5-31-2016

Identification and Use of Indicator Data to Develop Models for Marine-Sourced Risks in Massachusetts Bay

Marin M. Kress
University of Massachusetts Boston

IDENTIFICATION AND USE OF INDICATOR DATA TO DEVELOP MODELS FOR MARINE-SOURCED RISKS IN MASSACHUSETTS BAY

A Dissertation Presented

by

MARIN M. KRESS

Submitted to the Office of Graduate Studies,
University of Massachusetts Boston,
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2016

The Intercampus Marine Science Graduate Program

IDENTIFICATION AND USE OF INDICATOR DATA TO DEVELOP MODELS FOR MARINE-SOURCED RISKS IN MASSACHUSETTS BAY

A Dissertation Presented

by

MARIN M. KRESS

Approved as to style and content by:

_____
Robert Bowen, Associate Professor
Chairperson of Committee

_____
Jarrett Byrnes, Assistant Professor
Member

_____
Helen Poynton, Assistant Professor
Member

_____
Todd Swannack, PhD., U.S. Army Corps of Engineers
Member

_____
Ellen Douglas, Campus Coordinator
Intercampus Marine Science Graduate Program

_____
Robyn Hannigan, Dean
School for the Environment

ABSTRACT

IDENTIFICATION AND USE OF INDICATOR DATA TO DEVELOP MODELS

FOR MARINE-SOURCED RISKS IN MASSACHUSETTS BAY

May 2016

Marin M. Kress, B.A., Smith College
M.S., University of Massachusetts Boston
Ph.D., University of Massachusetts Boston

Directed by Professor Robert Bowen

The coastal watersheds around Massachusetts Bay are home to millions of people, many of whom recreate in coastal waters and consume locally harvested shellfish. Epidemiological data on food-borne illness and illnesses associated with recreational water exposure are known to be incomplete. Of major food categories, seafood has the highest recorded rate of associated foodborne illness. In total, the health impacts from these marine-sourced risks are estimated to cost millions of dollars each year in medical expenses or lost productivity. When recorded epidemiological data is incomplete it may be possible to estimate abundance or prevalence of specific pathogens or toxins in the source environment, but such environmental health challenges require an interdisciplinary approach.

This dissertation is divided into four sections: (1) a presentation of two frameworks for organizing research and responses to environmental health issues; (2) an exploration of human population dynamics in Massachusetts Bay coastal watersheds from 2000 to 2010 followed by a review of, and identification of potential indicators for, five marine-sourced risks: *Enterococcus* bacteria, *Vibrio parahaemolyticus* bacteria, Hepatitis A Virus, potentially toxigenic *Pseudo-nitzschia* genus diatoms, and anthropogenic antibiotics; (3) an introduction to environmental health research in the context of a changing data landscape, presentation of a generalized workflow for such research with a description of data sources relevant to marine environmental health for Massachusetts Bay; and (4) generation of models for the presence/absence of *Enterococcus* bacteria and *Pseudo-nitzschia delicatissima* complex diatoms and model selection using an information-theoretic approach.

This dissertation produced estimates of coastal watershed demographics and usage levels for anthropogenic antibiotics, it also demonstrated that *Pseudo-nitzschia delicatissima* complex diatoms may be present in any season of the year. Of the modeling generation and selection, the *Enterococcus* model performed poorly overall, but the *Pseudo-nitzschia delicatissima* complex model performed adequately, demonstrating high sensitivity with a low rate of false negatives. This dissertation concludes that monitoring data collected for other purposes can be used to estimate marine-sourced risks in Massachusetts Bay, and such work would be improved by data from purpose-designed studies.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance, support, thoughtful discussions, and intellectual sparkle provided by friends and family over the years. The list below is an incomplete record of all those who have my sincere thanks. First and foremost to my parents and family for support. Also first is my 'academic parent' and advisor Dr. Bob Bowen for everything since we met at a UMB Open House, and of course to my amazing committee members Dr. Helen Poynton, Dr. Jarrett Byrnes, and Dr. Todd Swannack for their guidance and feedback.

The inter-dependent nature of scientific work has never been more apparent to me than during this process, thanks to specialists in multiple fields who shared their time and expertise with me. Special thanks to Wendy Leo of the MWRA for data that was essential to this work, to Dave Borkman of the University of Rhode Island for information about phytoplankton monitoring in Massachusetts Bay, Mingshun Jiang for discussions about physical processes in Massachusetts Bay and cogent answers to my out-of-the-blue email questions, and J. Duff for.

Thanks to the wonderful faculty and staff in the UMB EEOS Dept/SfE, and the Urban Harbors Institute, you are inspirational in the way you make interdisciplinary work happen. Thanks to all my classmates for being great people to learn with, especially K. Cialino, S. Sheldon, S. Gazda, C. Macintyre, B. Broadaway, Y. Yang, E. Sun, S. Rouhani, R. Barakin, B. Fradkin, S. Kichefski, D. Oglavie for their good humor -

especially during long hours in the GIS lab. Special thanks to Tom Angus for years of collegial support, inspiration, and many many conversations about *Pseudo-nitzschia*.

For friends who receive me with good humor despite my faults, the Library of Congress would be too small to catalog everything I should thank you for, this is for you Ali, Anne, Adrienne, Kay, Grace, Jenn, and Michelle for being my core support and inspiration. Thanks to A. Jones for biology conversations and measured thought; to JW for infinite patience and kind assistance; to B. Lapointe and family for friendship and support; to the inspirational Jurneys and Alaina, Britta, and Hardy – simply the best teachers and friends underwater and on land – I would not be on this path without you. Thanks to my Knauss class – such talented people who continue to remind me how awesome science can be. Special thanks to E. Wilkinson, T. Davenport, M. Freedman, T. Dolan, and L. Windecker for extra support in the final stages of this process, food is love.

Thanks to others for being models of accomplishment and offering unwavering support for my academic endeavors; especially J. & L. Lillycrop, J. Rosati, K. Touzinsky, E. Vuxton, and colleagues at USACE; to Dr. Cami Graham, Dr. Margaret Koziel, and Dr. Qi He for foundational scientific and professional mentorship; my boundless gratitude to the profound Mr. Meta for 1000% awesomeness and being a sounding board for all the thoughts in my head. Lastly, thanks to the inspirational professors at Smith, but especially R. Lim, R. Dorit, J. Wopereis, M. Anderson, and C. White-Ziegler who expected the best from me and taught me so much. Thank you, thank you, thank you, for being such wonderful friends, colleagues, and sources of wisdom.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER 1

INTRODUCTION

Background.

The concepts of 'health' and 'well-being' are intertwined but not identical. We often think of 'health' as physical and physiologic health, as something which can be diminished or damaged during a disease or after an injury.[1] The concept of well-being is not as clearly defined as that of health, but can include multiple aspects of a person's quality of life, including economic vitality, social and cultural connectedness, psychological stability and strength, and happiness.[1] Although health and well-being may be defined separately, the two concepts are so connected that the World Health Organization (WHO) defines health as a "state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity."[2] Reduced health may lead to reduced well-being, and vice-versa, but this is not always the case. People may have impaired health but a satisfactory level of well-being if there is enough social and cultural support.[3] Identifying and improving well-being is an active area of research in the medical and public health communities.[3-5] Research linking health and well-being to the environment, recognizing the value of natural systems in supporting health and well-being, is also ongoing.[1; 6-9]

Like well-being, a person's health can be improved or adversely affected by the surrounding environment. For those living in coastal areas the marine environment is part of their local environment and the potential influences are much more apparent.[1] Those living far from the coast may still directly interact with ocean products such as seafood, or be indirectly linked to the ocean through weather or climate impacts. Human interaction with the ocean and its products can improve health and well-being through scenic enjoyment[8], recreation opportunities[9], spiritual and cultural practices[10], and healthy seafood consumption.[11] Conversely, we recognize that ocean interactions may contain risks such as consumption of contaminated seafood, physical trauma from large waves or strong currents, interactions with poisonous animals such as jellyfish, and storm or flood damage to coastal communities.[1] In addition to visible risks such as flooding, microbiological risks have been recognized from multiple taxa, including bacteria, viruses, toxigenic dinoflagellates or diatoms, helminthes, and yeasts.[12] While this dissertation will focus on marine-sourced risks and their potentially negative impacts on human health the benefits of ocean interaction should not be forgotten.

Today we recognize a wide variety of risks to human health. Some of these risks are man-made (e.g., anthropogenic pollutants, cultural norms), some are intrinsic parts of the natural environment (e.g., hurricanes, tornadoes), and some are a combination of the two; risks that exist independent of human activity, but which can be made more serious by human behaviors. These risks, at the intersection of socio-economic factors and environmental conditions, are challenging to study using a traditional single-discipline approach. Interdisciplinary questions require interdisciplinary approaches. For

environmental health questions an interdisciplinary approach requires understanding relevant aspects of the human population at the risk and the biology and ecology of the risks themselves. As new risks are identified policy-makers must consider the level of resources that can, or should, be devoted to minimizing exposure to these risks. Exposure to certain risks may be embedded in cherished cultural practices or economically significant industries, or both.

Although we recognize a wide variety of health risks, we are not equally informed about each risk's prevalence or impact. For certain types of illness traditional epidemiological data are not accurate in capturing the true number of disease cases. Infectious disease completeness reporting ranges from 9 to 99 percent, with greatest completeness for high profile diseases like tuberculosis, AIDS, and certain sexually transmitted diseases. [13] For environmentally-linked illnesses, such as seafood-borne illness and illness associated with recreational waters, public health experts estimate that as few as ten percent of cases are reported.[14-18] For diseases with incomplete epidemiological data we must find other ways to estimate the true burden of disease until reporting rates improve. One alternate approach is to understand the abundance of the underlying risk factor as it exists in the environment. Examples of this may include identifying viral strains circulating among a population, measuring the abundance of indigenous marine bacteria in areas where humans harvest shellfish, or identifying seasonal variation in the presence of toxigenic phytoplankton in coastal waters.[19-22] For marine-sourced risks understanding the underlying risk factor means estimating their presence in the wild (e.g., natural abundance), or estimating the extent of human loading

of a risk into the marine environment (e.g., enteric bacteria and viruses released through wastewater flows). Additional data on the extent of human exposure (e.g., the number of swimmers exposed to polluted water) would further help to refine disease burden estimates, but such data may not be available.

This dissertation is motivated by the following question:

How can investigation of multiple marine-sourced risks best be organized in terms of the identification of useful indicators? When epidemiological data is lacking, can we identify and use proxy data from other sources to understand potential risks, and can we use that data to develop predictive models that could serve to protect public health?

Using examples for specific marine-sourced risks known to exist in Massachusetts Bay, and the proximate human population living in the surrounding coastal watersheds, this dissertation is divided into four chapters that treat different aspects of the motivating question.

Chapter 1 Frameworks.

The research question for Chapter 1 is: *How can we organize our understanding of the pathways that create risks to human health?*

Identifying a problem is only the first step in solving it, long-term success depends on addressing the root cause, which for environmental health risks may be an intertwined combination of environmental and human factors. We focus on coastal

systems because they are home to large numbers of people and the coastal environment is influenced by natural variability, episodic events, and anthropogenic forcing. Chapter 1 is built around two essential themes, 1) the recognition that coastal systems are complex, but we can reveal the underlying structure and use that understanding to make informed management choices, and 2) management choices that are both socially inclusionary and data-supported are likely to be successfully implemented. Chapter 1 describes how using a comprehensive, yet flexible, organizing framework is a useful was to address environmental health problems. The two frameworks discussed are:

- Driver-Pressure-State-Impact-Response (DPSIR) framework
- Driver-Pressure-State-Exposure-Effect-Action (DPSEEA) framework

Both of these frameworks allow the user to place a specific problem within a larger system context to identify where to target response actions. Chapter 1 examines the elements of the DPSIR and DPSEEA frameworks, and applications of this approach in research and policy settings.

To illustrate the wide applicability of this framework, chapter 1 summarizes 11 case studies which utilize the DPSIR framework to evaluate different types of environmental, health, and management challenges around the globe. Those 11 cases discuss the following topics:

- Evaluating success under the European Water Framework Directive
- Common environmental challenges of coastal megacities
- Urban infrastructure development and groundwater use and quality

- Historical development drivers in South African municipalities
- Coastal management in three South American coastal sites
- Environmental challenges facing the ecosystem of the Ebrié Lagoon, Ivory Coast
- Recommending indicators to understand reef fishing in Kenya
- Evaluating sustainable aquaculture options in South Africa
- Linking upstream drivers and downstream impacts in Venetian bathing beaches
- Understanding declines in coastal wetlands in Xiamen, China
- Integrating indicators to assess Marine Protected Areas in Malta

In addition to the 11 DPSIR examples, Chapter 1 presents two human health focused applications of the DPSEEA framework. The two examples are:

- The GEO Health Pilot study in São Paulo, Brazil that brought together stakeholders from the medical, waste management, environmental, and residential communities to identify waste and water problems adversely affecting human health.
- The Good Places, Better Health program in Scotland, initiated to make better connections between the built environment and human health and well-being.

