmental and public health communication beyond this specific application.

## 4.2 Data Description

### 4.2.1 *In situ* *Vibrio parahaemolyticus* measurements

*In situ* sampling for *Vibrio parahaemolyticus* enumeration was conducted during the months April, July, and October for years 2007 through 2010. Surface water samples (0.5 m depth) were taken at 148 monitoring stations throughout the Chesapeake Bay and its major tributaries over this time period. Total *V. parahaemolyticus* abundance in units of genomic equivalents of colony-forming units (CFU) per milliliter (GE/ml) was determined via quantitative polymerase chain reaction (qPCR) with a limit of detection of 0.14 CFU/ml. Many of the water samples (80.4%) did not have *V. parahaemolyticus* present in the sample or had abundances below the instrument's limit of detection. Along with bacterial analysis, these same samples were also analyzed for numerous water quality parameters. Further detail on these measurements can be found in Jacobs et al. (2014) and Davis et al. (2017).

#### 4.2.2 Satellite data processing

Satellite data used in this study was derived from the Moderate Resolution Imaging Spectroradiometer on the Aqua platform (MODIS-Aqua). Level 1A MODIS Aqua (R2018) data over the Chesapeake Bay was downloaded from NASA's ocean color archive (<http://ocean.color.gsfc.nasa.gov/>) for the 4-year period from 2007 through 2010. This data was batch processed in NASA's SeaDAS (version 7.5) software in order to obtain to binned 8-day composite files for ocean color products and sea surface temperature (SST). Changes to the default processing scheme in SeaDAS, developed for open ocean waters, in order to process coastal Chesapeake Bay waters are described in detail in Chapter 2. These changes to the cloud detection band, high light mask, and straylight mask increase the number of valid MODIS pixels around the Chesapeake Bay's coastlines, within tributaries, and in turbid waters, where *Vibrio* are commonly found. Only 8-day weekly composites for months March through October (30 weekly composites in total) were used to overlap with the *in situ* measurement dates.

Ocean color products were derived from MODIS weekly composites, including chlorophyll-a, total suspended solids (TSS), and sea surface salinity (SSS). The algorithms used to derive these products are found in Werdell et al. (2007), DeLuca et al. (2018), and Urquhart et al. (2012), respectively. Sea surface temperature (SST) is derived from MODIS by transforming thermal infrared radiances into degrees Celsius (°C) using the Planck function.

#### 4.2.3 Shellfish harvesting areas

Shellfish harvesting areas in this study were obtained from the Maryland Department of Natural Resources (<https://gisapps.dnr.state.md.us/AquaDocs/coordinates.pdf>) and the Virginia Marine Resources Commission (<https://mrc.virginia.gov/Regulations/FR720.shtm>). Here we represent the polygonal Chesapeake Bay shellfish lease sites as single point locations that are based on a corner of the recorded boundaries. Ninety-eight shellfish harvesting areas located throughout the Chesapeake Bay and major tributaries are used in the subsequent analyses. Satellite data were matched to these shellfish harvesting locations using a nearest neighbor approach within the R package “*yaImpute*” (Crookston & Finley, 2007; R Core Team, 2013). Each of the 98 harvesting areas were therefore matched with one satellite pixel for each of the 30 weeks in the study.

### 4.3 Methods

#### 4.3.1 Modeling methodology

*Vibrio parahaemolyticus* data are commonly zero-inflated, with non-detects being the majority class and detected presence comprising only 20% of the *in situ* dataset. To account for the zero-inflation, we use a two-step approach to modeling bacterium abundance from the satellite data. This methodology is explained in detail in Chapter 2. For this study, we use a binary classification model to first predict whether the bacterium is present or absent at the harvesting site. For those sites where presence is predicted, a quantile regression model then predicts bacterial abundance. Both of these models are trained using the satellite-*in situ* matched dataset described in Chapter 3 and tested on the

30-week satellite dataset described in Section 2.2. A Random Forest model was used for classification and a Quantile Regression Forest (QRF) model was used for quantile regression. Both of these models used sea surface temperature, sea surface salinity, total suspended solids and chlorophyll-a as predictor variables. All modeling was performed in the R Statistical Computing Environment (version 3.3.2) using the package “quantregForest” for QRF analyses (R Core team, 2013; Meinshausen, 2017).

#### *4.3.2 Quantile Regression Forest*

QRFs are an extension of a Random Forest (RF), a decision tree-based machine learning model that can be used for both classification and regression (Breiman, 2001). The decision trees in a RF follow rules at data-defined nodes in the covariates in order to predict a response when fed with new data. The RF differs from a classic Classification and Regression Tree (CART) model because it comprises an ensemble of these trees, typically in excess of 500 trees, thereby making a “forest” of predictions. Each tree in this ensemble selects a random subset of training data (bagging) and predictor variables to be used at the decision-making nodes. When used for classification, the predicted response is the class with the majority vote from all trees in the forest. When doing regression, the predicted response is the mean of the predictions made by the ensemble of trees. The strengths of a Random Forest include the ability to manage non-linear and non-parametric data, typically high predictive performance, the lack of intricate tuning of parameters, and robustness even when handling noisy predictor data (Tyralas et al., 2019). A QRF model is an extension of the RF regression model because it stores and outputs information from all of the trees in the ensemble. Instead of providing one

prediction value representing the mean of all predictions, a QRF provides information about the distribution of all predictions in the ensemble and can produce prediction values using the median (50<sup>th</sup> quantile) or another quantile from the distributions. While the QRF has this flexibility in prediction options, they can overfit a model and should be used in conjunction with cross-validation.

#### 4.3.3 Model evaluation

We evaluate the performance of the QRF model on the satellite-*in situ* matched *V. parahaemolyticus* dataset described in Chapter 3 using a Continuous Ranked Probability Score (CRPS) (Brown, 1974; Matheson & Winkler, 1976; Unger, 1985). The CRPS is a statistical tool that compares a distribution of predictions or a probabilistic forecast to an observed outcome. In order to calculate the value of the CRPS, both the prediction distribution and the outcome are represented as cumulative distribution functions (CDF). The total area between the two CDFs is the CRPS value given in units of the predicted variable. The CRPS for a deterministic prediction corresponds to a Mean Absolute Error (MAE) evaluation. A CRPS value of zero indicates perfect predictive accuracy.

Here the CRPS was used to evaluate QRF model performance using a Monte Carlo holdout scheme. The satellite-*in situ* dataset was randomly split into 80% training data and 20% holdout data over 100 iterations. The two-step classification and regression approach described in Section 3.1 provided QRF prediction distributions for all data where presence was predicted. A CRPS was calculated for each of the data in an iteration of the Monte Carlo scheme and the mean CRPS value from that iteration was recorded.



The mean CRPS from all 100 iterations is then calculated and represents the performance of our QRF model.

## 4.4. Results and Discussion

### 4.4.1 Model performance

The performance of our QRF model on the satellite-*in situ* matched dataset was evaluated using a Continuous Ranked Probability Score (CRPS), where a score of zero indicates perfect model performance because the cumulative distribution functions of the predictions and observed outcomes are equal. Over the 100 Monte Carlo iterations, the range of CRPS from the QRF model was 1.16 to 2.26 GE/ml with a mean value of 1.59 GE/ml. This mean CRPS value represents 8.98% of the range of *V. parahaemolyticus* abundances (0.40 to 17.71 GE/ml) in the satellite-*in situ* dataset. We determine this to be an acceptable level of performance for this study. A full evaluation of the Random Forest approach upon which the QRF model is an expansion can be found in Chapter 3.

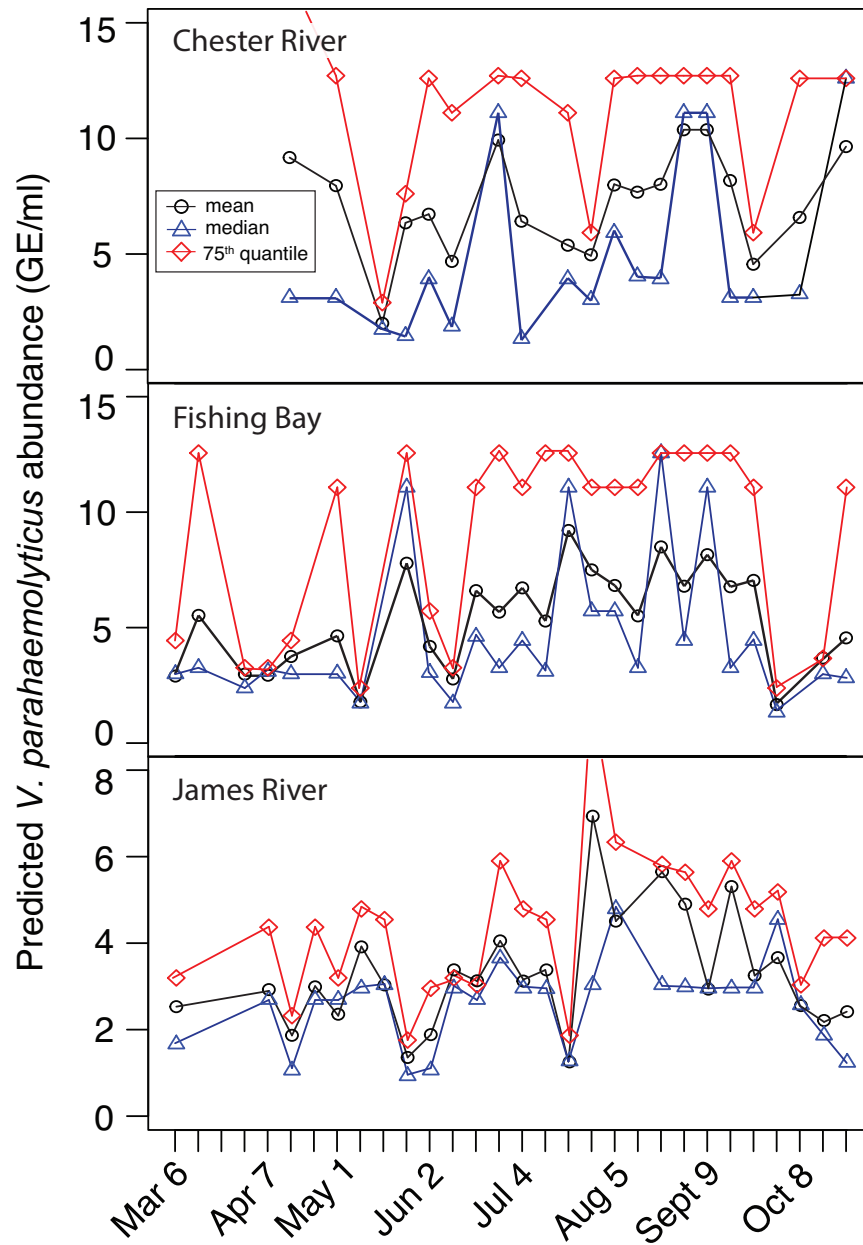
### 4.4.2 Mean versus quantile predictions

One advantage of quantile regression techniques is the ability to predict outcomes at a specific quantile instead of the mean (Cook & Manning, 2013). While a skewed distribution or outlier can influence the mean prediction value, the median (50<sup>th</sup> quantile) prediction value is less susceptible to those effects. To investigate the utility of using a mean versus quantile prediction value, we chose one harvesting area from three locations in various regions of the Chesapeake Bay where a large percentage of samples in the *in situ* dataset had detected the presence of *V. parahaemolyticus* (Davis et al., 2017).