These examples show that the DPSIR and DPSEEA frameworks are flexible in their application, allowing users to organize and share their thinking about complex social and environmental issues. In addition, these frameworks allows for establishing measurement criteria to evaluate response actions before those actions are taken. This facilitates

transparency for all stakeholders involved, an important aspect of problem solving where environmental and health issues may be intimately linked to social norms.

Chapter 2 Human Population Demographics and Marine-sourced Risks in Massachusetts Bay.

The topics addressed in Chapter 2 are: *What are the demographics of people living in the six watersheds that border Massachusetts Bay? For these watersheds, what was composition of the population with the greatest opportunity for coastal water interaction in 2010? For a set of 5 marine-sourced risks known to exist in Massachusetts Bay what are the known or suspected environmental and socio-economic influences on their abundances identified in the existing scientific literature? Which of these influencing factors should be, or could be, monitored through direct or proxy indicators to provide the most valuable public health value information about the changes in risk potential in nearshore coastal waters frequented by residents and tourists?*

This chapter examines data from multiple sources to make demographic estimates about the population living in the coastal watersheds around Massachusetts Bay, the same population expected to have to highest level of interaction with coastal recreational waters and locally harvested seafood. Marine-sourced risks may take multiple forms, five commonly identified categories include enteric bacteria, indigenous marine bacteria, enteric viruses, natural marine toxigenic organisms, and anthropogenic pollutants. This chapter reviews the available epidemiological and biological data for five marine-sourced risks to human health that exist in the Massachusetts Bay area representative of a

different category, the specific risks are *Enterococcus* bacteria [23], *Vibrio parahaemolyticus* bacteria[24], Hepatitis A Virus [25], *Pseudo-nitzschia* genus diatoms[26], and anthropogenic antibiotics[27]). All of these risks may exist at varying scales and abundance across the Bay, but water quality testing is driven primarily by enteric bacteria, primarily *Enterococcus*. Through this review we create a matrix identifying high-value data types for influencing factors on the five marine-sourced risks that could be useful in developing predictive models.

Chapter 3 Interdisciplinary Data Science.

The topic addressed in Chapter 3 is: *How inter-disciplinary questions in environmental health and infectious disease research may be addressed through the use of data beyond traditional medical and epidemiological sources.*

This chapter discusses the changing landscape of data availability and the emerging practice of data science. Increases in computing power have allowed for both the rise of 'big data' and the generation of crowdsourced data. Crowdsourced data in particular may offer unique insights on certain topics, but is not necessarily informative for every research question. In addition, this chapter presents a general 3-phase workflow for the type of interdisciplinary environmental health work that utilizes multiple disparate data types. This chapter also presents a list of the data collected from public sources to support marine-source risk modeling in Massachusetts Bay.

Chapter 4 Marine-sourced Risk Models.

The research question in this chapter is:

Is it possible to use existing public data to build a model that can predict the presence or absence of Pseudo-nitzschia delicatissima *complex diatoms and* Enterococcus *bacteria in Massachusetts Bay with reasonable accuracy? Are there data gaps that limit the predictive ability of these models? Does there appear to be any correlation between the presence of these taxa in the northern part of Massachusetts Bay?*

This chapter discusses the information-theoretic approach used to develop and select predictive models for the presence/absence of two marine-sourced risks known to exist in Massachusetts Bay. The first model is for *Pseudo-nitzschia delicatissima* complex diatoms as measured at two stations in Massachusetts Bay between the years 1995 and 2014. The second model is for *Enterococcus* bacteria as measured at three marine beaches along the north coast of the Bay during the summer bathing seasons of 2007 to 2014. After presenting the results of the model testing and the potential utility of these models, this chapter closes with suggestions for ways to improve these and other marine-sourced risk modeling efforts.

Literature Cited.

1. Wheeler, B., White, M. P., Fleming, L. E., Taylor, T., Harvey, A., Depledge, M. H. 2014. Influences of the Oceans on Human Health and Well-Being. *In* Oceans and Human Health: Implications for Society and Well-being. Bowen R. E., Depledge, M. H., Carlarne, C. P. *et al*, Eds.: 3-22. John Wiley & Sons, Ltd. Oxford, England.

2. World Health Organization. 1946. Preamble to the Constitution of the World Health Organization. Vol. 2: 100.

3. Kiefer, R. A. 2008. An integrative review of the concept of well-being. Holist. Nurs. Pract. 22: 244.

4. Kahneman, D., Krueger, A. B. 2006. Developments in the Measurement of Subjective Well-Being. J. Econ. Perspectives. 20: pp. 3-24.

5. Vemuri, A. W., Costanza, R. 2006. The role of human, social, built, and natural capital in explaining life satisfaction at the country level: Toward a National Well-Being Index (NWI). Ecol. Econ. 58: 119-133.

6. Millennium Ecosystem Assessment, Reid, W. V., Mooney, H. A., Cropper, A., Capistrano, D., Carpenter, S. R., Chopra, K., Dasgupta, P., Dietz, T., Duraiappah, A. K., Hassan, R., Kasperson, R., Leemans, R., May, R. M., McMichael, T., Pingali, P., Samper, C., Scholes, R., Watson, R. T., Zakri, A. H., Shidong, Z., Ash, N. J., Bennett, E., Kumar, P., Lee, M. J., Raudsepp-Hearne, C., Simons, H., Thonell, J., Zurek, M. B. 2005. Ecosystems and Human Well-being: Synthesis. Island Press. Washington, D.C.

7. Coon, J. T., Boddy, K., Stein, K., Whear, R., Barton, J., Depledge, M. H. 2011. Does participating in physical activity in outdoor natural environments have a greater effect on physical and mental wellbeing than physical activity indoors? A systematic review. Environ. Sci. Technol. 45: 1761-1772.

8. White, M., Smith, A., Humphryes, K., Pahl, S., Snelling, D., Depledge, M. 2010. Blue space: The importance of water for preference, affect, and restorativeness ratings of natural and built scenes. J. Environ. Psychol. 30: 482-493.

9. Depledge, M. H., Bird, W. J. 2009. The Blue Gym: Health and Wellbeing from our Coasts. Mar. Pollut. Bull. 58: 947-948.

10. Costanza, R., d'Arge, R., De Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R., Paruelo, J. 1997. The value of the world's ecosystem services and natural capital. Nature. 387: 253-260.

11. Mozaffarian, D., Rimm, E. B. 2006. Fish intake, contaminants, and human health - Evaluating the risks and the benefits. JAMA. 296: 1885-1899.

12. Bienfang, P. K., DeFelice, S. V., Laws, E. A., Brand, L. E., Bidigare, R. R., Christensen, S., Trapido-Rosenthal, H., Hemscheidt, T. K., McGillicuddy Jr, D. J., Anderson, D. M. 2011. Prominent Human Health Impacts from Several Marine Microbes: History, Ecology, and Public Health Implications. Int. J. Microbiol. 2011.

13. Doyle, T. J., Glynn, M. K., Groseclose, S. L. 2002. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. Am. J. Epidemiol. 155: 866-874.

14. Centers for Disease Control and Prevention. 2015. National Enteric Disease Surveillance: COVIS Annual Summary, 2013. Centers for Disease Control and Prevention. Atlanta, GA. 1-13.

15. Hlavsa, M. C., Roberts, V. A., Kahler, A. M., Hilborn, E. D., Mecher, T. R., Beach, M. J., Wade, T. J., Yoder, J. S. 2015. Outbreaks of illness associated with recreational water—United States, 2011–2012. MMWR. 64: 668-672.

16. Centers for Disease Control and Prevention. 2013. Surveillance for foodborne disease outbreaks--United States, 2009-2010. MMWR. 62: 41-47.

17. Centers for Disease Control and Prevention. 2012. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet Surveillance Report for 2011 (Final Report). U.S. Department of Health and Human Services, CDC. Atlanta, GA. 1-53.

18. Shuval, H. 2003. Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment. J. Water & Health. 1: 53-64.

19. McNeil Jr., D. G. 2015. Oysters may Serve as Link in Transmission of Norovirus. Section D: HealthNYTThe New York Times Company. New York, NY.

20. Williams, W. W., Lu, P. J., O'Halloran, A., Bridges, C. B., Kim, D. K., Pilishvili, T., Hales, C. M., Markowitz, L. E., Centers for Disease Control and Prevention (CDC). 2015. Vaccination coverage among adults, excluding influenza vaccination - United States, 2013. MMWR. 64: 95-102.

21. Kirkpatrick, B., Pierce, R., Cheng, Y. S., Henry, M. S., Blum, P., Osborn, S., Nierenberg, K., Pederson, B. A., Fleming, L. E., Reich, A. 2010. Inland transport of aerosolized Florida red tide toxins. Harmful algae. 9: 186-189.

22. Johnson, C. N., Bowers, J. C., Griffitt, K. J., Molina, V., Clostio, R. W., Pei, S., Laws, E., Paranjpye, R. N., Strom, M. S., Chen, A., Hasan, N. A., Huq, A., Noriea, N. F.,3rd, Grimes, D. J., Colwell, R. R. 2012. Ecology of *Vibrio parahaemolyticus* and *Vibrio vulnificus* in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washington (United States). Appl. Environ. Microbiol. 78: 7249-7257.

23. Commonwealth of Massachusetts, Department of Public Health. 2015. Marine and Freshwater Beach Testing in Massachusetts, Annual Report: 2014 Season. Commonwealth of Massachusetts. Boston, M.A. 1-151.

24. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Vibrio Control Plan. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/vibrio.html (Accessed October 19, 2015).

25. Centers for Disease Control and Prevention. 2015. Surveillance for Viral Hepatitis – United States, 2013. U.S. Department of Health and Human Services. Atlanta, G.A. http://www.cdc.gov/hepatitis/statistics/2013surveillance/commentary.htm#hepatitisA (Accessed September 26, 2015).

26. Libby, P. S., Borkman, D. G., Geyer, W. R., Turner, J. T., Costa, A. S. 2014. 2013 Water Column Monitoring Results. Report 2014-17. Massachusetts Water Resources Authority. Boston, MA. 1-43.

27. Hicks, L. A., Taylor Jr, T. H., Hunkler, R. J. 2013. U.S. outpatient antibiotic prescribing, 2010. N. Engl. J. Med. 368: 1461-1462.

CHAPTER 2

INTEGRATING FRAMEWORKS TO ASSESS HUMAN HEALTH AND WELL-BEING IN MARINE ENVIRONMENTAL SYSTEMS

Note: This paper has been published as the following citation: R. E. Bowen, M. Kress, G. Morris, and D. Rothman. 2014. Chapter 2: Integrating Frameworks to Assess Human Health and Well-Being in Marine Environmental Systems. In *Oceans and Human Health: Implications for Society and Well-Being. R.E. Bowen, M.H. Depledge, C.P. Carlarne, and L.E. Fleming, eds. Hoboken, NJ: John Wiley & Sons, Ltd. 304 pages. ISBN: 978-1-119-94131-6.*

Introduction.

The previous chapter characterized the differences and interconnections between human health and well-being. One of the earliest conscious connections made by humans was that their health and well-being were influenced by nature. We framed our lives in the context of the environment in which we lived, and still do so today. When it did not rain, food was scarce; seeking shelter from storms helped initiate social systems; rivers and the coastal ocean provided swifter movement and opened the opportunity for connections between distant communities of people. We discovered that eating certain foods during certain times of the year might make us ill, while other plants held curative properties. As the populations of humans grew larger, so did our understanding of the

diversity of ways in which we were connected to the natural systems around us. As our social systems became more sophisticated so did our capacity to define and respond to environmental change.

However, that knowledge did not always lead to sophisticated, beneficial social action. Indeed, many would argue that our social actions were both too infrequent and too ineffective. It is not within the purview of this chapter to assess the origin, reasoning and consequences of historic social choices. Rather, the focus here is to examine the value of integrating frameworks that afford a more mature, inclusive view of complex relationships between environmental conditions, human health, and well-being.

Integration, Complexity and Need for New Frameworks.

The fact that a system is complex does not mean it lacks a structure that one can reveal and act upon. Indeed, it has been nearly fifty years since Herbert Simon wrote his famous paper describing the "architecture of complexity,"[1] and those insights are as valuable today in considering coastal systems as they were then. Simon has said of complex systems that "in the face of complexity, an in-principle reductionist may be at the same time a pragmatic holist."[1] It is simply pragmatic to embrace the idea that information on all parts of this system need to be acquired if an understanding of the whole is to be achieved.

Coastal environments are systems that interact in non-simple ways but nonetheless, hold an underlying structure that can be better understood. And, this structure can be used to direct and integrate efforts to acquire, assess, and communicate information linking the environment, human health and well-being. To achieve these

outcomes, we need to broaden the field of investigation to ensure that information on communities, on the structure of society, on coastal environmental change, on the social gains and losses influenced by the environment, and the responsive activity driven by that knowledge, are all included in our pragmatic approach to managing the whole of coastal systems.

Background.

During the past several years, numerous broad-ranging national and international reports have assessed the state of marine and coastal systems with the goal of contributing to more integrative and sustainable views. The overall goals motivating these reports were quite broad and included, *inter alia*: assessing climate change;[2-4] illustrating global ecological themes;[5] developing a strategy for the sustained monitoring of global environmental change;[6-8] conducting national assessments of coastal and ocean management;[9] and using indicators to assess change in coastal systems.[10; 11] The last decadal assessment provided by the United Nations Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection (GESAMP)[12] states:

> Humanity's future, just like its past, will continue to depend on the oceans, on the intricate interchanges between land and water. Yet the relationship has changed. Over most of human history it has been dominated by the sea's influence on people. But from now on humanity's effect on the state of the sea is probably at least as important. And, by and large, this is getting worse.

> The state of the world's seas and oceans is deteriorating. Most of the problems identified decades ago have not been resolved, and many are worsening. New threats keep emerging. The traditional uses of the seas and coasts – and benefits that humanity gets from them – have been widely undermined.
>
> . . .
>
> More hopefully, perhaps, there is a dawning realization that neither individual problems, nor the crisis of the seas as a whole can be dealt with in isolation. They are intricately interlinked both with themselves and with social and economic development on land. Policy decisions, research, and management programs are all shifting their focus accordingly.

The GESAMP report is quoted here not because of its unique conclusions, but rather because it provides a notable example of the kinds of conclusions that reside in virtually all the significant broad-view reports assessing environmental systems released since the turn of the new millennium. Aware of the challenges of oversimplification, we argue that two themes, in particular, emerge and provide essential organizing tools for the study of the oceans, human health and well-being.

First, it is clear that **effective and efficient environmental management needs to embrace an integrated, ecosystem perspective that includes humans**. Traditional sector management (including the public health sector) ignores the "reality of interdependence" faced by current managers. These realities are clear and well reflect our natural and social scientific understanding of environmental systems. However, to fully embrace this system view brings with it a level of complexity and uncertainty for which we are, too often, ill-prepared. Consequently, the second essential theme is that **management should move toward a more inclusive, information-driven system of decision-making and assessment.** With these themes clearly in mind, we can now begin this chapter's discussion of analytical frameworks.

Integrating Frameworks for Human Health and Well-Being.

A starting point to reduce the barriers of these complex challenges resides in the acceptance and use of simple integrating frameworks designed to ensure that all the forces that contribute to a functional understanding are accounted for and considered. This chapter describes and illustrates two such frameworks. They are the:

DPSIR – Driver, Pressure, State, Impact, Response; and,

DPSEEA – Driver, Pressure, State, Exposure, Effect, Action.

These frameworks emerged and evolved concurrently, both during the early 1990's. Within the present context, the DPSIR is used to assess the broader issues of environmentally influenced human well-being, while the DPSEEA is used most generally by the public health community. These two frameworks focus the remainder of this

discussion for several reasons. First, the measure of conceptual similarity between the two is high. This reflects the substantial symmetry in the attributes and indicators where human health and well-being converge (e.g., those commonalities influencing critical areas of environmental change). The two frameworks diverge where the major consequence of that environmental change is primarily in assessing only the more illness related view of human health (DPSEEA) or is viewed more broadly within the context of well-being (DPSIR).

Assessing the Influences on Environmental Change.

An essential step in understanding of the architecture of complexity is to identify the sources of environmental change. Figure 2-1 (below) represents the general forcing functions influencing the state of environmental conditions. Natural variability, episodic events, and anthropogenic forcing all play a role in the dynamics of coastal environmental systems. Therefore, one simple goal for a successful framework is for it to be able to better discern the relative contribution of the various drivers and pressures altering the state of environmental conditions. Environmental conditions can influence a change in human health and in overall well-being – as well as the dynamics between them. This view is used to convey the importance in understanding that both human and natural factors can be the primary sources of environmental state changes. And, since both human health and well-being can be influenced by those changes in environmental state, our capacity to responsibly act or respond is dependent on an understanding the associated drivers and pressures.

Effective arguments can be made that both the DPSIR and the DPSEEA meet the framework needs of the management community and have acquired broad and general support.[7; 8; 11; 13-18] We appear to have reached a point of general consensus on the attributes necessary for successful management framework even if marginal details may differ slightly from effort to effort.

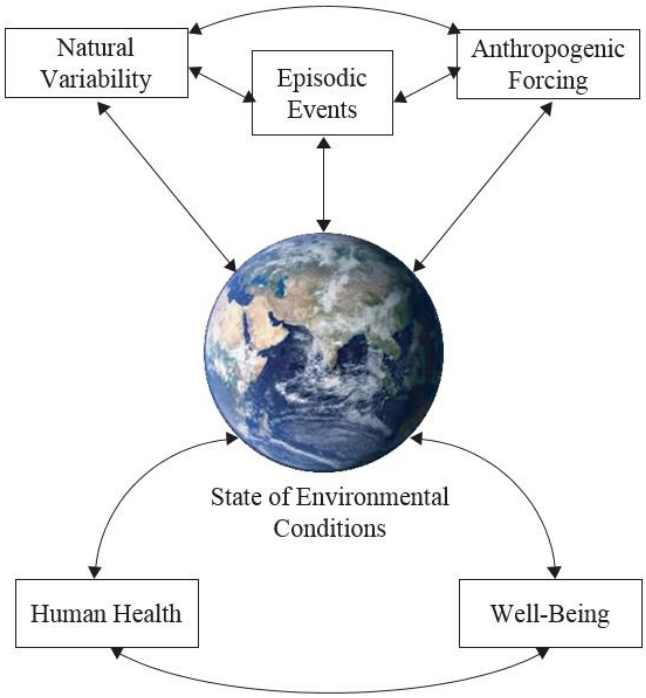

Figure 2-1. Forcing functions, global environmental systems, human health and well-being. The forcing functions influencing change on the state of environmental system (including coastal and marine systems) natural variability, episodic events, and anthropogenic activity. Therefore, one simple goal for a successful framework is for it to be able to better discern the relative contribution of the various drivers and pressures altering the state of environmental conditions. Environmental conditions can influence a change in human health and in overall well-being – as well as the dynamics between them. And, since both human health and well-being can be influenced by those changes in environmental state, our capacity to responsibly act or respond is dependent on an understanding the associated drivers and pressures.

The Driver-Pressure-State-Impact-Response Framework.

The current core of what is known as the DPSIR can most easily be traced to work in the early 1990s carried out by the Organization for Economic Cooperation and Development (OECD) when it focused on developing a more common, structured view of how to assess the relationship between humans and the environment.[19-22] This holistic view was embraced and expanded by the United Nations and the European Commission (among others) to include a broader view of the root causes of environmental change and the impacts this change has on ecosystems and on humans.[23-25] The DPSIR framework, as described here, was first elaborated by the European Environmental Agency in 1995.[26; 27]

The integrative view conveyed by this framework is that:

> "The way a country or community is broadly structured and organized is defined by a suite of large-scale social drivers which can impose various forms of pressure altering the state of environmental conditions. The changing state of the environment can consequently impact social benefit values (notably social well-being). Responsible social sustainability requires that any responses to enhance sustainability and overall well-being account for all attributes of this system."[26; 27]

Figure 2-2 (below) is a simple diagram of the DPSIR, wherein:

• ***Driver*** refers to large scale socio-economic conditions and sectoral trends such as patterns in coastal land use and land cover, and population growth, economic growth and energy use patterns;

• ***Pressure*** include patterns of development-driven habitat alteration, the introduction of industrial POPs/metals and fertilizer use, wastewater management can affect environmental quality;

• ***State*** indicators describe observable changes in environmental conditions. If assessed over time *state* indicators afford a view of view of environmental system change;

• ***Impacts*** are the discrete measured changes in social benefit values and in ecosystem service values. In short, within the present context the focus of *impacts* is on attributes of human/social well-being; and,

• ***Response*** indicators are described as the institutional response to changes within the whole of this system.

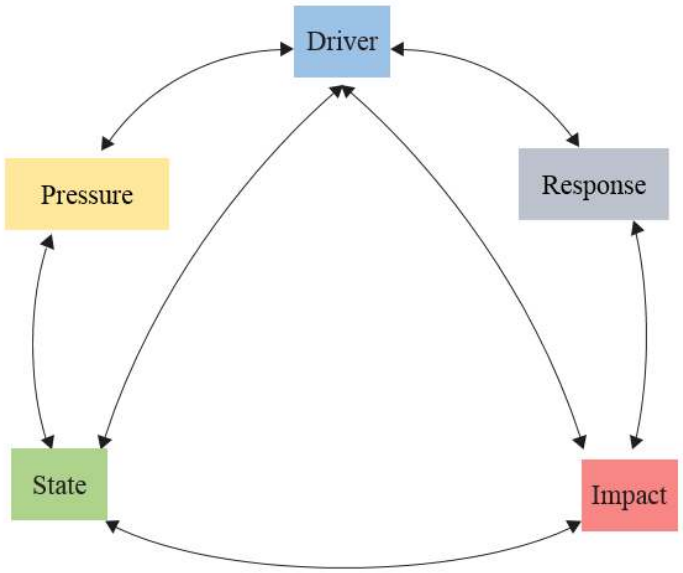

Figure 2-2. The DPSIR framework: Driver– Pressure– State– Impact– Response. The DPSIR represents a structured view of the relationship linking large-scale social organization (drivers) and the con- sequential pressures society can impose on the state of the environment. In the current context impacts are viewed as the associate changes to human health and well-being. Response represents the nature of management action based on this social-environmental system.

The primary value of the DPSIR framework resides in how it serves to ensure that scientific assessment, policy development, and regulatory construction incorporate environmental changes as well as the social benefits that are linked to that change. In the context of coastal ecosystem functions, Kerry Turner and colleagues[13] have argued that the DPSIR is useful for:

> the scoping of biodiversity management issues and problems. It can make tractable the complexity of causes of habitat/species degradation or loss and the links to socio-economic activities, across the relevant spatial and temporal scales. It also provides the important conceptual connection

> between ecosystem change and the effects of that change (impacts) on people's economic and social well-being. Relevant indicators of environmental change can be derived and the loss of ecosystem function provision in terms of goods and services (direct and indirectly received) can be translated into human welfare loss and quantified in monetary and/or other more qualitative ways.[13]

The reader is reminded that the context of this volume is a focus on human health and well-being. Accordingly, our assessments of *impact* are, by definition, associated with human/social well-being within the constructs of the DPSIR. However, as we have already noted, if human health served to singularly focus the development of integrated policy formulation, we acknowledge that the DPSEEA framework is viewed as being the more effective construct.

The Driver-Pressure-State-Exposure-Effect-Action Framework.

The Driver-Pressure-State-Exposure-Effect-Action (DPSEEA) framework was developed by the World Health Organization, along with the United Nations Environment Programme and the United States Environmental Protection Agency, as part of the Health and Environmental Analysis for Decision-making Linkage Analysis and Monitoring Project, or HEADLAMP.[23; 28] Despite huge advances in understanding how health is created and destroyed (including the interdependence between human and environmental health), governments still encounter difficulty in developing coherent, evidence-informed, and effective policies on environment and human health.