A time series of predicted bacterial abundance at the three locations, the Chester River, Fishing Bay, and the James River, is shown in Figure 4.2. The length of the time series for each location depends on the number of weeks in which the bacterium was predicted to be present using the binary classification model described previously. The Fishing Bay and James River harvesting areas saw predicted bacterial presence in March, while the Chester River area did not have predicted presence until late-May. The Chester River harvesting sees fewer weeks with predicted presence than either of the other two harvesting areas. The increase in the number of weeks of predicted bacterial presence between harvesting areas in the Upper Bay (Chester River) and the Mid to Lower Bay Chesapeake Bay (Fishing Bay and James River) may be due to the north-south gradient of SST in the Chesapeake Bay corresponding with the seasonal cycle of temperature conditions in which *V. parahaemolyticus* thrive.

In general, the median prediction values are lower than the mean prediction values. This is particularly apparent during the summer months of June through September, when notable differences between the mean and median prediction values occur in the Chester River and Fishing Bay. This difference is also observed to a lesser degree in the James River from August to September. Also shown in Figure 4.2 is the 75<sup>th</sup> quantile prediction, representing the high end of the prediction distribution from all of the trees in the Quantile Regression Forest.

The 75<sup>th</sup> quantile predictions illustrate how the mean prediction can be influenced by the shape of the prediction distribution throughout the 30-week time series. Where the mean prediction values are higher than the median values in summer weeks, the 75<sup>th</sup> quantile prediction values show a larger range of predictions than during weeks where the



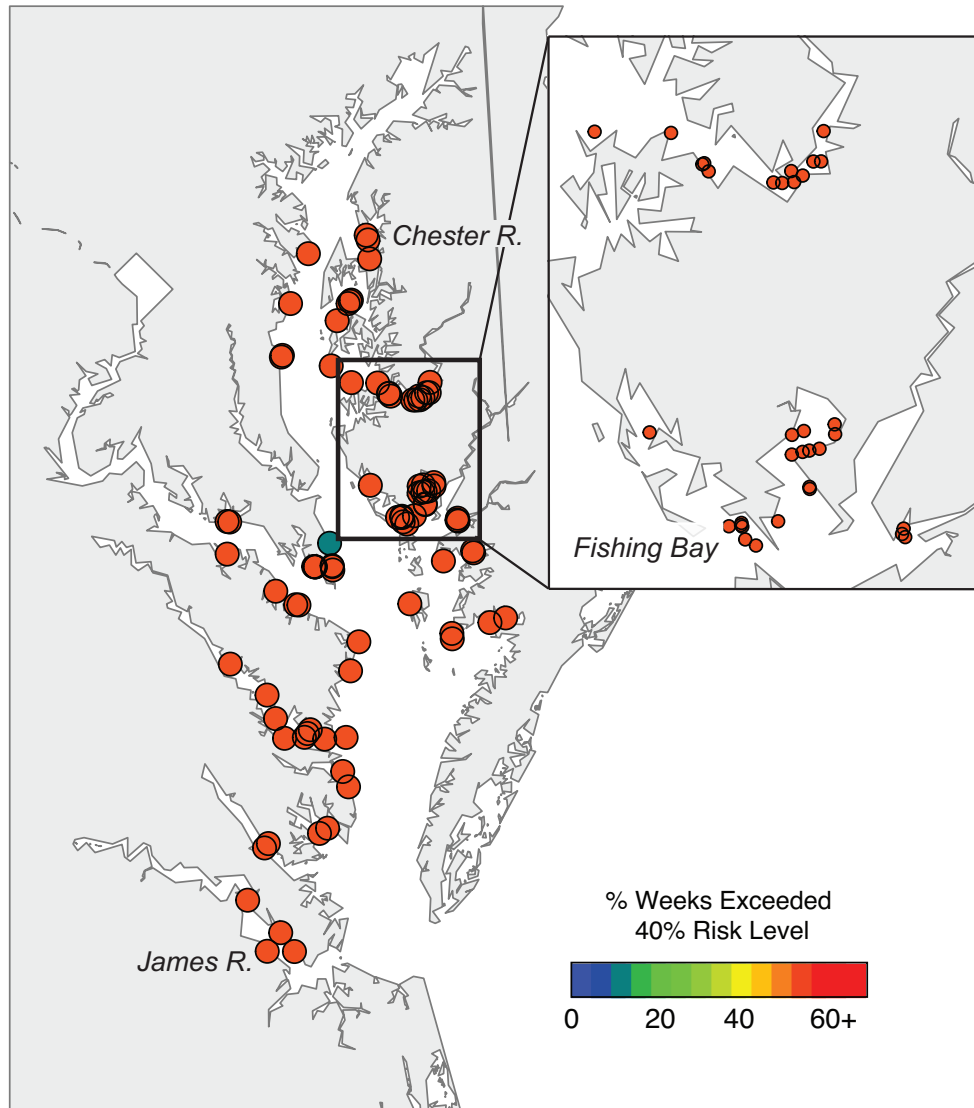
**Figure 4.2** Time series of predicted *V. parahaemolyticus* abundance (GE/ml) at the Chester River, Fishing Bay, and the James River over 30 weeks in 2010 showing mean RF prediction values (black circles), 50<sup>th</sup> quantile or median QRF prediction values (blue triangles), and 75<sup>th</sup> quantile prediction values (red diamonds). The number of data points for each location indicates the number of weeks in which the bacterium was predicted to be present.

mean and median values are more similar. This suggests that the models are producing a larger range in predictions, and therefore more uncertainty, during these summer weeks than at other times of the year. This could be important to consider when delivering abundance predictions to users and decision-makers during this summer period when illness cases from *V. parahaemolyticus* increase (U.S. Centers for Disease Control, 2013). Mean predictions could present an abundance value that garners more cautionary action for a harvesting area than would a median prediction.

#### *4.4.3 Scenarios based on level of risk*

Quantile predictions provide the ability to communicate the risk of a prediction exceeding some user-defined threshold, as described in Section 3.2. The risk levels in the following scenarios represent the percentage of the prediction distribution above a arbitrary threshold of bacterial abundance (1.5 GE/ml). In Figures 4.3 – 4.5, we examine how often shellfish harvesting areas throughout the Chesapeake Bay would be closed over the 30-week period in 2010 for three theoretical levels of acceptable risk. For example, the percentage of closures in Figure 4.3 indicates the frequency of a harvesting area having an abundance prediction distribution in which 40% of the distribution is above the 1.5 GE/ml threshold.

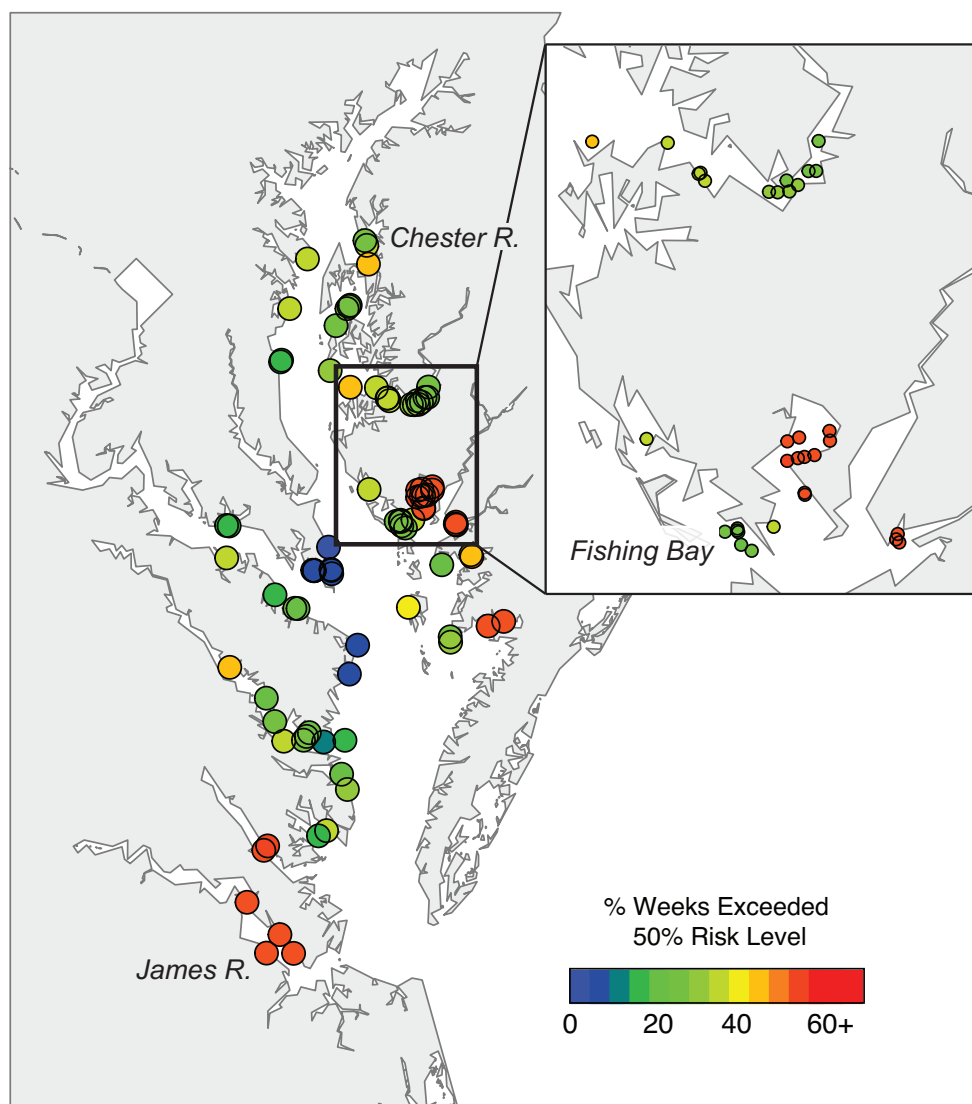
When the risk level is held at 40% (Figure 4.3), all but one of the harvesting areas in the Bay are closed more than 60% of the 30-week period. This 40% risk level scenario is the most conservative of the three scenarios indicating that either harvesting area closures or stricter post-harvesting measures would be implemented often throughout the Bay.



**Figure 4.3** Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 40% level of acceptable risk, where at least 40% of the QRF prediction distribution is above a 1.5 GE/ml threshold, over a 30-week period in 2010.

A similar mapping of the frequency of harvesting areas closures using a 50% risk level is shown in Figure 4.4. Here, at least half of the prediction distribution must be above the bacterial abundance threshold in order to implement harvesting areas closures

or strict post-harvesting measures. A 50% risk level shows more variation in how often harvesting areas are subject to action than did a 40% risk level. There are several areas that are affected more than 60% of the 30-week time period, located primarily in Fishing Bay and the James River. Many other areas exceed the 50% risk level for 30-40% of the time period, while several areas along the western shore of the Mainstem do not exceed this level more than 5% of the time period.



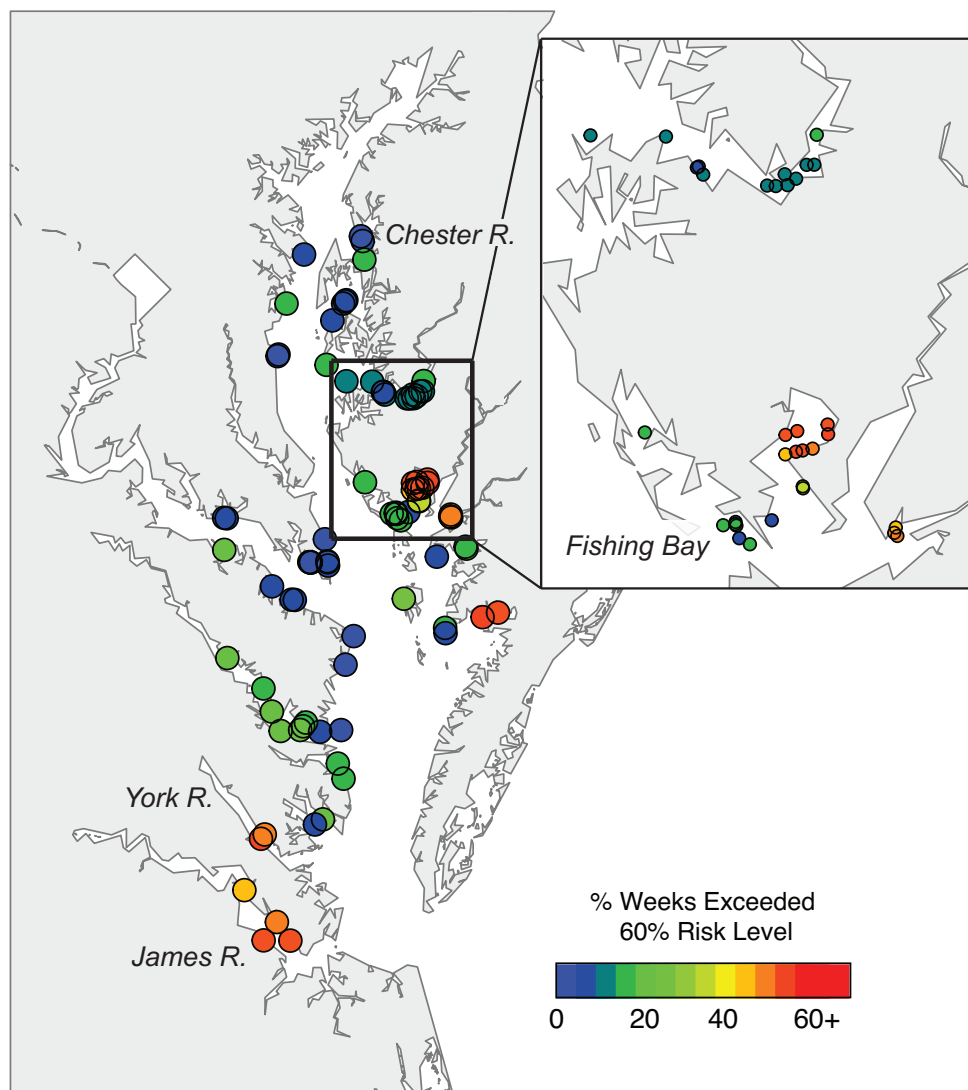
**Figure 4.4** Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 50% level of acceptable risk, where at least 50% of the QRF prediction distribution is above a 1.5 GE/ml threshold, over a 30-week period in 2010.

The highest percent risk level explored in this study (60%), representing the highest risk tolerance and greatest willingness to keep harvesting areas active, is shown in Figure 4.5. In this scenario, most of the harvesting areas exceed the risk level less than 30% of the time period. The several harvesting areas that exceed the risk level more often are located in the James River, Fishing Bay, and the York River.

We then zoomed in to the three harvesting sites evaluated in Section 4.1 to examine how the frequency of weeks that exceed the risk level changes across the entire range of risk levels for different sections of the Chesapeake Bay (0% to 100%) (Figure 4.6). Here, a 0% risk level indicates that a harvesting area would be closed if any portion, or none, of the quantile prediction distribution is above the *V. parahaemolyticus* abundance threshold of 1.5 GE/ml. A 100% risk indicates that a harvesting area would be affected only if the entire quantile prediction distribution is above the *V. parahaemolyticus* threshold. These values are shown in Figure 4.6 for completeness, but are excluded from the analysis below.

The Chester River harvesting site located in the Upper Bay is affected most frequently from risk level levels of 10% to 30%, where the frequency of weeks that exceed the risk level remains constant at 57%. The frequency then gradually slopes downward as the percent risk level increases to 80%. The following 90% risk level would not close Chester River harvesting area during any weeks in the period. Fishing Bay and

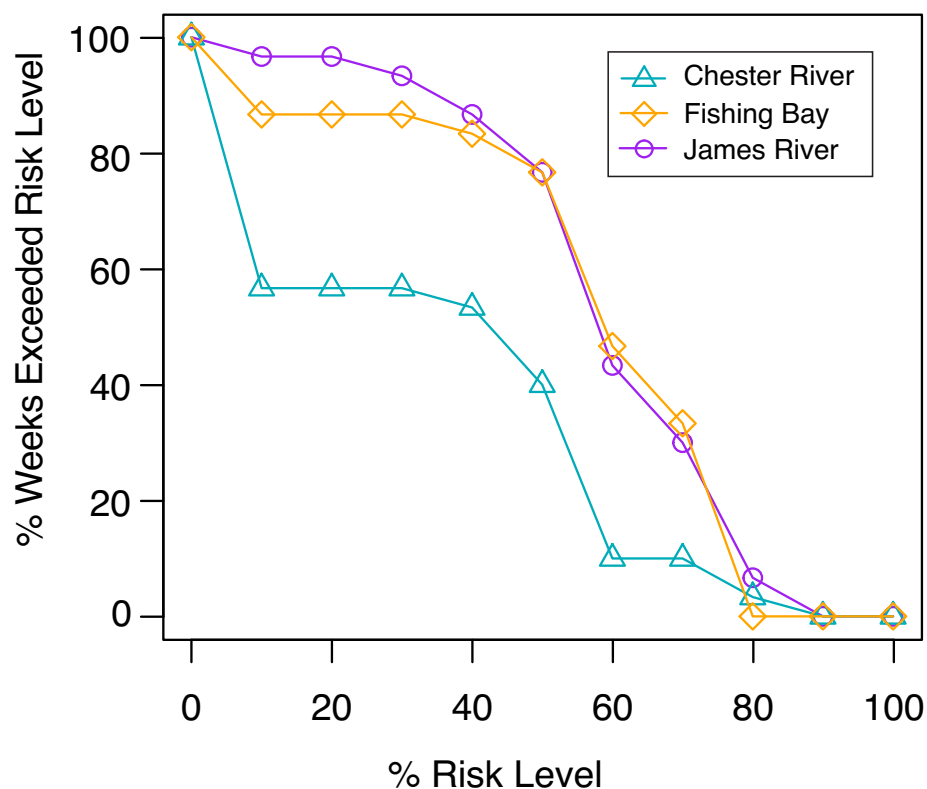
the James River are more frequently affected (87% and 97% of the time period, respectively) at low risk levels than is Chester River. The frequency of closures begins decreasing at the 30% risk level at the James River site, whereas the frequency begins to decrease at 40% risk at the Fishing Bay site. The frequency gradually decreases until an 80% risk level at Fishing Bay and 90% risk level at James River, where both harvesting areas not closed for any portion of the time period.