In general, both the DPSIR and DPSEEA can be viewed as concurrent recognitions by different stakeholder communities of the need for tools to better integrate views of multidimensional systems. The DPSEEA evolved to meet the more specific need of the global public health community, while the DPSIR was viewed as meeting the needs of broader communities with interests in changing environmental conditions. Both frameworks hold in common an acceptance of the influential value of a changing state of the environment. They diverge in terms of the areas of emphasis they associate with environmental system change.[23; 28-31]

The convergent/divergent attributes of the DPSIR/DPSEEA relationship are illustrated in both Figures 2-3 and 2-4. The basic structure of the DPSEEA is represented in Figure 2-3. Here, the differences and similarities between the two frameworks are emphasized by the use of visual cues. We chose to emphasize these synergies by constructing views of them wherein their commonalities are clearly evident. Both frameworks share a view that large-scale social *drivers*, and consequential *pressures* can alter environmental conditions. If one compares Figures 2-2 (The DPSIR) and 2-3 (The DPSEEA) the boxes depicting *Driver*, *Pressure* and *State* are located in the same place in both figures and delineated in the same colors (*Driver*/Blue; *Pressure*/Yellow; *State*/Green). This visual commonality should reinforce the idea that the motivating forces behind, and stakeholder communities served by, the two frameworks are linked. They hold not only important similarities, but essential differences as well.

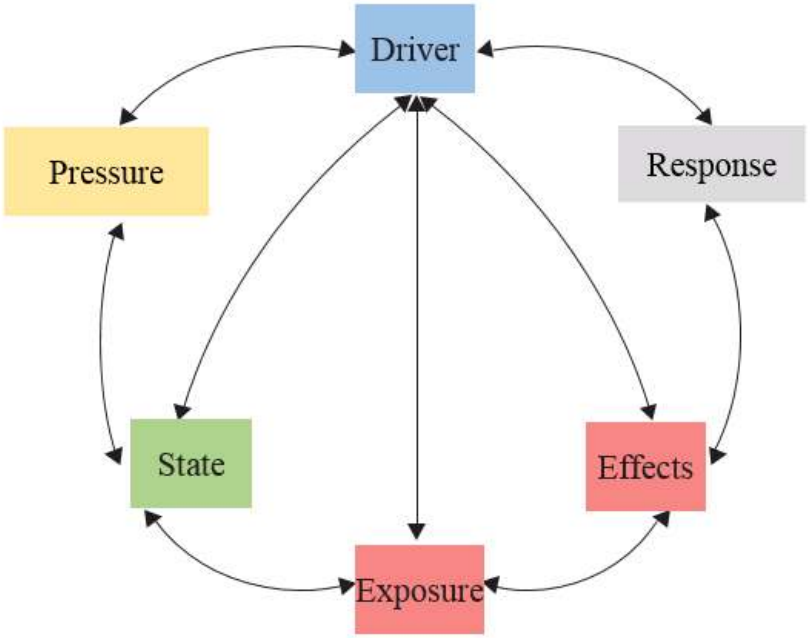

Figure 2-3. The DPSEEA framework: Driver – Pressure – State – Exposure – Effect – Action. The DPSEEA was proposed to represent the same intellectual approach to structuring complex systems as the DPSIR. The difference between the two frameworks resides in the social goals they engage. The DPSEEA is structured to address human health risks and thereby engages terms used by the medical and public health communities.

As already noted above, Figure 2-2 defines *Impacts* and *Response* as relating to a broader sweep of human/social impact of environmental change and integrative and contemplative response by various levels of governance to system change. In Figure 2-3 the three framework attributes defining *Exposure*, *Effect,* and *Action* serve as the divergent attributes of the framework. Here, with a determinative focus on the medical and public health communities:

• ***Exposure*** reflects the vectors of risk exposure (either risk elevation of diminution) that emerge as a consequence of environmental change;

• ***Effect*** is a measure of changes in health resulting from changes in risk exposure; and,

• ***Action*** assesses the nature and scope of regulatory, clinical or personal response changes in human health conditions.

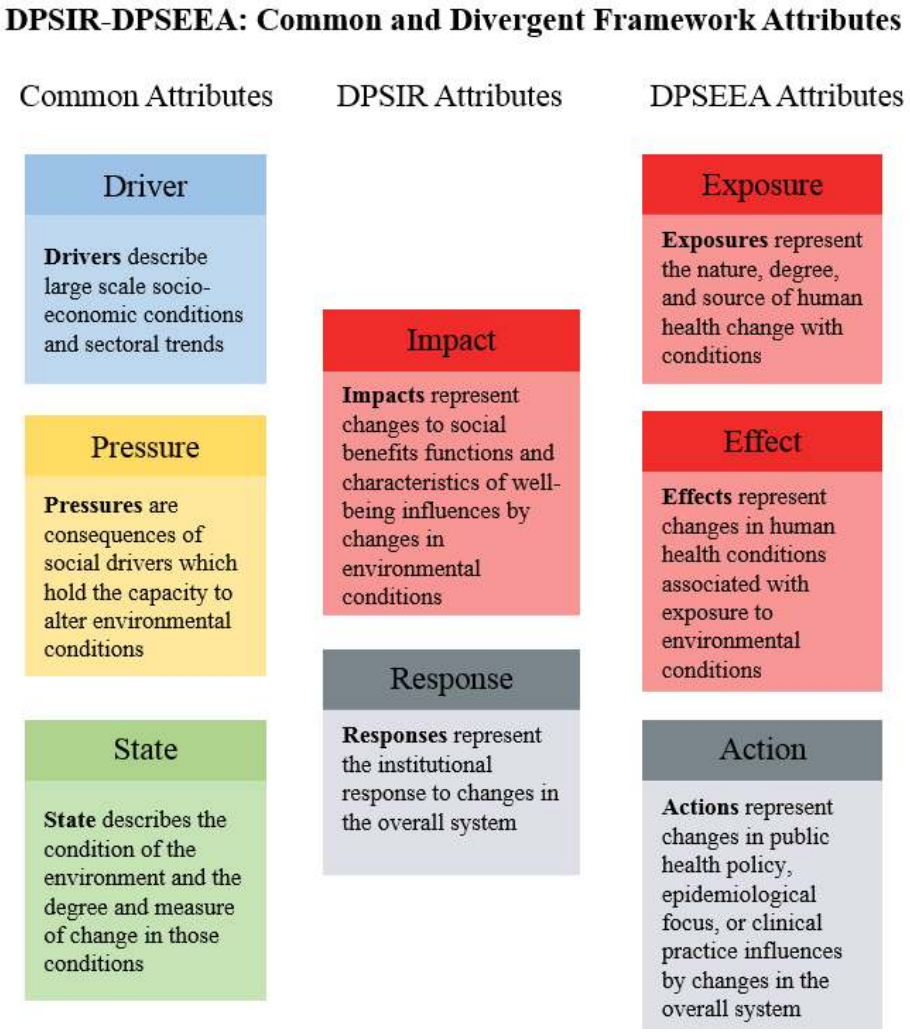

Figure 2-4. Similarities and differences between the DPSIR/DPSEEA frameworks. As noted, the two frameworks are similar in their approach to organizing complex system – as well as supportive information – however, they do contain important convergent/divergent attributes. This figure is an attempt to illustrate the relationship between the two frameworks

Together these two frameworks present the capacity to reveal the structure of the complex social/environmental themes that define the overall storyline and discrete components of this book. However, they provide more than just conceptual tools to be used in the abstract. They have been used often during particularly the last decade by researchers and policy-makers to assess these kinds of problems and to evaluate the best ways to respond. We have reviewed a large portion of this rich literature and selected the following case studies as representative of the diverse questions that have drawn benefit from the use of these frameworks.

The DPSIR in Case-study Literature.

The DPSIR has been, and continues to be, utilized in hundreds of studies around the world. The geographic scale of use varies broadly as does, not surprisingly, the broad range of the questions to which it has been applied. Here, we have selected and briefly summarized eleven case studies which we view as representative.

The European Water Framework Directive. Borja et al. (2006) used the DPSIR framework to forecast whether or not a waterbody would be likely to 'fail' in achieving 'good ecological status' under the European Water Framework Directive (WFD) by the year 2015.[32] A case in the Basque Country in northern Spain was selected to illustrate *pressures* and *impacts* (but not *responses* to *impacts*) on water quality at the regional level.[32] The identified coastal waterbodies were assigned a risk status of: Significant, Not Significant, Low (L), Moderate (M), High (H), or Without [Risk]. The relevant *pressures* and *state* changes were defined in terms of nutrients, water pollution, sediment pollution, water abstraction, dredged sediments, shoreline reinforcement, intertidal losses,

berths, alien species, and "global pressure." The waterbodies were then assigned an overall risk assessment rating (Low, Medium or High).

The authors concluded "that the main 'driving factors' explaining the variability of pressures upon Basque water bodies are…population density and industry concentration. … Use of the DPSIR analysis in the Basque Country, together with the methodologies in identifying relevant pressures and impacts, has been demonstrated as a useful approach in assessing the risk of failing the WFD objectives."[32]

Assessing Common Challenges of Coastal Megacities. The DPSIR was used to illustrate general environmental challenges common to the "megacity" (or, by inference, large urban areas).[33] The study gives greater weight to coastal megacities through the inclusion of a limited number of environmental considerations specific to coastal areas. The overall goal of this work was to examine the opportunities and challenges of sustainability in areas of highly clustered human habitat. "Sustainable development in coastal megacities faces various obstacles . . . which makes their planning and regulation actions extremely difficult."[33] These obstacles include: the influence that maritime transport emissions can have on air quality; loss of coastal and marine habitats; coastline stability; coastal erosion; and, sea level rise.

One attribute considered by the authors as common to all "megacities" and to most urban areas in the emerging economies is the informal expansion around them. A notable contribution of this work is the articulated need for the "establishment of unique set of indicators, in order to make monitoring of environmental state and impacts in

megacities clearer, therefore finding a way to more appropriate management responses easier."[33]

While the call for standardized indicators is not new, it is possible that megacities face unique challenges because of their sheer size and relatively recent appearance on the urban scene. The authors also provided a table giving examples of urban management responses to various sectoral challenges, captured in the *response* section of the study. For example, under the *response* options for addressing "Water Quality and Quantity" issues, potential governance actions included: changes in water pricing; enacting 'polluters pay' rules; improved wastewater treatment; better detection of leaking water conduits; promoting new technologies for saving water; and reuse of storm water and wastewater. These are challenging governance decisions that the authors argued benefit from range of social and environmental indicators the DPSIR is designed to reveal.

Urban Infrastructure Development and Groundwater Access. This study examined groundwater use and quality changes in relation to urban development in 7 major Asian cities (Bangkok, Jakarta, Manila, Osaka, Seoul, Taipei and Tokyo).[34] The DPSIR framework was used to link problems of groundwater quality and quantity to extractive activities, loss of aquifer recharge, possible saltwater intrusion, and subsidence deemed sufficient to be of risk to the surface built environment. Jago-on et al. (2009) documented that in some cities, the creation of laws restricting or regulating groundwater use (when enforced with effect) has helped reduce or stop subsidence rates.[34] Interestingly, in some places (Tokyo) implementation of such rules has revealed unanticipated feedback responses. Implementation of groundwater recharge activities can

affect infrastructure that was built during 'drawdown' times when the water table was lower.

Although published in 2009, the authors presciently anticipated the probability and scope of severe flooding that affected Bangkok during 2011. Bangkok is a city located at current sea level and had been flooded before. The previous notable Bangkok flood events were listed as occurring in 1983, 1995, and 1996, all caused substantial economic loss. It was also noted that flooding causes waste and garbage to be disbursed, creating conditions that can lead to an increase in human disease and an increase in disease vectors such as mosquitoes.[34] An effective use of the DPSIR in this study resides in the linkage of both groundwater QUANTITY and QUALITY with the same social drivers. The probability of tracking back both these often unrelated indicators can reveal governance options with greater policy and economic efficiencies.

The Historical Context of Development Drivers in South African Municipalities. This study argued that a temporal perspective is necessary when examining the development trajectories of two neighboring and environmentally similar, but socio-economically divergent South African municipalities, Ndlambe and Ngqushwa.[35] While describing the past and present situations of both municipalities, the authors focused on the differences in *drivers* as being key to building a compatible sustainable system for both communities.

This focus on historical differences in large-scale social system themes (*drivers*) appears to be somewhat novel within the literature, but appropriate given similarities and

differences between the jurisdictions. Palmer et al. (2011) used the DPSIR to examine land use changes, the nature of economic investment, and land-tenure as critical drivers in both the socio-economic and environmental factors that have clustered urban development in sensitive estuary areas. *Drivers* were broken out into sub-categories of 'Economic', 'Social,' and 'Legislative.' All three of these interlink/overlap most clearly in the issue of land tenure/ownership as a major driver of formal development (or lack thereof) and subsequent human migration. The authors made recommendations for government actions to support town planning and managerial capacities at the local level in order to enhance local plans and implement existing national/provincial guidelines.