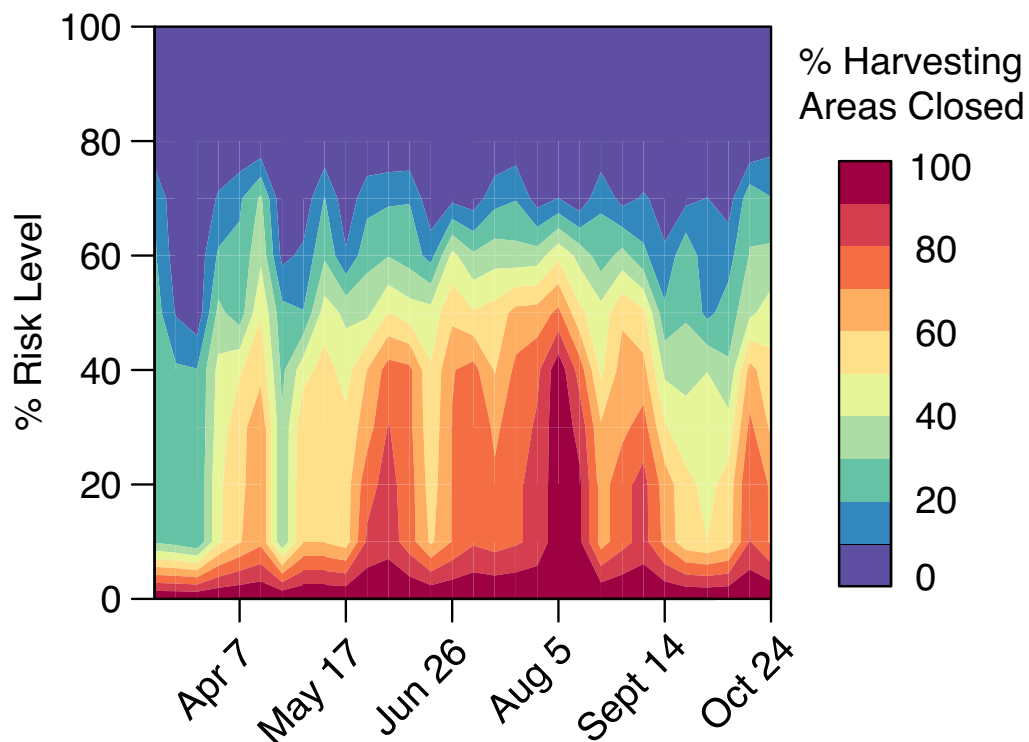
**Figure 4.5** Map of the frequently that shellfish harvesting areas throughout the Chesapeake Bay would be affected by a 60% level of acceptable risk, where at least 60% of the QRF prediction distribution is above a 1.5 GE/ml threshold, over a 30-week period in 2010.

The previous figures show that the frequency of a harvesting area exceeding some percent risk level can vary by the enforced percent risk level decided upon and can also vary through space in different regions of the domain. This variation could affect whether decisions about an acceptable level of risk are made Chesapeake Bay-wide, regionally, or more locally for each harvesting area. We then removed the factor of spatial variation and examined how many of the harvesting areas would be closed or affected by different percent risk levels chosen throughout the 30-week period (Figure 4.7).



**Figure 4.6** Frequency of shellfish harvesting areas at Chester River (turquoise triangles), Fishing Bay (orange diamonds), and James River (purple circles) affected by a range of risk levels from 0% to 100%. The risk level indicates the percentage of the QRF prediction distribution that is above a 1.5 GE/ml abundance threshold.

The percentage of closed harvesting areas increases during the summer months of June through September, when an increase in *V. parahaemolyticus* concentrations have also been observed. Over 90% of the harvesting locations would be closed during August when using up to a 42% risk level. As the risk level increases past 50% in the summer period, the number of closed sites gradually decreases until the 80% risk level when there would be zero sites are closed. A majority (> 50%) of the harvesting areas would be closed during the 30-week time period for any risk level below 50%.

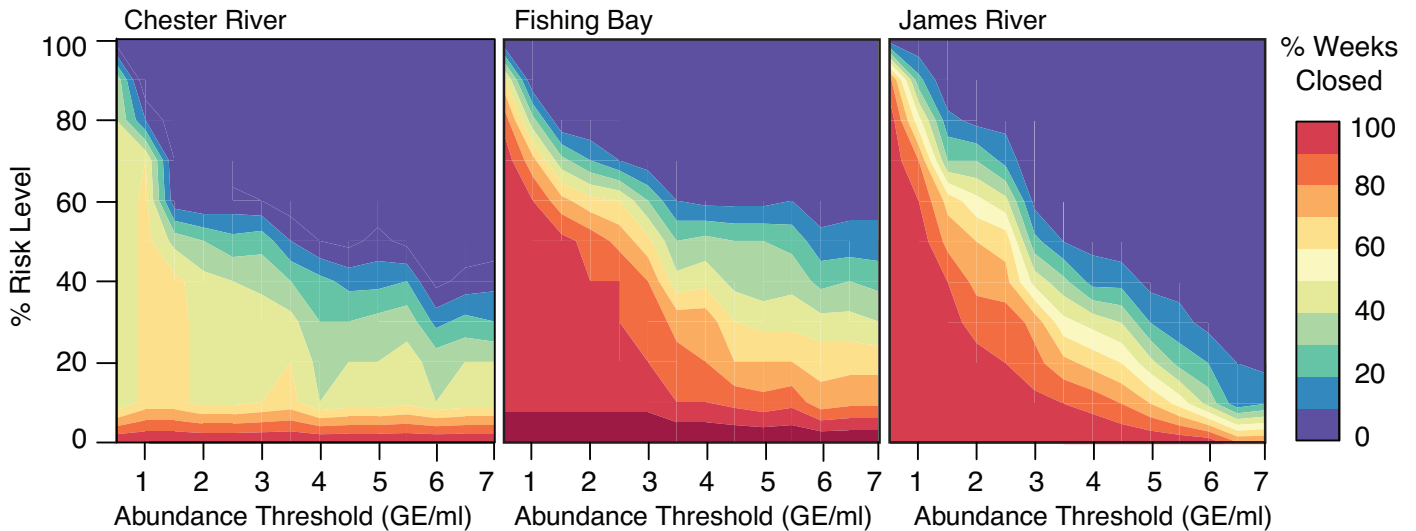


**Figure 4.7** Contours showing the proportion of Chesapeake Bay shellfish harvesting areas affected by a range of risk levels over the 30-week period in 2010. The risk level indicates the percentage of the QRF prediction distribution that is above a 1.5 GE/ml abundance threshold.

#### 4.4.4 Scenarios varying abundance threshold and level of risk

Finally, we examine scenarios where both the level of risk and *V. parahaemolyticus* abundance threshold could be chosen to determine how often harvesting areas would be closed or affected by variations in those decisions. Figure 4.8 shows this analysis for the three harvesting area sites used previously: Chester River, Fishing Bay, and James River. At the James River site, there is a distinct relationship between the shellfish manager's two decisions. As the bacterial abundance threshold and percent risk level decrease, the harvesting area is closed less frequently. Fishing Bay shows a similar relationship up to an abundance threshold of about 3.5 GE/ml, when the frequency of closure at each risk level appears to plateau. The Chester River harvesting area is closed less frequently than the other two sites at all percent risk levels above 10%. At low abundance thresholds (< 2.0 GE/ml), the Chester River site is closed up to 60% of the time period at risk levels up to 70%. At higher abundance thresholds, the Chester River harvesting area is closed 40% of the period at risk levels below 20%. Similar to the Fishing Bay site, frequency of closures at the Chester River site appears to plateau at risk levels less than 40% with increasing abundance thresholds. Prior to implementing any policy based on *V. parahaemolyticus* abundance, then, decision makers would need to consider these two dimensions of risk: what concentration of bacteria is considered to be

dangerous, and what risk of realizing that concentration is acceptable before a harvesting area needs to be closed.



**Figure 4.8** Contours showing the frequency of shellfish harvesting areas at Chester River, Fishing Bay, and James River affected by a range of risk levels from 0% to 100% and a range of abundance thresholds from 0.5 GE/ml to 7 GE/ml over the 30-week period in 2010. The risk level indicates the percentage of the QRF prediction distribution that is above the abundance threshold.

## 4.5 Conclusions

This study examines a method for communicating environmental model predictions to users and decision-makers using a Quantile Regression Forest. A Quantile Regression Forest model provides information about the distribution of each prediction made by the ensemble of decision trees in the Random Forest. While a Random Forest model only provides the mean prediction from these trees, a Quantile Regression Forest is able to provide predictions

from the median or another specific quantile from the distribution that is less likely to be affected by outliers. A point estimate can be a valuable prediction for some applications, but another potential use of this method is to provide a degree of uncertainty to the user or decision-makers. These end-users may require information about the likelihood of the model's prediction exceeding some biological or public health-based threshold, which can be determined by the proportion of the prediction distribution from the Random Forest ensemble that is above their threshold.

We applied the Quantile Regression Forest method to a case study in the Chesapeake Bay, where the abundance of *Vibrio parahaemolyticus* at shellfish harvesting areas poses a public health concern. The QRF model output showed acceptable performance on a dataset with known *V. parahaemolyticus* abundances using a Continuous Ranked Probability Score. The median (50<sup>th</sup> quantile) predictions of *V. parahaemolyticus* abundance were generally lower than the mean predictions for a satellite-derived dataset matched to known harvesting area locations. This indicates that the mean prediction values, those produced by a Random Forest model, are likely influenced by outliers at the high end of the prediction distribution. This was a particularly important difference in predictions during summer months, when the bacterium's concentration and associated illness cases increase.

We also explored the application of QRF models to a subjective, user-defined *V. parahaemolyticus* abundance threshold and the associated level of risk. The spatial and temporal variation of harvesting areas affected by the abundance threshold changed as the acceptable level of risk, or percentage of the prediction distribution that is above the abundance threshold, was varied. At three specific harvesting areas chosen to represent different regions of the Chesapeake Bay, the frequency of those sites being closed or targeted

for strict post-harvesting measures risk decreased as the acceptable level of risk increased. These three sites also showed an increase in the frequency of being affected as the bacterial abundance threshold and percent risk level decreased.

The results of this study are not meant to suggest a bacterial abundance threshold or risk level for either this case study's application to *V. parahaemolyticus* or other environmental health applications. Indeed, in the specific case of *V. parahaemolyticus* it is not clear that water column abundance is even the optimal variable for informing harvesting decisions. Rather, our study aims to provide an example for how such results can be presented to those who will ultimately be making these decisions in real-world operations. The economic impact and cost-benefit analysis of quantile-based decisions could also be accounted for in future studies that move closer toward using these types of methods operationally. In summary, this study shows how Quantile Regression Forests can be used to make complex modeling output more transparent to users and encourage better-informed decisions.

## 5. CHAPTER 5: CONCLUSIONS

The Chesapeake Bay has a decades-long history of continuous *in situ* water quality measurements. *In situ* sampling is crucial for monitoring the Bay's water quality, ecological processes, and environmental health, but it can lack the temporal and spatial resolution necessary for dynamic estuaries and environmental management applications. Satellite remote sensing can address these limitations by providing continuous, near-daily coverage for the entire Chesapeake Bay estuary. Remotely sensed data can also provide data in near-real time, whereas many *in situ* water quality measurements can be delayed due to the time involved with laboratory measurements. The goal of this dissertation is present methodology and insights for satellite remote sensing to be used as an additional tool for water quality monitoring and the prediction of *Vibrio parahaemolyticus* bacteria in the Chesapeake Bay. Water quality variables like total suspended solids are important to monitor in the Chesapeake Bay because high concentrations can affect submerged aquatic vegetation, pathogen abundance, and habitats for other aquatic life. Water quality parameters are also important for the prediction of *Vibrio* bacteria, which has become increasingly important as the rate of illnesses attributed to these taxa has increased over the last few decades (U.S. Centers for Disease Control and Prevention, 2019).