A special concern was raised regarding the potential for 'ribbon development' along the coastline in the two municipalities examined. This type of development, where construction is concentrated in a narrow band immediately adjacent to the coast, has occurred in other parts of South Africa to the detriment of coastal systems. The authors concluded by cautioning that, "it is important to remember that development in the coastal zone is inevitable and instead of attempting to conserve the entire coastal zone, conservationists need to work together with town planners and developers to ensure that development pressure is controlled."[35]

Using the DPSIR Framework and Numerical Modeling to Examine Coastal Management in Three Contrasting South American Coastal Sites. The authors of this study compared three very different coastal zones and the management challenges facing each. First, the Santos estuary in Brazil, a sub-tropical area with the largest port and industrial complex on the Brazilian coast and an estimated population of 1 million

people, characterized as highly modified and polluted.[36] Second, the Bahía Blanca estuary in Argentina, located in a semi-arid climate and home to an important deep water port and petrochemical industry. With a population of 350,000 people its marine environment is characterized as modified and polluted.[36] The third site was the Aysén Fjord in southern Chile, a complex sub-polar estuarine system where the main economic activities are salmon aquaculture and artisanal fishing. The Aysén Fjord area was classified as near pristine and unpolluted.[36]

Despite the variety of environmental and socio-economic characteristics, all three sites face problems of habitat transformation, sewage and garbage disposal to varying degrees. Campuzano et al. (2011) argued that the goal of engaging and empowering local stakeholders in order to facilitate full implementation of existing environmental laws and procedures is currently under-practiced. They also posit that if governance could be more inclusive the effective *responses* to these common problems could acquire a higher level of sustainability.[36]

Using the DPSIR to Study the Largest Coastal Estuary Ecosystem in Western Africa – the Ebrié Lagoon, Ivory Coast. The focus of this piece is the Ebrié lagoon - the largest coastal estuarine ecosystem in Western Africa.[37] Using the DPSIR framework to structure related factors allowed the authors to draw "generic but reliable conclusions on the basis of limited data."[37] The paper divided the study area into 7 analytical units. The major environmental challenge in the lagoon is eutrophication derived from a variety of primary sources. Identified *impacts* from eutrophication include

fish kills, bad smells, floating debris, and an increase in waterborne diseases (typhoid, salmonella, cholera) because of high temperatures and noxious conditions in the lagoon.

While all study areas had similar upstream inputs – they share a general catchment zone - localized impacts and circulation differences may require different governance responses. For example, the city of Abidjan (a major port and economic hub) is the source of most industrial outputs and a large amount of domestic wastewater, but is also near the Vridi canal leading to the ocean and so is more strongly influenced by seawater and some tidal flushing than other parts of the lagoon. The authors noted that even though much data are missing, they can still make recommendations that could work to counter the effects of rapid unplanned urbanization and increased agricultural development and fertilizer use.[37]

Recommending Indicators to Understand Reef Fishing in Kenya Using the DPSIR Framework. The authors examined reef fishing activities in Kenya, where "the level of compliance to most…fisheries regulations by fishers has been low due to increased poverty, poor enforcement, and in some cases the rules are unknown and unclear."[38] They used the DPSIR framework to describe the selection of indicators "based on their relevance and priority for fisheries assessment and management."[38] While artisanal reef fishing supports 5000-6500 fishers (with each having an average 7 dependents), "marine fisheries comprise less than 5% of the national fisheries production, dwarfed by catch from inland lakes (predominantly Lake Victoria) and rivers. As a result, despite declaring some fishing gears illegal for many years, enforcement has been irregular, as the government has played little part in active management."[38] The

identification of a *response* option as simply enforcement of existing rules is an important point, since "reasonable legislative framework for fisheries management clearly exists" already in Kenya.[38]

It appears that the functional value of the DPSIR in this case was to assess impacts of environmental conditions on relatively marginal economic groups (marine fishers), and to confirm that implementation and enforcement of existing regulatory tools could be sufficient to mitigate existing challenges. The study also was able to emphasize both the importance of international tourism in Kenya's coastal zone and to identify tourism as a potential indicator needing greater attention from the government if reef fishing is to be sustainably retained as a viable economic contributor.[38]

Using the DPSIR Framework to Evaluate Aquaculture Options in South Africa. With an increasing amount of the world's seafood being produced through aquaculture[39], there is a growing interest in sustainable production methods. In this case study[40], the authors used a modified DPSIR framework to compared land-based systems focused on single-species aquaculture (abalone) to multi-species aquaculture (abalone + seaweed). They identified *pressures* from this aquaculture operation as *nutrient loads in aquaculture effluents* (released into the open ocean)*, harvesting of wild kelp,* and *greenhouse gas (GHG) emissions*. Indicators measured for different aspects of the framework included: nitrogen, phosphorus, oxygen, pH, temperature and turbidity of effluents, GHG emissions from electricity consumption under various scenarios, the hectares of kelp harvested per year, the investment costs to implement multi-species aquaculture, and changes in profit under different scenarios.

The authors found that switching from single-species to multi-species aquaculture would have clear economic, environmental, and societal benefits (in the form of increased profits, reduced effluents, and increased employment respectively). They noted that even without considering the environmental and societal benefits, multi-species aquaculture would be more profitable than single-species. The integrated and broadly inclusive orientation of the DPSIR helped to identify a more sustainable aquaculture system that better guaranteed a higher level of social well-being (in this case, economic valuation and job security).[40]

Linking Upstream Drivers and Downstream Impacts in Venetian Bathing Beaches. Venice draws visitors to its historic centers and coastal beaches.[41] Tourism is, essentially, the only viable economic activity supporting the Venice region. While cultural tourism dominates, the regional tourism of the Adriatic near to the City remains an important part of the economic landscape. With this motivation, researchers at the Veneto Regional Prevention and Protection Agency (ARPAV) performed a historical analysis of marine bathing waters for a 7-year timespan (2000-2006). Their analysis used the DPSIR framework to structure the relationship "considering water quality status and existing pressure sources."[41]

Recognizing that continuing development in the area has contributed to an increasing wastewater burden, the authors examined levels of specific bacteria in wastewater treatment plant (WWTP) effluents, rivers, offshore marine sites, and bathing waters. Levels of microbial contamination were identified as being linked to WWTPs that discharged into rivers or canals that then emptied into the Adriatic near bathing

areas. The best bathing-water quality was in an area without a river mouth and where WWTP effluent was released through an offshore submarine outfall pipe 4 km from land. Absent moving recreational bathing beaches to areas more distant from riverine influence, the researchers concluded that "submarine outfalls seem to be the best solution to guarantee good bathing water quality on the coast" and that "the issue of microbiological impacts must be studied following a river basin approach according to the influences of river loads on coastal areas."[41] Here, too, the system perspective of DPSIR helped to ensure that a broad and inclusive attribute set were included in the analysis.

Understanding Declines in Coastal Wetlands in Xiamen, China. Xiamen City on the southeastern coast of China has roughly 230km of coastline and is one of many areas around the world facing an apparent 'conflict between economic development and wetland conservation.'[42] Using the DPSIR framework to assess coastal wetland changes, the authors identified 4 time periods for comparative analysis of individual indicators. Because of specific concerns with coastal wetlands, they divided the *State* category into 3 sub-categories of *Physical State, Chemical State*, and *Biological State.* A total of 33 indicators were measured, examples included human population (*driver),* coastal reclamation area and industrial water use (*pressures*), suspended solids, organic pollutants, and species abundance (*states*), number of red tides and siltation in navigation channels (*impacts*), followed by indicators such as wastewater emission control, the establishment of conservation areas, and scientific support ability (*responses*).[42]

The authors concluded that, "On the whole, the state of the Xiamen coastal wetland is getting worse and the negative impacts are becoming more severe,"[42] despite the fact that "great human efforts have been expended to protect the coastal wetland."[42] These efforts have not been strong enough to counter the "pressures from human population growth and economic development"[42] that have driven the observed declines in wetland habitats. This study is particularly notable in the detail afforded the indicator structure and the complex architecture of this system.

Integrating Indicators to Assess Marine Protected Areas: A Malta Case. This study presented "a method for selecting and prioritizing socio-economic indicators, using a bottom-up approach involving stakeholder input. This technique [was] developed further to measure the effectiveness of integrated coastal management, using a Marine Protected Area (MPA) as an example. Stakeholder input is essential at an early stage to ensure MPA management success, providing the opportunity to include public participation and ensure community support."[43]

In this work, the DPSIR framework was used "to integrate environmental and socioeconomic indicators derived through stakeholder participation and contributing to the evaluation of management effectiveness."[43] The methodology employed by the organizers of this process was described as being able to identify "the socio-economic indicators that measure the success of MPA management in attaining goals that are important to the maximum number of stakeholder groups."[43]

One notable contribution of this study was the effort to reveal stakeholder preferences for both management goals and assessment indictors through the stakeholder influenced management plan developed to establish the MPA. Using a qualitative content analysis the plan was deconstructed to identify management goals and assessment indicators at the core of the plan recommendations. The emphasis on socio-economic indicator ranking is also an unusual contribution to the literature.

The DPSEEA in the Case-study Literature.

While the DPSIR has acquired broad international acceptance the DPSEEA does not hold the breadth of use in the literature; however, in cases where it has been engaged it has been used to strong effect. The relative under-use of the DPSEEA relative to the DPSIR is due in some significant part to the simple fact that the number of sectors to which the DPSIR can be applied exceeds the more focused human health core of the DPSEEA. Accordingly, we have selected two cases to represent the application of the DPSEEA; one from Brazil and the other from Scotland.

GEO Health Pilot Study, São Paulo, Brazil. São Paulo, Brazil - a metropolitan region of approximately 11 million people across 96 Administrative Districts – faces considerable challenges in the areas of water supply, sewage collection, and waste disposal.[44] Recognizing that these complex problems overlap to influence human health, a pilot project was undertaken by a broad range of local, national and international organizations, including: the São Paulo Municipal Health Secretariat (SMS), the city's Green and Environment Secretariat (SVMA), United Nations Environment Programme (UNEP), and the National School of Public Health (ENSP) from the Oswaldo Cruz

Foundation (FIOCRUZ) of the Ministry of Health of Brazil. The program also partnered with other stakeholders (including the academic community, medical doctors, and representatives of the environmental community). The inclusion of such a broad cohort of knowledgeable professionals clearly contributed to both the structural diversity at the core of the effort but to an enhanced probability of implementation success as well.

The goal was to identify specific water and waste problems in the city of São Paulo adversely affecting human health, and to build the indicators and indices that best depict the environment-health relationship of interest. This was accomplished by applying the core themes of the DPSEEA framework. Examples of important indicators identified for this situation included: *Share of heads of households without schooling per Administrative District* (driver); *Share of households without sewage system* (state); *Index of rodent infestation in buildings per administrative district* (exposure); *Average rate of hospitalization per waterborne disease among children less than 5 years of age per 100,000 inhabitants* (effect); and *Average rate of leptospirosis per 100,000 Inhabitants* (effect). The subsequent integrated indicators allowed officials to identify "in which Administrative District actions that change the pattern of the Driver, Pressure, or State components would have the most impact on population health, because of reduced exposure and/or recomposition of the environmental quality of affected sites."[44] [emphasis added]

The response to this pilot exercise was the "governments of the city of São Paulo and of the state of São Paulo, through their competent bodies,… adopting a series of measures to minimize or resolve environmental problems related to the degradation of water streams and the presence of waste (domestic and debris) in public areas."[44] These *actions* involved divisions such as the Public Works and Services, the Basic Sanitation Company of the State of São Paulo, the Municipal Housing Secretariat, along with the Green and Environment Secretariat.[44]

Scotland – Good Places, Better Health Program. The *Good Places, Better Health*[45] policy initiative in Scotland was developed to make better connections between health, well-being and the physical environments in which people live, work, are educated, and spend their leisure time (see Table 2-1). This initiative relied on an approach to framing issues in environment and health with explicit reference to the many factors which bear upon human health and well-being. As noted in guidance documents for this project, "the expansion of public health interest beyond the usual areas of immediate and discrete harms, such as toxics exposure, into physical and operational designs that shape the way people live, work, and interact with their communities is a recognition that when it comes to health- everything matters."[45]

The DPSEEA framework is cited as an organizing principle behind *Good Health, Better Places.*[46] The DPSEEA-based approach adopted in this initiative forms the basis for intelligence and data[47] gathering, for analysing relationships, and for developing clear, evidence-informed advice to the policy constituency (e.g. on the efficacy of existing policies and actions and those which are under consideration).