In Chapter 2, I used 14-years of MODIS coverage of the Chesapeake Bay to validate and potentially improve satellite retrieval algorithms for total suspended solids concentration. I investigated whether unexploited spectral information, combined with advanced statistical and machine learning models, could improve MODIS-derived total suspended solids (TSS) concentration retrieval. Compared to a conventional algorithm that used only red reflectance band information and polynomial regression, I found that the additional of other MODIS spectral

bands in a Random Forest model could improve the accuracy of TSS retrievals in more turbid waters. For applications where turbid waters provide a biological or ecological importance, such as the modeling of *Vibrio* bacteria, this result could significantly improve predictions. However, the multi-band model showed little improvement over the conventional single-band algorithm in less turbid waters. I found that the machine learning method alone provided some of the improvement over the polynomial algorithm, but the additional of multispectral information further improved the model's performance. The study also showed that both the multispectral and single-band algorithms were similarly generalizable throughout various partitions of the Chesapeake Bay and times of year. Multispectral sensors are expensive to build and fly aboard large spacecraft relative to sensors with few bands that can fly aboard small spacecraft. Therefore, it is important to understand if and how the addition of spectral bands contributes to environmental monitoring of TSS and other parameters in optically complex waters like those in the Chesapeake Bay.

Chapter 3 investigated whether remotely sensed water quality variables in addition to SST could improve prediction and forecasting models for *V. parahaemolyticus* concentration in Chesapeake Bay surface waters. To do this I used satellite-derived ocean color (sea surface salinity, the improved total suspended solids algorithm from Chapter 2, and chlorophyll-a) and SST products, including remote sensing reflectances, as inputs to four statistical or machine learning models. Several versions of these models, using multiple structures of input predictors, were tested for their ability to perform classification and regression, predicting bacterial presence and abundance, respectively. Remote sensing data from the week prior to the *in situ* measurements were input to the models to evaluate whether *V. parahaemolyticus* concentration



could be forecasted, which tests the feasibility of an operational early warning system in the future. The additional ocean color products improved the prediction performance of the classification and regression models. The models using remote sensing reflectances as input predictors, thereby bypassing the need to transform reflectances into ocean color products using knowingly imperfect algorithms, also showed potential for use in these models. The ability of remote sensing data to aid in the prediction of *V. parahaemolyticus* and the improvement in predictions shown by the inclusion of multiple water quality variables are valuable information for future development of operational models that environmental managers and shellfish harvesters can use to reduce the risk of *Vibrio*-related illness.

Finally, Chapter 4 proposed a new method of quantifying prediction uncertainty using Quantile Regression Forests, employing remotely sensed *V. parahaemolyticus* predictions in the Chesapeake Bay as a case study. Instead of outputting a point prediction for bacterial concentration for a given data point, Quantile Regression Forests output a prediction interval. These prediction intervals can provide quantile predictions, such as the median (50<sup>th</sup> quantile) prediction, instead of the mean prediction provided by a conventional Random Forest. The prediction interval can also be used to communicate the probability of a prediction crossing a threshold, in this case study a subjective bacterial abundance threshold, for each data point. Using this, environmental management decisions could be better informed and based on a degree of uncertainty in the delivered model prediction or forecast. The study provides an example of how environmental modeling output can be presented to end-users and decision-makers who use the predictions for real-world operations in a way that allows them to visualize how model statistics and thresholds impact decision outcomes.

## 5.1 Future Directions

The work in this dissertation provides a useful foundation for using remote sensing approaches to develop future operational models for the Chesapeake Bay and beyond. In this dissertation, I used remote sensing data from MODIS-Aqua, a sensor that is currently operating well beyond its designed lifetime but has provided over a decade of continuous moderate resolution coverage of Chesapeake Bay waters. The MODIS-Aqua sensor is quickly becoming outdated by the recent launch of several new sensors aboard polar-orbiting satellite platforms that are designed for ocean color studies. Newer sensors can provide improved spatial resolution over the 1-km resolution used in this work, which could provide additional coverage of the Chesapeake Bay in smaller tributaries and near-shore regions. Temporal resolution would be improved because several of these new sensors are a part of a fleet, where multiple versions of the same sensor fly aboard more than one satellite. This allows for two or more times the coverage that is currently possible with the MODIS-Aqua sensor. Geostationary satellites that have orbits designed to keep them positioned over a particular region, providing sub-hourly data, have also been suggested for ocean color remote sensing work in dynamic water systems like the Chesapeake Bay estuary. Improving spectral resolution through the development of hyperspectral sensors is also an active area of research and engineering that has the potential ability to improve retrievals of optically interactive water quality parameters, such as TSS and colored dissolved organic matter (CDOM), and enable the distinction of different species of plankton.

Future modeling efforts could incorporate a combination of *in situ* data and remotely sensed data, along with hydrodynamic model output. The incorporation of these

different data types is known as data assimilation, which is an active area of research in the hydrology and weather forecasting fields (Ghil & Malanotte-Rizzoli, 1991; Hannart et al., 2016; Pathiraja et al., 2016; Li et al., 2018). The combined efforts of various model types have the potential to improve the spatial and temporal resolution of output. The accuracy of predictions can also be improved from the incorporation of “ground-truth” *in situ* measurements. Geostatistical kriging, where data points closer to each other are assumed to be more similar, is also able to provide model estimates in higher spatial and temporal resolutions than *in situ* or remote sensing data are able to on their own. Kriging can address the spatial inconsistencies in *in situ* data and remotely sensed data blocked by cloud cover, and it can be used to estimate the output for gaps in time between satellite overpasses.

*Vibrio parahaemolyticus* models would benefit largely from the availability of more *in situ* measurements that can be incorporated into model development. While the *in situ Vibrio parahaemolyticus* dataset used here is one of the most comprehensive for the Chesapeake Bay, a majority of the samples had bacterial concentrations below the limit of detection of the instrument. Newer instrumentation and updated procedures may also increase the number of bacterial measurements from surface water samples.

Another factor to consider for *V. parahaemolyticus* predictions is that the primary purpose for modeling this particular species of *Vibrio* is to prevent the incidence of gastrointestinal illness from the consumption of raw shellfish. While our models predict the abundance of the bacterium in surface waters, the abundance in the shellfish is what determines the risk of illness after ingestion. Oysters accumulate the bacterium through filtering these waters, but it is still unclear whether there is a relationship between the

concentrations in the two mediums (Nelisson et al., 2019). Additionally, the oyster-to-oyster variation in *V. parahaemolyticus* concentrations due to differences in physiology, immune response, and other individual functions has not yet been well studied. Therefore, organisms in the same harvesting area could have vastly different concentrations of the bacterium even if surface water abundance predictions are correct. However, the prediction of bacterial abundance in surface waters would certainly apply to other *Vibrio* species, such as *Vibrio vulnificus*, where a main mode of infection is through contact with open wounds (Klontz et al., 1988).

While the models presented in this dissertation were not intended to be suitable for immediate operational use, the findings from each chapter could aid in the development of future operational modeling efforts that incorporate satellite remote sensing data. A summary of recommendations for future models is as follows: 1) multispectral information and machine learning should be considered for water parameter retrievals, particularly in optically complex water like those found in the Chesapeake Bay, 2) water quality variables in addition to SST can improve remotely sensed *Vibrio* predictions, 3) environmental models using remote sensing reflectances can bypass the need to develop algorithms for future sensors and avoid the introduction of additional error, and 4) models that provide prediction intervals, such as Quantile Regression Forests, can provide the uncertainty information necessary for better informed environmental management decisions. Not only are the conclusions and suggestions from this dissertation applicable to the Chesapeake Bay, but I also anticipate that they would be helpful for modeling efforts in other water bodies with optically complex waters and environmental health concerns.

## REFERENCES

- Aurin, D., Mannino, A., & Franz, B. (2013). Spatially resolving ocean color and sediment dispersion in river plumes, coastal systems, and continental shelf waters. *Remote Sensing of Environment*, 137, 212-225.
- Bailey, S. W., Franz, B. A., & Werdell, P. J. (2010). Estimation of near-infrared water-leaving reflectance for satellite ocean color data processing. *Optics Express*, 18(7), 7521-7527.
- Baird, D., & Ulanowicz, R. E. (1989). The seasonal dynamics of the Chesapeake Bay ecosystem. *Ecological Monographs*, 59(4), 329-364.
- Baker-Austin, C., Trinanes, J. A., Taylor, N. G., Hartnell, R., Siitonen, A., & Martinez-Urtaza, J. (2013). Emerging *Vibrio* risk at high latitudes in response to ocean warming. *Nature Climate Change*, 3(1), 73.
- Banakar, V., De Magny, G. C., Jacobs, J., Murtugudde, R., Huq, A., Wood, R. J., & Colwell, R. R. (2011). Temporal and spatial variability in the distribution of *Vibrio vulnificus* in the Chesapeake Bay: A hindcast study. *EcoHealth*, 8(4), 456-467.
- Beven, K. (2018). *Environmental modelling: An uncertain future?*. CRC press.
- Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Quintana-Seguí, P., & Barella-Ortiz, A. (2018). A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula. *Hydrology and Earth System Sciences*, 22(2), 1371.
- Bogner, K., Pappenberger, F., & Zappa, M. (2019). Machine learning techniques for predicting the energy consumption/production and its uncertainties driven by meteorological observations and forecasts. *Sustainability*, 11(12), 3328.
- Boyd, C. E. (2015). *Water Quality: An Introduction*. Springer.

- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Chapman & Hall. *New York*.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Brown, T. A. (1974). Admissible Scoring Systems for Continuous Distributions.
- Brown, C. W., Connor, L. N., Lillibridge, J. L., Nalli, N. R., & Legeckis, R. V. (2007). An introduction to satellite sensors, observations and techniques. In *Remote Sensing of Coastal Aquatic Environments* (pp. 21-50). Springer, Dordrecht.
- Brush, G. S. (1989). Rates and patterns of estuarine sediment accumulation. *Limnology and Oceanography*, 34(7), 1235-1246.
- Bukata, R. P., Jerome, J. H., Kondratyev, A. S., & Pozdnyakov, D. V. (2018). *Optical properties and remote sensing of inland and coastal waters*. CRC press.
- Busch, D. S., O'Donnell, M. J., Hauri, C., Mach, K. J., Poach, M., Doney, S. C., & Signorini, S. R. (2015). Understanding, characterizing, and communicating responses to ocean acidification: Challenges and uncertainties. *Oceanography*, 28(2), 30-39.
- Caburlotto, G., Haley, B. J., Lleò, M. M., Huq, A., & Colwell, R. R. (2010). Serodiversity and ecological distribution of *Vibrio parahaemolyticus* in the Venetian Lagoon, Northeast Italy. *Environmental Microbiology Reports*, 2(1), 151-157.
- Carder, K. L., Cannizzaro, J. P., & Lee, Z. (2005, August). Ocean color algorithms in optically shallow waters: Limitations and improvements. In *Remote Sensing of the Coastal Oceanic Environment* (Vol. 5885, p. 588506). International Society for Optics and Photonics.

- Chen, S., Huang, W., Chen, W., & Chen, X. (2011). An enhanced MODIS remote sensing model for detecting rainfall effects on sediment plume in the coastal waters of Apalachicola Bay. *Marine Environmental Research*, 72(5), 265-272.
- Chen, J., D'Sa, E., Cui, T., & Zhang, X. (2013). A semi-analytical total suspended sediment retrieval model in turbid coastal waters: A case study in Changjiang River Estuary. *Optics Express*, 21(11), 13018-13031.
- Chesapeake Bay Program (2018A). *Discover the Chesapeake: Facts & Figures*. Web access: April 4, 2018. <https://www.chesapeakebay.net/discover/facts>
- Chesapeake Bay Program (2018B). *Learn the Issues: Oysters*. Web access: April 4, 2018. <https://www.chesapeakebay.net/issues/oysters>
- Chesapeake Bay Program Water Quality Database (1984-present). Available online: [http://www.chesapeakebay.net/what/downloads/cbp\\_water\\_quality\\_database\\_1984\\_present](http://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present) (accessed on 16 August 2017).
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298.
- Clean Water Act section 303(d): Notice for the establishment of the total maximum daily load (TMDL) for the Chesapeake Bay (EPA). Available online: <https://www.gpo.gov/fdsys/pkg/FR-2011-01-05/pdf/2010-33280.pdf> (accessed on 12 December 2017).
- Colwell, R. R., Kaper, J., & Joseph, S. W. (1977). *Vibrio cholerae*, *Vibrio parahaemolyticus*, and other vibrios: occurrence and distribution in Chesapeake Bay. *Science*, 198(4315), 394-396.

- Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, 240, 113-122.
- Crookston, N. L., Finley, A. O. (2007). yaImpute: An R Package for k-NN Imputation. *Journal of Statistical Software*. 23(10):1-16.
- Dall'Olmo, G., Gitelson, A. A., & Rundquist, D. C. (2003). Towards a unified approach for remote estimation of chlorophyll-a in both terrestrial vegetation and turbid productive waters. *Geophysical Research Letters*, 30(18).
- Davis, B. J., Jacobs, J. M., Davis, M. F., Schwab, K. J., DePaola, A., & Curriero, F. C. (2017). Environmental determinants of *Vibrio parahaemolyticus* in the Chesapeake Bay. *Applied and Environmental Microbiology*, 83(21), e01147-17.
- Davis, B. J., & Curriero, F. C. (2019). Development and Evaluation of Geostatistical Methods for Non-Euclidean-Based Spatial Covariance Matrices. *Mathematical Geosciences*, 1-25.
- Davis, B. J., Jacobs, J. M., Zaitchik, B., DePaola, A., & Curriero, F. C. (2019). *Vibrio parahaemolyticus* in the Chesapeake Bay: operational in situ prediction and forecast models can benefit from inclusion of lagged water quality measurements. *Applied and Environmental Microbiology*, 85(17), e01007-19.
- DeLuca, N., Zaitchik, B., & Curriero, F. (2018). Can multispectral information improve remotely sensed estimates of total suspended solids? A statistical study in Chesapeake Bay. *Remote Sensing*, 10(9), 1393.
- DePaola, A., Nordstrom, J. L., Bowers, J. C., Wells, J. G., & Cook, D. W. (2003). Seasonal abundance of total and pathogenic *Vibrio parahaemolyticus* in Alabama oysters. *Applied and Environmental Microbiology*, 69(3), 1521-1526.



- Dogliotti, A. I., Ruddick, K. G., Nechad, B., Doxaran, D., & Knaeps, E. (2015). A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters. *Remote Sensing of Environment*, 156, 157-168.
- Dorji, P., & Fearn, P. (2016). A quantitative comparison of total suspended sediment algorithms: A case study of the last decade for MODIS and Landsat-based sensors. *Remote Sensing*, 8(10), 810.
- Dou, X., & Yang, Y. (2018). Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. *Computers and Electronics in Agriculture*, 148, 95-106.
- Doxaran, D., Ehn, J., Bélanger, S., Matsuoka, A., Hooker, S., & Babin, M. (2012). Optical characterisation of suspended particles in the Mackenzie River plume (Canadian Arctic Ocean) and implications for ocean colour remote sensing. *Biogeosciences*, 9 (2012), 3213-3229.
- Feng, L., Hu, C., Chen, X., & Song, Q. (2014). Influence of the Three Gorges Dam on total suspended matters in the Yangtze Estuary and its adjacent coastal waters: Observations from MODIS. *Remote Sensing of Environment*, 140, 779-788.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1-67.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 1189-1232.
- Froelich, B. A., Phippen, B., Fowler, P., Noble, R. T., & Oliver, J. D. (2017). Differences in Abundances of Total *Vibrio* spp., *V. vulnificus*, and *V. parahaemolyticus* in Clams and Oysters in North Carolina. *Applied and Environmental Microbiology*, 83(2), e02265-16.

- Ghil, M., & Malanotte-Rizzoli, P. (1991). Data assimilation in meteorology and oceanography. *Advances in Geophysics*, 33, 141-266.
- Gitelson, A. A., Schalles, J. F., & Hladik, C. M. (2007). Remote chlorophyll-a retrieval in turbid, productive estuaries: Chesapeake Bay case study. *Remote Sensing of Environment*, 109(4), 464-472.
- Greenfield, D. I., Gooch Moore, J., Stewart, J. R., Hilborn, E. D., George, B. J., Li, Q., ... & Sandifer, P. A. (2017). Temporal and environmental factors driving *Vibrio vulnificus* and *V. parahaemolyticus* populations and their associations with harmful algal blooms in South Carolina detention ponds and receiving tidal creeks. *GeoHealth*, 1(9), 306-317.
- Han, B., Loisel, H., Vantrepotte, V., Mériaux, X., Bryère, P., Ouillon, S., ... & Zhu, J. (2016). Development of a semi-analytical algorithm for the retrieval of suspended particulate matter from remote sensing over clear to very turbid waters. *Remote Sensing*, 8(3), 211.
- Hannart, A., Carrassi, A., Bocquet, M., Ghil, M., Naveau, P., Pulido, M., ... & Tandeo, P. (2016). DADA: data assimilation for the detection and attribution of weather and climate-related events. *Climatic Change*, 136(2), 155-174.
- Hasan, M., & Benninger, L. (2017). Resiliency of the western Chesapeake Bay to total suspended solid concentrations following storms and accounting for land-cover. *Estuarine, Coastal and Shelf Science*, 191, 136-149.
- Hastie, T.; Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1, 297–310.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.
- Hu, C., Chen, Z., Clayton, T. D., Swarzenski, P., Brock, J. C., & Muller-Karger, F. E. (2004). Assessment of estuarine water-quality indicators using MODIS medium-resolution

- bands: Initial results from Tampa Bay, FL. *Remote Sensing of Environment*, 93(3), 423-441.
- Jacobs, J. M., Rhodes, M., Brown, C. W., Hood, R. R., Leight, A., Long, W., & Wood, R. (2014). Modeling and forecasting the distribution of *Vibrio vulnificus* in Chesapeake Bay. *Journal of Applied Microbiology*, 117(5), 1312-1327.
- Jacobs, J., Moore, S. K., Kunkel, K. E., & Sun, L. (2015). A framework for examining climate-driven changes to the seasonality and geographical range of coastal pathogens and harmful algae. *Climate Risk Management*, 8, 16-27.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). New York: springer.
- Johnson, C. N., Flowers, A. R., Noriega, N. F., Zimmerman, A. M., Bowers, J. C., DePaola, A., & Grimes, D. J. (2010). Relationships between environmental factors and pathogenic vibrios in the northern Gulf of Mexico. *Applied and Environmental Microbiology*, 76(21), 7076-7084.
- Johnson, C. N., Bowers, J. C., Griffitt, K. J., Molina, V., Clostio, R. W., Pei, S., ... & Hasan, N. A. (2012). Ecology of *Vibrio parahaemolyticus* and *Vibrio vulnificus* in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washington (United States). *Applied and Environmental Microbiology*, 78(20), 7249-7257.
- Johnson, C. N. (2015). Influence of Environmental Factors on *Vibrio* spp. in Coastal Ecosystems. *Microbiology Spectrum*, 3(3).
- Kaneko, T., & Colwell, R. R. (1973). Ecology of *Vibrio parahaemolyticus* in Chesapeake bay. *Journal of Bacteriology*, 113(1), 24-32.

- Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., & Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. *Environmental Modelling & Software*, 19(9), 845-855.
- Kemp, W. M., Boynton, W. R., Adolf, J. E., Boesch, D. F., Boicourt, W. C., Brush, G., ... & Harding, L. W. (2005). Eutrophication of Chesapeake Bay: historical trends and ecological interactions. *Marine Ecology Progress Series*, 303, 1-29.
- Klontz, K. C., Lieb, S., Schreiber, M., Janowski, H. T., Baldy, L. M., & Gunn, R. A. (1988). Syndromes of *Vibrio vulnificus* infections: clinical and epidemiologic features in Florida cases, 1981-1987. *Annals of Internal Medicine*, 109(4), 318-323.
- Konrad, S., Paduraru, P., Romero-Barrios, P., Henderson, S. B., & Galanis, E. (2017). Remote sensing measurements of sea surface temperature as an indicator of *Vibrio parahaemolyticus* in oyster meat and human illnesses. *Environmental Health*, 16(1), 92.
- Lê Cook, B., & Manning, W. G. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai Archives of Psychiatry*, 25(1), 55.
- Lee, K. Y., Cha, Y. T., & Park, J. H. (1992). Short-term load forecasting using an artificial neural network. *IEEE Transactions on Power Systems*, 7(1), 124-132.
- Li, B., Rodell, M., Kumar, S., Beaudoin, H. K., Getirana, A., Zaitchik, B. F., ... & Tian, S. (2018). Global GRACE data assimilation for groundwater and drought monitoring: Advances and challenges. *Water Resources Research*.
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), 145-151.

- Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *R Journal*, 6(1).
- MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8), 2248-2255.
- Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10), 1087-1096.
- Martinez-Urtaza, J., Bowers, J. C., Trinanes, J., & DePaola, A. (2010). Climate anomalies and the increasing risk of *Vibrio parahaemolyticus* and *Vibrio vulnificus* illnesses. *Food Research International*, 43(7), 1780-1790.
- Martinez-Urtaza, J., Baker-Austin, C., Jones, J. L., Newton, A. E., Gonzalez-Aviles, G. D., & DePaola, A. (2013). Spread of Pacific Northwest *Vibrio parahaemolyticus* strain. *New England Journal of Medicine*, 369(16), 1573-1574.
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun), 983-999.
- Meinshausen, N. (2017). Quantregforest: quantile regression forests. *R package version 1.3-7*.
- Mitchell, C., Hu, C., Bowler, B., Drapeau, D., & Balch, W. M. (2017). Estimating particulate inorganic carbon concentrations of the global ocean from ocean color measurements using a reflectance difference approach. *Journal of Geophysical Research: Oceans*, 122(11), 8707-8720.
- Morel, A., & Prieur, L. (1977). Analysis of variations in ocean color 1. *Limnology and Oceanography*, 22(4), 709-722.

- Muhling, B. A., Jacobs, J., Stock, C. A., Gaitan, C. F., & Saba, V. S. (2017). Projections of the future occurrence, distribution, and seasonality of three *Vibrio* species in the Chesapeake Bay under a high-emission climate change scenario. *GeoHealth*, 1(7), 278-296.
- Najjar, R., Patterson, L., & Graham, S. (2009). Climate simulations of major estuarine watersheds in the Mid-Atlantic region of the US. *Climatic Change*, 95(1-2), 139-168.
- NASA Goddard Space Flight Center, Ocean Biology Processing Group; (2018): Moderate Resolution Imaging Spectrometer (MODIS) Ocean Color Data, NASA OB.DAAC, Greenbelt, MD, USA. [http://doi.org/10.5067/ORBVIEW-2/SEAWIFS\\_OC.2014.0](http://doi.org/10.5067/ORBVIEW-2/SEAWIFS_OC.2014.0). Accessed: 2018/07/30. Maintained by NASA Ocean Biology Distributed Active Archive Center (OB.DAAC), Goddard Space Flight Center, Greenbelt MD.
- NASA OceanColor Web. Available online: <https://oceancolor.gsfc.nasa.gov> (accessed on 16 August 2017).
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.
- Newton, A., Kendall, M., Vugia, D. J., Henao, O. L., & Mahon, B. E. (2012). Increasing rates of vibriosis in the United States, 1996–2010: review of surveillance data from 2 systems. *Clinical Infectious Diseases*, 54(suppl\_5), S391-S395.
- Nilsson, W. B., Paranjpye, R. N., Hamel, O. S., Hard, C., & Strom, M. S. (2019). *Vibrio parahaemolyticus* risk assessment in the Pacific Northwest: it's not what's in the water. *FEMS Microbiology Ecology*, 95(4), fiz027.
- Nychka, D.; Furrer, R.; Paige, J.; Sain, S. (2015) “fields: Tools for spatial data.” *R package version 9.0*. doi:10.5065/D6W957CT.

- Ondrusek, M., Stengel, E., Kinkade, C. S., Vogel, R. L., Keegstra, P., Hunter, C., & Kim, C. (2012). The development of a new optical total suspended matter algorithm for the Chesapeake Bay. *Remote Sensing of Environment*, 119, 243-254.
- Pathiraja, S., Marshall, L., Sharma, A., & Moradkhani, H. (2016). Hydrologic modeling in dynamic catchments: A data assimilation approach. *Water Resources Research*, 52(5), 3350-3372.
- Patra, A. K., Acharya, B. C., & Mohapatra, A. (2009). Occurrence and distribution of bacterial indicators and pathogens in coastal waters of Orissa. *Indian Journal of Marine Sciences*, 38(4), 474–480.
- Phillips, A. M. B., DePaola, A., Bowers, J., Ladner, S., & Grimes, D. J. (2007). An evaluation of the use of remotely sensed parameters for prediction of incidence and risk associated with *Vibrio parahaemolyticus* in Gulf Coast oysters (*Crassostrea virginica*). *Journal of Food Protection*, 70(4), 879-884.
- Qiu, Z. (2013). A simple optical model to estimate suspended particulate matter in Yellow River Estuary. *Optics Express*, 21(23), 27891-27904.
- R Core Team. R: A Language and Environment for Statistical Computing. Available online: <http://www.R-project.org>
- Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., ... & Griffin, P. M. (2011). Foodborne illness acquired in the United States—major pathogens. *Emerging Infectious Diseases*, 17(1), 7.
- Schalles, J. F. (2006). Optical remote sensing techniques to estimate phytoplankton chlorophyll a concentrations in coastal. *Remote Sensing of Aquatic Coastal Ecosystem Processes* (pp. 27-79). Springer, Dordrecht.

- Shashaani, S., Guikema, S. D., Zhai, C., Pino, J. V., & Quiring, S. M. (2018). Multi-Stage Prediction for Zero-Inflated Hurricane Induced Power Outages. *IEEE Access*, 6, 62432-62449.
- Shen, F., Zhou, Y., Peng, X., & Chen, Y. (2014). Satellite multi-sensor mapping of suspended particulate matter in turbid estuarine and coastal ocean, China. *International Journal of Remote Sensing*, 35(11-12), 4173-4192.
- Shen, J., Qin, Q., Wang, Y., & Sisson, M. (2019). A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. *Ecological Modelling*, 398, 44-54.
- Smith, R. C., & Baker, K. S. (1978). Optical classification of natural waters 1. *Limnology and Oceanography*, 23(2), 260-267.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222.
- Sokoletsky, L., Fang, S., Yang, X., & Wei, X. (2016). Evaluation of empirical and semianalytical spectral reflectance models for surface suspended sediment concentration in the highly variable estuarine and coastal waters of East China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(11), 5182-5192.
- Stumpf, R. P. (1988). Sediment transport in Chesapeake Bay during floods: analysis using satellite and surface observations. *Journal of Coastal Research*, 1-15.
- Stumpf, R. P., & Pennock, J. R. (1989). Calibration of a general optical equation for remote sensing of suspended sediments in a moderately turbid estuary. *Journal of Geophysical Research: Oceans*, 94(C10), 14363-14371.



- Thickman, J. D., & Gobler, C. J. (2017). The ability of algal organic matter and surface runoff to promote the abundance of pathogenic and non-pathogenic strains of *Vibrio parahaemolyticus* in Long Island Sound, USA. *PloS one*, 12(10), e0185994.
- Turner, J. W., Malayil, L., Guadagnoli, D., Cole, D., & Lipp, E. K. (2014). Detection of *Vibrio parahaemolyticus*, *Vibrio vulnificus* and *Vibrio cholerae* with respect to seasonal fluctuations in temperature and plankton abundance. *Environmental Microbiology*, 16(4), 1019-1028.
- Tzortziou, M., Herman, J. R., Gallegos, C. L., Neale, P. J., Subramaniam, A., Harding Jr, L. W., & Ahmad, Z. (2006). Bio-optics of the Chesapeake Bay from measurements and radiative transfer closure. *Estuarine, Coastal and Shelf Science*, 68(1-2), 348-362.
- Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5), 910.
- Unger, D. A. (1985). A method to estimate the continuous ranked probability score. Preprints. In *Ninth Conf. on Probability and Statistics in Atmospheric Sciences* (pp. 206-213).
- Urquhart, E. A., Zaitchik, B. F., Hoffman, M. J., Guikema, S. D., & Geiger, E. F. (2012). Remotely sensed estimates of surface salinity in the Chesapeake Bay: A statistical approach. *Remote Sensing of Environment*, 123, 522-531.
- Urquhart, E. A., Hoffman, M. J., Murphy, R. R., & Zaitchik, B. F. (2013). Geospatial interpolation of MODIS-derived salinity and temperature in the Chesapeake Bay. *Remote Sensing of Environment*, 135, 167-177.

- Urquhart, E. A., Zaitchik, B. F., Guikema, S. D., Haley, B. J., Taviani, E., Chen, A., Brown, M. E., Huq, A., & Colwell, R. R. (2015). Use of Environmental Parameters to Model Pathogenic Vibrios in Chesapeake Bay. *Journal of Environmental Informatics*, 26:1–13.
- U.S. Centers for Disease Control and Prevention (2019). *Foodborne Diseases Active Surveillance Network (FoodNet)*. Web access: May 13, 2019. <https://wwwn.cdc.gov/foodnetfast/>
- U.S. Environmental Protection Agency. (1996). Recommended Guidelines for Sampling and Analyses in the Chesapeake Bay Monitoring Program. EPA, Washington, DC.
- U.S. Environmental Protection Agency. (2011). Clean Water Act section 303(d): notice for the establishment of the total maximum daily load (TMDL) for the Chesapeake Bay (EPA). EPA, Washington, DC.
- Uusitalo, L., Lehtikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, 63, 24-31.
- Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, 291, 55-64.
- Virginia Department of Health. (2013). *Virginia Department of Health issues emergency closing of Fisherman's Island to shellfish harvesting*. Web access: April 26, 2019. <http://www.vdh.virginia.gov/news/PressReleases/2013/071213Shellfish.htm>
- Vogel, R. L., & Brown, C. W. (2016). Assessing satellite sea surface salinity from ocean color radiometric measurements for coastal hydrodynamic model data assimilation. *Journal of Applied Remote Sensing*, 10(3), 036003.