One key goal of the *Good Health, Better Places* initiative was to present coherent and unified messages to policy-makers across multiple disciplines, based on a deep understanding of the larger social context. This strategy was implemented after elements of the model framework were filled in. This process results in a 'populated model,' based on the concerns raised by stakeholders during workshops facilitated by topic experts and practitioners. After they have been validated with reference to scientific, epidemiological etc. literature; and appraised for practicality and coherence in workshops of field practitioners, the populated models are sometimes said to represent "maps of the environmental health territory."[30]

In keeping with the cross-cutting aspirations of the *Good Places, Better Health*[45] initiative, recommendations that emerge from the expert group are directed to a spectrum of policy interests across the government. These policy interests range from education, justice, planning, transport, and under-served communities, to economists, and of course the health and environmental policy-makers. The messages to these policy-makers relate to, for example, a damaging absence of data about a key variable bearing upon the problem; a knowledge gap (indicating a need for further research or evaluation); a discernible policy void; or perhaps an existing policy which has been found to be poorly targeted, lacking in impact or impeded in its implementation.

This Scottish case is the most wide-ranging and inclusive illustration of how a structured framework such as DPSEEA can serve as both a design and self-auditing tool. In this example, it helped provide a way for attributes of both traditional human health concern and other considerations of well-being to be effectively integrated with, and

communicated to, the medical community (traditionally holding a more singular focus on illness and harm) along with other policy constituencies and stakeholder groups (whose purview may not traditionally include human health and well-being).

Table 2-1. **Good Places, Better Health: A New Approach to Environment and Health in Scotland. Implementation Plan**

http://www.scotland.gov.uk/Publications/2008/12/11090318/0
"The Scottish Government is committed to creating a wealthier and fairer, smarter, healthier, safer and stronger, and greener Scotland. Through these strategic objectives we aim to deliver on the central purpose of creating a more successful country, with opportunities for all of Scotland to flourish, through increasing sustainable economic growth. *Good Places, Better Health* recognizes that to deliver on the Government's purpose, themes, and national outcomes there is a need for greater connections around how physical environment influences health. In *Equally Well*,[47] the Health Inequalities Task Force highlighted the need to work to reduce further people's exposure to factors in their physical and social environments that cause stress, damage health and wellbeing and lead to inequalities. We know that the physical environment that surrounds us is key to our health and well-being. Historically, we have focused (very successfully) on creating environments free from significant hazards. Whilst this continues to be important we now recognize an additional need to create positive physical environments which nurture better health and well-being. The relationship between environment and health is complicated and creating safe and positive environments for health requires us to think, plan and deliver in new and more effective ways. The Scottish government has established National Outcomes that it sees as part of good governance for "creating safe and positive environments which nurture better and more equal health and wellbeing." These core National Outcomes are supported by an understanding that seeks to integrate sectors as diverse as health, transportation, public safety, and economic development. To measure progress towards the National Outcomes the Scottish government has selected 45 indicators which most clearly show progress towards the achievement of a more successful and prosperous Scotland."[48]

Flexibility in Applying Frameworks.

The DPSIR/DPSEEA frameworks are meant to be flexible in their applications. Their purpose is to organize thinking about complex social and environmental issues, not to limit them. Niemeijer and de Groot[49] recently argued for a move from causal chains to causal networks in framing environmental indicators. They posit that the DPSIR and related frameworks rely on simple uni-directional chains of causality, ignoring feedbacks and emphasizing one-to-one relationships at the expense of one-to-many, many-to-one, and many-to-many relationships. While there is a danger that this might occur in practice, it is by no means inherent in the frameworks themselves. Rothman and Robinson[50] had already pointed to the importance of feedbacks and complex dynamics in early discussions of conceptual frameworks for integrated assessments and studies such as the *North American Environmental Outlook to 2030*[51] have been explicit about the role of common set of drivers causing multiple environmental pressures and impacts. In the latter, an additional set of "meta-forces", representing important socio-economic developments and global environmental changes were also added to better clarify global forces in what was essentially a regional report.

Many countries have built active and detailed information acquisition and management systems to better understand conditions and trends. Those data, if used in a more systemic and integrated fashion, can provide the backbone of a regulatory environment that is both more transparent to the stakeholder community and based on a clearer understanding of the nature and pace of change in social and environmental systems.

An earlier quote from Turner, et al. (2000)[13] introduced the concept and challenges of the use of the DPSIR to organize relevant socio-economic, environmental, and governance indicators. Given the complexity of understanding the relationship between the environment and human health and well-being (as well as the embedded nuances of that relationship), embracing the idea that the DPSIR or DPSEEA can be used as an organizing framework for indicator identification and use is critical.

Where the starting point is an environmental *state*, the procedure for applying the DPSEEA framework is essentially the same. Using the example of coastal water contaminated with fecal pathogens, the DPSEEA model demands consideration of the manmade *pressures* and *drivers* which create that environmental *state*. It is then necessary to consider the nature of any potential human *exposure* (e.g. ingestion of contaminated seawater or seafood), and any plausible health *effect*(*s*) (i.e. in this case, gastrointestinal illness). The contextual factors which influence exposure in this instance might include engagement by the individual in water sports or shellfish harvesting; and different contextual factors, such as immune status, might also influence likelihood of disease in the exposed individual.

Irrespective of the sequence in which a framework is populated, the final step is to incorporate within the model any existing policies or *actions*; and, if required, any additional policies or *actions* which might be considered likely to provide benefit value. This example reflects a common situation, the identification of an environmental health (*state*) concern, and the subsequent population of the model elements. While the

identification of a problem is usually the first step, the entire framework cycle, including indicator monitoring, follow-up of any *action/response*, and evaluation of success, should be a part of a comprehensive management plan.

This flexibility to modify the DPSIR/DPSEEA frameworks, while still maintaining their essential character, is also important when considering the focus of any particular application and the placement of any particular indicator within the framework. In viewing the real system as a complex causal network encompassing many causal chains, the same indicators may fit into different points along the chain. For example, while coastal population in China may be a significant *driver* indicator from the perspective of coastal pollution, it can also be seen as a *state* variable driven by economic imbalances leading to migration from the interior of the country. Which is the case will depend on the issue context and the questions of concern.

In short, the DPSIR and DPSEEA can be used as effective tools in both ensuring: (i) the full range of applicable attributes are considered in addressing the complex interdependencies linking the coastal environment and human well-being; and, (ii) that critical indicators are assessed to better understand the sources and consequences of this nuanced system.

Conclusion.

The value of any framework as a research tool is primarily through its ability to allow for organization of data and information at the start of the analysis process, and at the end of the process, for an auditing of outcomes (whether theoretical or actual). In

addition, using a defined framework can allow for the comparison of analysis outcomes when the subject matter is altered.

Ocean and Human Health questions are by their nature interdisciplinary, variously combining aspects of fields such as ecology, biology, chemistry, economics, psychology, toxicology, statistics, and oceanography. Since data from these disciplines can be structured in a wide variety of formats, the product of specialized research methods with their own assumptions, the results can sometimes be inaccessible to the non-specialist. Decision-makers (e.g., politicians, natural resource managers, planning committees) however, must be able to interpret or use specialized data in order to meet policy goals within their sphere of influence. A framework then, is a tool that allows people use information within their decision-making process. Not all frameworks are created equal. Internalized judgment frameworks may suffer from a wide variety of biases (e.g. imaginability, illusory correlation, anchoring bias, or examples of prior outcomes[52]). Frameworks used in public decision-making should therefore strive to be transparent about assumptions and any value judgments embedded in the framework itself. The frameworks detailed in this chapter, the Driver-Pressure-State-Impact-Response framework (DPSIR) and the Driver-Pressure-State-Exposure-Effect-Action Framework (DPSEEA), are two organizing and auditing frameworks that emerged from the intergovernmental community and are in broad use around the world.

The DPSIR has been described as "a useful tool to support decision making by means of showing solid evidence with alternatives and decision options, rather than by

presenting predetermined solutions."[53] The DPSEEA framework is structurally and philosophically similar to the DPSIR, but has been modified through its use by the public health community in light of their disciplinary focus and language.[30]

While it can take longer to identify indicators that are appropriate to address the problems identified within a DPSIR/DPSEEA framework, and require more input from a wider variety of stakeholders, in theory the indicators chosen should be better representatives of the problem at hand, as they will be the result of a more comprehensive understanding of the inputs to those very problems.

While the DPSIR and DPSEEA frameworks may seem simplistic and uni-directional to critics,[54] they can provide flexibility and transparency in decision-making processes if all parts of the framework are described. By examining the full spectrum of causal relationships that lead to a specific problem of interest decision-makers should be prompted to fully understand the tradeoffs between different *responses/actions*. This should lead to a more efficient use of resources than choices made without a framework, because they will be directed at the solutions that have the greatest possible impact given the resources at hand. By having a built-in auditing function, in the form of a feedback-loop structure, the DPSIR/DPSEEA framework can then be used to assess the subsequent success or failure of any policy/program. Without such assessment, interested parties would have no way of measuring the results of policy choices that can impact their health and well-being.

Literature Cited.

1. Simon, H. 1965. The architecture of complexity. General Systems Yearbook. 10: 43-64.

2. IPCC, Intergovernmental Panel on Climate Change. 2007. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Paris, France.

3. IPCC, Intergovernmental Panel on Climate Change. 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Paris, France.

4. IPCC, Intergovernmental Panel on Climate Change. 2007. Climate Change 2007: Mitigation of Climate Change. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Paris, France.

5. Millennium Ecosystem Assessment, Reid, W. V., Mooney, H. A., Cropper, A., Capistrano, D., Carpenter, S. R., Chopra, K., Dasgupta, P., Dietz, T., Duraiappah, A. K., Hassan, R., Kasperson, R., Leemans, R., May, R. M., McMichael, T., Pingali, P., Samper, C., Scholes, R., Watson, R. T., Zakri, A. H., Shidong, Z., Ash, N. J., Bennett, E., Kumar, P., Lee, M. J., Raudsepp-Hearne, C., Simons, H., Thonell, J., Zurek, M. B. 2005. Ecosystems and Human Well-being: Synthesis. Island Press. Washington, D.C.

6. GOOS, Coastal Panel of the United Nations' Global Ocean Observing System. 2003. The Integrated Strategic Design Plan for the Coastal Observations Module of the Global Ocean Observing System. GOOS Report No. 125. GOOS Report No. 125. GOOS, Coastal Panel of the United Nations' Global Ocean Observing System.

7. GOOS, Coastal Panel of the United Nations' Global Ocean Observing System. 2005. An Implementation Strategy for the Coastal Module of the Global Ocean Observing System. GOOS Report No. 148. GOOS Report No. 148. IOC Information Documents Series, No. 1217.

8. GTOS, Christian, R., Baird, D., Bowen, R. E., Clark, D., DiGiacomo, P., de Mora, S., Jimenez, J., Kineman, J., Mazzilli, S., Servin, G., Talaue-McManus, L., Viaroli, P., Yap, H. 2005. Coastal GTOS Draft Strategic Design and Phase I Implementation Plan. GTOS Report No. 36. FAO, Rome. Rome, Italy.

9. Panetta, L. E., Pew Oceans Commission. 2003. America's Living Oceans: Charting A Course For Sea Change: A Report to the Nation: Recommendations for a New Ocean Policy. Executive Summary. Pew Oceans Commission. Washington, D.C. 1-9.

10. IOC, Intergovernmental Oceanographic Commission, ICAM Programme. 2003. A Reference Guide on the Use of Indicators for Integrated Coastal Management. ICAM Dossier No. 1, 2003. ICAM Dossier No. 1, 2003. Paris, France.

11. Belfiore, S., Barbière, J., Bowen, R., Cicin-Sain, B., Ehler, C., Mageau, C., McDougall, D., Siron, R. 2006. A Handbook for Measuring the Progress and Outcomes of Integrated Coastal and Ocean Management. UNESCO Intergovernmental Oceanographic Commission (IOC) Manuals and Guides, 46; ICAM Dossier, 2. .

12. GESAMP, United Nations Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection, Advisory Committee on Protection of the Sea. 2001. A sea of troubles: Rep. Stud. GESAMP No. 70. 70. Printed by GRID-Arendal for the United Nations Environment Programme.

13. Turner, R. K., Brouwer, R., Georgiou, S., Bateman, I. J. 2000. Ecosystem functions and services: an integrated framework and case study for environmental evaluation. CSERGE Working Paper GEC 2000-21. Centre for Social and Economic Research on the Global Environment (CSERGE). 36.

14. Agard, J., Alcamo, J., Ash, N., Arthurton, R., Barker, S., Barr, J., Baste, I., Chambers, W. B., Dent, D., Fazel, A., Gitay, H., Huber, M., Jäger, J., Kuylenstierna, J. C. I., King, P. N., Kok, M. T. J., Levy, M. A., Mafuta, C., Martino, D., Panwar, T. S., Rast, W., Rothman, D. S., Varughese, G. C., Zommers, Z. 2007. Global Environment Outlook 4: Environment for Development (GEO4). United Nations Environment Programme. Nairobi, Kenya. 978-992.

15. Depledge, M. H., Bird, W. J. 2009. The Blue Gym: Health and Wellbeing from our Coasts. Mar. Pollut. Bull. 58: 947-948.

16. Bauman, A., Smith, B., Stoker, L., Bellew, B., Booth, M. 1999. Geographical influences upon physical activity participation: evidence of a 'coastal effect'. Aust. N. Z. J. Public Health. 23: 322-324.

17. White, M., Smith, A., Humphryes, K., Pahl, S., Snelling, D., Depledge, M. 2010. Blue space: The importance of water for preference, affect, and restorativeness ratings of natural and built scenes. J. Environ. Psychol. 30: 482-493.

18. Food and Agriculture Organization of the United Nations, Fisheries and Aquaculture Department. 2009. The State of World Fisheries and Aquaculture 2008. Electronic Publishing Policy and Support Branch. Rome, Italy. 1-1-196.

19. Organization for Economic Cooperation and Development (OECD). 1991. Environmental Indicators, a Preliminary Set.

20. OECD, Organization for Economic Cooperation and Development. 1993. OECD core set of indicators for environmental performance reviews. Paris, France.

21. IIED, International Institute for Environment and Development. 2002. National strategies for sustainable development: the pressure state response framework.

22. LEAD, Livestock Environment and Development Initiative. 2002. Livestock and environmental toolbox; pressure state response framework and environmental indicators. www.Virtualcentre.org/en/dec/toolbox/Refer/EnvIndi.htm (Accessed July 1, 2007).

23. Kjellstrom, T., Corvalan, C. 1995. Framework for the development of environmental health indicators. World Health Stat. Q. 48: 144-154.

24. European Commission. 2002. Towards Environmental Performance Indicators for the European Union (EU): A European System of Environmental Indicators; First Publication.

25. United Nations, Department of Economic and Social Affairs, Division for Sustainable Development, Commission on Sustainable Development. 2001. Indicators of Sustainable Development: Framework and Methodologies: Background Paper No. 3 for the 9th Session of the Commission on Sustainable Development. United Nations. New York, NY, USA.

26. European Environment Agency. 1995. Europe's Environment: The Dobris Assessment. European Environment Agency.

27. Holten-Andersen, J., Paalby, H., Christensen, N., Wier, M., Andersen, F. M. 1995. Recommendations on strategies for integrated assessment of broad environmental problems. Report submitted to the European Environment Agency (EEA) by the National Environmental Research Institute (NERI). European Environment Agency. Copenhagen, Denmark.

28. Corvalán, C., Briggs, D., Kjellstrom, T. 1996. Development of environmental health indicators. *In* Linkage Methods for Environment and Health Analysis: General Guidelines. D. J. Briggs, C. Corvalán, and M. Nurminen, Ed.: 19-53. World Health Organization. Geneva, Switzerland.

29. Jordan, H., Dunt, D., Dunn, L., Verrinder, G. 2008. Evaluating the Actions towards Environmental Health using DPSEEA and Program Logic. Environ. Health. 8: 11-25.

30. Morris, G. P., Beck, S. A., Hanlon, P., Robertson, R. 2006. Getting strategic about the environment and health. Public Health. 120: 889-903.

31. Briggs, D. J. 2008. A framework for integrated environmental health impact assessment of systemic risks. Environ. Health. 7: 61.

32. Borja, Á., Galparsoro, I., Solaun, O., Muxika, I., Tello, E. M., Uriarte, A., Valencia, V. 2006. The European Water Framework Directive and the DPSIR, a methodological approach to assess the risk of failing to achieve good ecological status. Estuar. Coast. Shelf Sci. 66: 84-96.

33. Sekovski, I., Newton, A., Dennison, W. C. 2012. Megacities in the coastal zone: Using a driver-pressure-state-impact-response framework to address complex environmental problems. Estuar. Coast. Shelf Sci. 96: 48-59.

34. Jago-on, K. A. B., Kaneko, S., Fujikura, R., Fujiwara, A., Imai, T., Matsumoto, T., Zhang, J., Tanikawa, H., Tanaka, K., Lee, B. 2009. Urbanization and subsurface environmental issues: An attempt at DPSIR model application in Asian cities. Sci. Total Environ. 407: 3089-3104.

35. Palmer, B. J., Hill, T. R., Mcgregor, G. K., Paterson, A. W. 2011. An Assessment of Coastal Development and Land Use Change Using the DPSIR Framework: Case Studies from the Eastern Cape, South Africa. Coast. Manage. 39: 158-174.

36. Campuzano, F. J., Mateus, M. D., Leitão, P. C., Leitão, P. C., Marín, V. H., Delgado, L. E., Tironi, A., Pierini, J. O., Sampaio, A. F. P., Almeida, P., Neves, R. J. 2011. Integrated coastal zone management in South America: A look at three contrasting systems. Ocean Coast. Manage.

37. Scheren, P. A. G. M., Kroeze, C., Janssen, F. J. J. G., Hordijk, L., Ptasinski, K. J. 2004. Integrated water pollution assessment of the Ebrié Lagoon, Ivory Coast, West Africa. J. Mar. Syst. 44: 1-17.

38. Mangi, S. C., Roberts, C. M., Rodwell, L. D. 2007. Reef fisheries management in Kenya: Preliminary approach using the driver–pressure–state–impacts–response (DPSIR) scheme of indicators. Ocean Coast. Manage. 50: 463-480.

39. Food and Agriculture Organization of the United Nations (Fisheries and Aquaculture Department). 2010. The State of World Fisheries and Aquaculture 2010. Rome, Italy. 1-218.

40. Nobre, A. M., Robertson-Andersson, D., Neori, A., Sankar, K. 2010. Ecological–economic assessment of aquaculture options: Comparison between abalone monoculture and integrated multi-trophic aquaculture of abalone and seaweeds. Aquaculture. 306: 116-126.

41. Ostoich, M., Aimo, E., Fassina, D., Barbaro, J., Vazzoler, M., Soccorso, C., Rossi, C. 2011. Biologic impact on the coastal belt of the province of Venice (Italy, Northern Adriatic Sea): preliminary analysis for the characterization of the bathing water profile. Environ Sci Pollut Res. 18: 247-259.

42. Lin, T., Xue, X. Z., Lu, C. Y. 2007. Analysis of coastal wetland changes using the "DPSIR" model: a case study in Xiamen, China. Coast. Manage. 35: 289-303.

43. Vella, P., Bowen, R. E., Frankic, A. 2009. An evolving protocol to identify key stakeholder-influenced indicators of coastal change: the case of Marine Protected Areas. ICES Journal of Marine Science. 66: 203-213.

44. Hacon, S. (. 2008. GEO Health: City of São Paulo: Green and Health Environments Project: Summary and Lessons Learned. City Government of São Paulo; FIOCRUZ; United Nations Environmental Program. São Paulo, Brazil. 1-48.

45. The Scottish Government. 2008. Good Places, Better Health: a New Approach to the Environment and Health in Scotland: Implementation Plan. The Scottish Government. Edinburgh. 1-22.

46. The Scottish Government. 2010. Publications: 2010: July: Scottish Social Attitudes Survey 2009: Sustainable-Part 3. The Scottish Government. Edinburgh, Scotland, UK. http://www.scotland.gov.uk/Publications/2010/07/02134238/3 (Accessed September 26, 2012).

47. The Scottish Government, Ministerial Task Force on Health Inequalities. 2008. Equally Well: Report of the Ministerial Task Force on Health Inequalities. The Scottish Government. Edinburgh, Scotland. 1-75.

48. The Scottish Government. 2012. About: Performance: Scotland Performs: National Outcomes (2007). Edinburgh, Scotland, UK. http://www.scotland.gov.uk/About/scotPerforms/indicators (Accessed September 26, 2012).

49. Niemeijer, D., de Groot, R. S. 2008. Framing environmental indicators: moving from causal chains to causal networks. Environ. Dev. Sustainability. 10: 89-106.

50. Rothman, D. S., Robinson, J. B. 1997. Growing pains: a conceptual framework for considering integrated assessments. Environ. Monit. Assess. 46: 23-43.

51. Commission for Environmental Cooperation of North America. 2010. North American Environmental Outlook to 2030. Communications Department of the CEC Secretariat. Montreal, Canada. 1-84.

52. Tversky, A., Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. Science. 185: 1124-1131.

53. Tscherning, K., Helming, K., Krippner, B., Sieber, S. 2012. Does research applying the DPSIR framework support decision making? Land Use Policy. 29: 102-110.

54. Svarstad, H., Petersen, L. K., Rothman, D., Siepel, H., Wätzold, F. 2008. Discursive biases of the environmental research framework DPSIR. Land Use Policy. 25: 116-125.

CHAPTER 3

WATERSHED DEMOGRAPHIC ESTIMATES AND IDENTIFICATION OF KEY INDICATORS FOR MULTIPLE MARINE-SOURCED RISKS IN MASSACHUSETTS BAY

Abstract. This chapter starts with a discussion about the current gaps in epidemiological data and the need for better ways of understanding and forecasting multiple types of marine-sourced risks to support public health efforts regarding recreation and shellfishing in coastal waters. Because a population's overall risk depends in part on its demographic characteristics, this chapter describes demographic trends in six coastal watersheds around Massachusetts Bay from 2000 to 2010 and discusses how these may be changing the population vulnerability to marine-sourced risks. This chapter then presents five marine-sourced risks known to exist in the coastal waters of Massachusetts Bay through sections describing their biology, known epidemiology, and known environmental or socio-economic influences reported in scientific literature. All of the marine-sourced risks described are known to exist in Massachusetts. This chapter then presents the results of this exercise in the form of a matrix showing high-value data types for the five specific risks addressed in this chapter which are: 1) *Enterococcus*

species bacteria; 2) *Vibrio parahaemolyticus* bacteria*;* 3) Hepatitis A Virus; 4) *Pseudo-nitzschia* species diatoms; and 5) Anthropogenic antibiotics. Based on the compilation of influencing factors we present a matrix of key indicators and a diagram illustrating the conceptual relationships between indicators and risks. These indicators can guide the development of an environmental model to forecast changes in these risks. Given the biological variety of these known marine-sourced risks, it is likely that current recreational- and seafood harvesting- water quality monitoring protocols do not fully account for changes in the true potential for exposure.

Chapter 2 Research Topics.

Onshore activities in coastal watersheds and offshore ocean processes can both influence the nearshore marine environment. The nearshore marine environment in turn can influence the health of humans who interact with it, either directly through physical contact or indirectly through consumption of local seafood. Therefore, it is important to understand the size and demographics of a human population in a coastal watershed, its estimated impact on the nearshore environment, and the marine risks which that population may encounter. The two research topics in this chapter deal with human populations and marine risks in the study area of Massachusetts Bay and its six neighboring coastal watersheds.

1) This chapter presents an original estimate of the number of people living in six watersheds bordering Massachusetts Bay along with key demographic characteristics that are known to influence population level vulnerability, especially to infectious diseases. This chapter argues that current water quality standards for recreational and shellfish

harvesting waters are not reflective of the full suite of known microbial risks and could be strengthened to better protect human health.

2) This chapter then presents the known epidemiology and natural history for a suite of 5 marine-sourced risk agents as examples of the diversity of marine-sourced risk categories that exist in any nearshore coastal environment where humans might recreate or harvest shellfish. For 5 specific marine-sourced risks known to exist in Massachusetts Bay we have assembled known or suspected environmental and socio-economic influences on their abundances as identified in scientific literature. Some of these influencing factors could be monitored through direct or proxy indicators to provide information for a model that estimates changes in risk potential in nearshore coastal waters. At present no such model exists. Work to develop a model for two of these risks is described in Chapters 3 and 4.

Introduction.

Environment-human health linkages are slowly being revealed in greater detail and complexity. Foundational ecological information that can help reveal these linkages likely already exists within individual scientific disciplines but is not being fully utilized to inform public health. Biological and ecological knowledge is traditionally found in the natural sciences, where it develops largely in isolation from the medical and social sciences. These disciplines have different cultures, terminology, and standards which present challenges to knowledge transfer. However, by using a synthesis approach that looks across disciplines to identify influences on the presence or abundance of a pathogen

or toxin where it might impact human health we can move beyond a singular reliance on historical epidemiological data to understand current health risks.