- Wang, H., Hladik, C. M., Huang, W., Milla, K., Edmiston, L., Harwell, M. A., & Schalles, J. F. (2010). Detecting the spatial and temporal variability of chlorophyll-a concentration and total suspended solids in Apalachicola Bay, Florida using MODIS imagery. *International Journal of Remote Sensing*, 31(2), 439-453.
- Werdell, P. J., Franz, B. A., Bailey, S. W., Harding Jr, L. W., & Feldman, G. C. (2007, October). Approach for the long-term spatial and temporal evaluation of ocean color satellite data products in a coastal environment. In *Coastal Ocean Remote Sensing* (Vol. 6680, p. 66800G). International Society for Optics and Photonics.
- Werdell, P. J., Bailey, S. W., Franz, B. A., Harding Jr, L. W., Feldman, G. C., & McClain, C. R. (2009). Regional and seasonal variability of chlorophyll-a in Chesapeake Bay as observed by SeaWiFS and MODIS-Aqua. *Remote Sensing of Environment*, 113(6), 1319-1330.
- Werdell, P. J., Franz, B. A., & Bailey, S. W. (2010). Evaluation of shortwave infrared atmospheric correction for ocean color remote sensing of Chesapeake Bay. *Remote Sensing of Environment*, 114(10), 2238-2247.
- Zhao, H., Chen, Q., Walker, N. D., Zheng, Q., & MacIntyre, H. L. (2011). A study of sediment transport in a shallow estuary using MODIS imagery and particle tracking simulation. *International journal of remote sensing*, 32(21), 6653-6671.
- Zimmerman, A. M., DePaola, A., Bowers, J. C., Krantz, J. A., Nordstrom, J. L., Johnson, C. N., & Grimes, D. J. (2007). Variability of total and pathogenic *Vibrio parahaemolyticus* densities in northern Gulf of Mexico water and oysters. *Applied and Environmental Microbiology*, 73(23), 7589-7596.

## CIRRICULUM VITAE

**Nicole M. DeLuca**

Born February 28, 1992 in Denville, NJ

### EDUCATION

- 2014 – Present    *Ph.D. Candidate, Earth & Planetary Sciences, Johns Hopkins University*  
                          **Thesis Advisor:** Benjamin F. Zaitchik  
                          **Research:** Applications of marine remote sensing for water quality monitoring and predictive modeling of pathogenic bacteria using machine learning and statistical analysis; Chesapeake Bay, USA and Lake Tana, Ethiopia
- 2014 – 2016    *M.A, Earth & Planetary Sciences, Johns Hopkins University*  
                          **Advisor:** Naomi E. Levin  
                          **Research:** Isotopic variation in Ethiopian precipitation; Identifying changes in Andean hydrology and climate using triple oxygen isotopes and clumped isotope laboratory analyses
- 2011 – 2014    *B.S., Chemistry, York College of Pennsylvania*  
                          Minor in Sustainability and Environmental Science  
                          **Advisors:** Gregory P. Foy and Keith E. Peterman  
                          **Research:** Exploring the effects of tree type on absorption of drinking water pollutants in a reservoir; Increasing climate change literacy among undergraduates and local communities
- 2010 – 2011    Studied Meteorology and Marine Science, **North Carolina State University**

### FIELD EXPERIENCE

- 2018                    24-hour cruise aboard R/V Atlantic Explorer
- 2017                    Water quality sampling in Lake Tana, Ethiopia
- 2016 – 2017           Water quality sampling and spectroradiometer measurements in Chesapeake Bay, USA
- 2016                    Water sampling for colored dissolved organic matter and dissolved inorganic carbon in Everglades National Park, FL, USA

### TEACHING EXPERIENCE

- 2019                    *Teaching Assistant*, Johns Hopkins University, “Climate Change: Science and Policy” course for undergraduates
- 2018 – 2019           *Teaching Assistant*, Johns Hopkins University at BIOS Bermuda, Field Experience in Oceanography study abroad course for undergraduates
- 2015                    *Co-instructor*, Johns Hopkins University, Engineering Innovations Fundamentals course for Baltimore area high school students

- 2013 *Student mentor, American Chemical Society, Student delegates to the United Nations Framework Convention on Climate Change 19<sup>th</sup> Conference of Parties in Warsaw, Poland*
- 2012 *Chemistry laboratory assistant, York College of Pennsylvania*

### PROFESSIONAL EXPERIENCE

- 2016 – 2019 *Journal Club Coordinator, Johns Hopkins University, Department of Earth & Planetary Sciences*
- 2015 – 2019 *Graduate Research Assistant, Johns Hopkins University, Department of Earth & Planetary Sciences*
- 2013 – 2014 *Secretary, American Chemical Society Chapter at York College of Pennsylvania*
- 2013 – 2014 *Student Delegate, American Chemical Society at the United Nations Framework Convention on Climate Change 18<sup>th</sup> Conference of Parties in Doha, Qatar and 19<sup>th</sup> Conference of Parties in Warsaw, Poland*
- 2013 *National Science Foundation Research Experience for Undergraduates Intern, University of California Irvine, Irvine, CA*
- 2010 – 2011 *Student Representative, American Meteorological Society Chapter at North Carolina State University*

### VOLUNTEER EXPERIENCE

- 2019 – 2019 *Therapy Dog Handler, Pets on Wheels*
- 2018 – 2019 *Equine Program Volunteer, Wellspring of Life Farm Equine and Canine Therapy*
- 2016 – 2017 *Public Programs Volunteer, Maryland Science Center*

### AWARDS AND FELLOWSHIPS

- 2014 – 2017 *National Science Foundation Integrative Graduate Education & Research Traineeship - Water, Climate, and Health Fellow, Johns Hopkins University*
- 2017 *Woods Hole Oceanographic Institute's Ocean Carbon and Biogeochemistry Program support for Cornell Satellite Remote Sensing Training course*
- 2016 *Johns Hopkins University Department of Earth & Planetary Sciences Journal Club Award*
- 2014 *York College of Pennsylvania Senior Honor Society*
- 2013 *American Geophysical Union Fall Meeting Travel Grant*
- 2013 *National Science Foundation-Research Experience for Undergraduates Internship at University of California Irvine, Department of Earth System Science*

## PROFESSIONAL DEVELOPMENT

- 2017                      Cornell Satellite Remote Sensing Training course  
2017                      Presentation Skills for Scientists and Engineers course at Johns Hopkins University

## REFEREED PUBLICATIONS

- (In review) DeLuca NM, Zaitchik BF, Guikema SD, Jacobs JM, Davis BJK, Curriero FC. Evaluation of remotely sensed *Vibrio parahaemolyticus* prediction and forecast models for Chesapeake Bay. *Remote Sensing of Environment*.
- DeLuca NM, Zaitchik BF, Curriero FC. 2018. Can multispectral information improve remotely sensed estimates of total suspended solids? A statistical study in Chesapeake Bay. *Remote Sensing*, 10, 1393.
- DeLuca NM. "Science of the Anthropocene." *Climate Change Literacy and Education: The Science and Perspectives from the Global Stage Volume 1*. Ed. Peterman KE, Ed. Foy GP, Ed. Cortes MR. 2017. Washington, DC: Oxford, 2017. 49-65. Print.
- Prather MJ, Hsu J, **DeLuca NM**, Jackman CH, Oman LD, Douglass AR, Fleming EL, Strahan SE, Steenrod SD, Søvde OA, Isaksen ISA, Froidevaux L, Funke B. 2015. Measuring and modeling the lifetime of nitrous oxide including its variability. *Journal of Geophysical Research: Atmospheres*, 120, doi:10.1002/2015JD023267.

## PRESENTATIONS (\*oral presentation)

- \*DeLuca NM, Using remote sensing for the prediction and forecasting of *Vibrio parahaemolyticus* in the Chesapeake Bay. *NASA Goddard Applied Sciences Seminar*, Laurel, MD, USA June 2019. (Invited)
- DeLuca NM et al. Evaluation of remotely sensed *Vibrio parahaemolyticus* prediction and forecast models for Chesapeake Bay. *International Ocean Colour Science Meeting*, Busan, Republic of Korea, April 2019.
- DeLuca NM, et al. Can multispectral information improve remotely sensed estimates of total suspended solids? A statistical study in the Chesapeake Bay. *American Geophysical Union Fall Meeting*, Washington, DC, USA December 2018.
- \*DeLuca NM, et al. Can multispectral information improve remotely sensed estimates of total suspended solids? A statistical study in the Chesapeake Bay. *Chesapeake Research & Modeling Symposium*, Linthicum, MD, USA June 2018.
- \*DeLuca NM, et al. Trends and variability of water quality in Lake Tana, Ethiopia using MODIS-Aqua. *American Geophysical Union Annual Meeting*, New Orleans, LA, USA December 2017.
- DeLuca NM, et al. Spatial and temporal variation in the isotopic composition of Ethiopian rainfall. *American Geophysical Union Annual Meeting*, San Francisco, CA, USA December 2016.

\*DeLuca NM, et al. American Chemical Society student delegates' outreach at the United Nations Framework Convention on Climate Change. *American Chemical Society Fall National Meeting*, San Francisco, CA, USA, August 2014.

DeLuca NM, et al. An improved atmospheric lifetime of nitrous oxide based on measurement and constrained models. *American Geophysical Union Fall Meeting*, San Francisco, CA, USA, December 2013.

\*DeLuca NM, Peterman KE. Climate change: The physical science and public policy disconnect. *Uptown York Rotary Club Meeting*, York, PA, USA, February 2013.

\*DeLuca NM, et al. Saving the planet by increasing undergraduate climate science literacy. *Downtown York Rotary Club Meeting*, York, PA, USA, February 2013.

## WORKSHOPS

NASA Interagency Workshop on Societal Applications of Satellite Data for the Chesapeake Bay. Laurel, MD, USA, August 2018.

## JOURNAL REVIEWER

Remote Sensing of Environment  
Remote Sensing

## PROFESSIONAL MEMBERSHIPS

American Geophysical Union  
Earth Science Woman's Network

## COMPUTATION AND VISUALIZATION SKILLS

- ☐ Machine Learning, Data Analytics, Statistical Analysis, Predictive Modeling
- ☐ Programming experience in R (advanced), MATLAB, Python, Linux
- ☐ ArcMap GIS, QGIS
- ☐ SeaDAS, BEAM ocean color remote sensing processing software
- ☐ Microsoft Office suite
- ☐ Adobe Illustrator