For many illnesses existing epidemiological data are not an accurate reflection of the true number of disease cases. Though highly variable, disease reporting completeness appears to be most strongly related to the disease or condition being reported.[1] In a review of disease reporting completeness from 1970 to 1999, researchers found that reporting completeness ranged from 9 to 99 percent with greatest completeness for tuberculosis, AIDS, and sexually transmitted diseases.[1] Surprisingly, increased completeness of reporting does not correspond to the number of people affected by a disease or category of illness, but instead seems to be influenced by the perceived seriousness of a disease or to the level of financial and human resources devoted to treatment and prevention.[1] There are whole disease categories that public health experts believe to be persistently under-reported. One such category is foodborne illness, estimated to affect 1 in 6 Americans annually.[1; 2] The highest rate of reported foodborne illness is associated with seafood consumption.[3] Another category of disease believed to be under-reported is that of marine-sourced diseases, which includes illnesses resulting from direct contact with harmful marine organisms and ingestion of contaminated seafood.[45] To clarify, some foodborne illness may have a marine-sourced origin but marine-sourced illness is not restricted to seafood-borne illnesses. Also, some pathogens may be transmitted through both marine-sourced and land-based pathways. A 2003 study estimated that globally, each year, there are over 120 million cases of marine-sourced gastrointestinal disease and more than 50 million cases of more severe respiratory

diseases caused by recreational exposure to polluted coastal waters.[5] Within the U.S. one study estimates that 5 million cases of gastrointestinal illness from beach exposure and over 3 million cases of seafood-borne illness occur in the U.S. annually.[6] The system for recording such diseases is described in the following section.

Marine-Sourced Diseases – Human Epidemiological Knowledge. In the U.S. many diseases of potential marine-sourced origin are considered 'reportable diseases', meaning if a healthcare provider or clinical laboratory suspects or confirms such an illness it must be reported to local or state public health authorities within a certain time frame (sometimes immediately). Some diseases are 'nationally notifiable,' meaning the U.S. Centers for Disease Control and Prevention (CDC) collects information on these diseases across the entire U.S. as they are reported by state and territorial public health agencies.[7] Many foodborne illness are nationally notifiable, data about these cases are assembled through a variety of CDC programs, some the programs relevant to this work are listed below.

- National Notifiable Diseases Surveillance System (NNDSS)[8], a nationwide collaboration between the CDC and all public health departments to share health information.
- Foodborne Disease Outbreak Surveillance System (FDOSS)[9] which collects data on foodborne disease outbreak reported by State and territorial public health departments.

- Cholera and Other Vibrio Illness Surveillance System (COVIS)[10] an online tool where health officials can report clinical data about *Vibrio* infections or cases of Cholera.
- National Electronic Norovirus Outbreak Network (CaliciNet)[7], a national surveillance network of 33 specially certified laboratories with the capacity to submit outbreak specimens for norovirus classification.

All disease outbreaks associated with recreational waters are notifiable to the CDC.

Through voluntary reporting by states and territories to the Waterborne Disease and Outbreak Surveillance System (WBDOSS)[11] the CDC collects outbreak data for treated waters (e.g., pools and spas) and untreated waters (e.g. lakes, rivers, ocean).[12] From 2007 to 2012 the WBDOSS received reports of 63 outbreaks associated with untreated waters, resulting in 1,261 reported cases of illness and at least 44 hospitalizations.[12-14] Of the 63 outbreaks associated with untreated waters 17 had unidentified etiology, in some cases these was a suspected, but never proven, causative agent.[12-14] Although the reported outbreak and case numbers are small relative to the millions of people that use recreational waters, these data demonstrate that both freshwater and ocean recreational waters continue to be a vector for human pathogens or other harmful compounds. Experts believe that the reported number of outbreaks and cases are much smaller than the true incidence due to multiple barriers to recording and reporting.[12-14]

Many factors may present barriers to outbreak reporting for recreational waterborne diseases including 1) mild illness; 2) small outbreak size; 3) long incubation

periods between exposure and onset of symptoms and subsequent attribution of illness to other sources; 4) the often transient nature of water contamination hindering traceability; 5) potential lack of communication between those who respond to outbreaks of chemical origin (e.g., hazardous materials personnel) and those who usually report outbreaks (e.g., infectious disease epidemiologists); and 6) many waterborne illnesses are self-limiting (not spread to another person) so medical advice is not sought.[414] In other words, even though a disease is reportable or notifiable there is no guarantee that all cases are reported to public health authorities. This gap in reporting results in a gap in our knowledge between the known (reported) burden of disease and the true burden of disease on a population. This situation may be self-reinforcing. A complicating factor in the understanding and management of recreational coastal waters is the difference in perceived seriousness and prevalence of different human health risks. Slovic (1987) noted that when it comes to evaluating hazards the majority of citizens rely on intuitive risk judgments, typically called "risk perceptions," which are largely based on the news media.[15] Risk perception and attitudes can be influenced by factors such as 'voluntariness of exposure', familiarity, control, catastrophic potential, and level of knowledge.[15] Risk perception contributes to the varying rates of disease completeness reporting, illnesses with a greater social stigma tend to have better reporting completeness.[1] Recreational pursuits and daily food choices are arguably perceived as voluntary, familiar, and under close control for most adults, so there is a low risk perception around these activities. If people do not think an illness is serious or significant enough to seek medical attention then cases are not recorded by public health

authorities, official records then underestimate the number of cases and true costs are difficult or impossible to quantify. If there is no true understanding of the problem's scope the public chooses to devote resources to other issues. As a consequence, knowledge about the DPSIR framework elements of *drivers*, *pressures*, and *states* that may influence marine-sourced risk exposure, and the true *impacts* of resulting illness does not improve and the ongoing costs to society go unrecorded. Thus, we are limited in our ability to quantitatively evaluate marine-sourced risks within the DPSIR framework (as described in Chapter 1) and must examine other evidence which can improve our understanding of these risks. This is true for any underreported illness, including those with obscured etiologies due to their environmental origins – the case many marine-sourced illnesses.

Despite these knowledge gaps, for some environmentally-linked illnesses there are estimates of their burden on society. One example of this is otitis externa (swimmer's ear), commonly caused by bacterial infection in the outer ear and associated with recreational activities that introduce water and bacteria into the ear canal.[16] In the U.S. in 2007 there were 2.4 million outpatient medical visits for otitis externa, and over 4,000 additional cases that required hospitalization. One study estimates annual costs to treat outpatient cases of otitis externa in the U.S. at $500 million, with hospitalization costs for severe cases totaling over $27 million. [16] Although otitis externa is not strictly marine-sourced it is strongly associated with recreational water exposure. Few estimates exist for the cost to society of illnesses attributed strictly to marine-sourced sources including marine pathogens and toxins. Ralston, Kite-Powell, and Beet (2011) conservatively

estimate that in the U.S. gastrointestinal illness from exposure to pathogens via beach recreation costs US$300 million annually, that food-borne disease from *identified* marine pathogens and toxins costs US$350 million annually, and that *unidentified* seafood-borne vectors cost US$300 million annually.[6] The recognition that illnesses which are not reflected in official epidemiological data still have a cost to society is an important one as these illnesses may warrant more attention than they currently receive. We suggest that marine-sourced illnesses fall into this group and that they warrant more attention from public health authorities. When multiple illnesses are linked to specific recreational activities or patterns of food consumption and it is not feasible to initiative extensive direct monitoring for all types of risk precursors, we suggest that improving the understanding of underlying drivers of disease risk potential can be used to help protect public health. The first step towards improved public health protection is to identify the risk(s) of interest, the routes of exposure, and the population(s) most at risk for exposure. To place this idea within a DPSIR framework, the local human population is part of the *pressure* and *state*. A *pressure* because of the impacts of humans on the environment and their potential for pathogen release into the environment, and a *state* because population characteristics influence the potential severity of effects of an exposure to marine-sourced risks (the *impact* of interest). We know that *impacts* from marine-sourced risks are under-reported. Therefore the goal of this work is to investigate the feasibility of using environmental modeling to assess the potential for changes in the *state* of environmental conditions and thereby imply changes to *impacts* on risk potential from certain risks. We propose to use environmental modeling because it has been used to successfully predict

risks from other environmentally influenced disease (examples are discussed in Chapter 3). Such modeling could serve public health interests in instances where multiple risks co-exist, can increase or decrease quickly, and epidemiological data is known to be insufficient to use as a predictor of future illness patterns. At present we are not aware of any such model or tool for the specific marine-sourced risks in Massachusetts Bay discussed in this chapter.

Organizing Marine-Sourced Risks by Category. In the marine environment there may be multiple microbiological health risks co-existing in space in time.[4] These marine-sourced risks may take different forms and have shared, or unique, factors influencing their risk potential at any given time (in DPSIR framework terms these correspond to *pressures* and *states*). One way to organize an assessment of these multiple co-existing risks is to group them based on similarities, either in their underlying biology or in the type of risk they present to humans. Based on the knowledge that there are multiple types of microbiological marine-sourced risks co-existing in the same marine space we build upon the work of Bienfang et al. (2011) and identify five major categories of microbiological risk[4]: human viruses, indigenous bacteria, introduced bacteria, natural marine toxins, and anthropogenic compounds. There are multiple examples of specific risks in each category:

1) Human viruses: Hepatitis A Virus, Norwalk/Noroviruses, Adenoviruses

2) Indigenous bacteria: *Vibrio parahaemolyticus*, *Vibrio vulnificus*, *Listeria* species

3) Introduced bacteria: *Enterococcus* species; *Escherichia coli*, *Streptococcus* species

4) Natural marine toxins: Domoic Acid, Ciguatoxins, Saxitoxins

5) Anthropogenic compounds: antibiotics, heavy metals, chlorinated chemicals

The specific risks used in this type of exercise would vary according to local conditions and the health concerns of interest. *This paper will use* the following risks as examples representative of their category, all are known to exist in Massachusetts:

1) *Enterococcus* bacteria, the current water quality monitoring standard;

2) *Vibrio parahaemolyticus* bacteria, all *Vibrio* infections are reportable in Massachusetts;

3) Hepatitis A Virus, a reportable disease in Massachusetts;

4) *Pseudo-nitzschia* genus diatoms, because they may produce a toxin that can accumulate in shellfish and cause a reportable foodborne illness; and

5) Anthropogenic antibiotics, because they are known to be released in the effluent of wastewater treatment plants and influence the development of antibiotic resistance in bacteria.

People may be exposed to specific marine-sourced risks through multiple pathways. The next section describes two important routes of exposure, coastal recreation and consumption of raw shellfish harvested in nearshore environments. We argue that current water quality standards for these activities are insufficient and would benefit from the outputs of a model able to forecast potential changes in marine-sourced risks.

Potential Exposure Routes for Marine-Source Risks: Coastal Recreation and Raw Shellfish Consumption.

Beach attendance, waterborne recreation, and consumption of raw shellfish present opportunities for contact with marine-sourced risks. Beach visits and sea bathing are popular recreational activities, with millions of visitors to U.S. beaches every year.[16] In Massachusetts, areas such as Cape Cod see large influxes of summer tourist visitors drawn largely by ocean-based recreational activities.[17] A study based on a survey of Massachusetts residents estimated that there are 111 million person-trips to Massachusetts coastal beaches and shorelines every year.[17] Coastal beaches are popular recreation sites for residents and tourists alike, Massachusetts residents reported a median of 12 visits per year and people residing in close proximity to beaches reported more visits in general.[17] In addition, Massachusetts fisheries land shellfish worth millions of dollars every year, some of which is consumed locally.[18] If people who live in coastal areas visit shoreline beaches, and potentially consume locally produced shellfish, more often than others this suggests that residents of coastal areas have a higher likelihood of encountering marine-sourced risks than the general population and so might benefit from more targeted information.

Coastal Recreation: Activities and Risk Exposure. The spatial extent of public marine beaches and semi-public marine beaches (i.e., where a landowner may charge a fee for public access) around Massachusetts Bay is shown in Figure 1 below. There are numerous marine beaches bordering Massachusetts Bay, providing ample opportunity for coastal recreation. People can be exposed to marine-sourced hazards through multiple

routes, including accidental ingestion during sea bathing or sand-contact activities. Therefore, it is valuable to understand conditions within the source environment (*state* in DPSIR terms) for these risks.

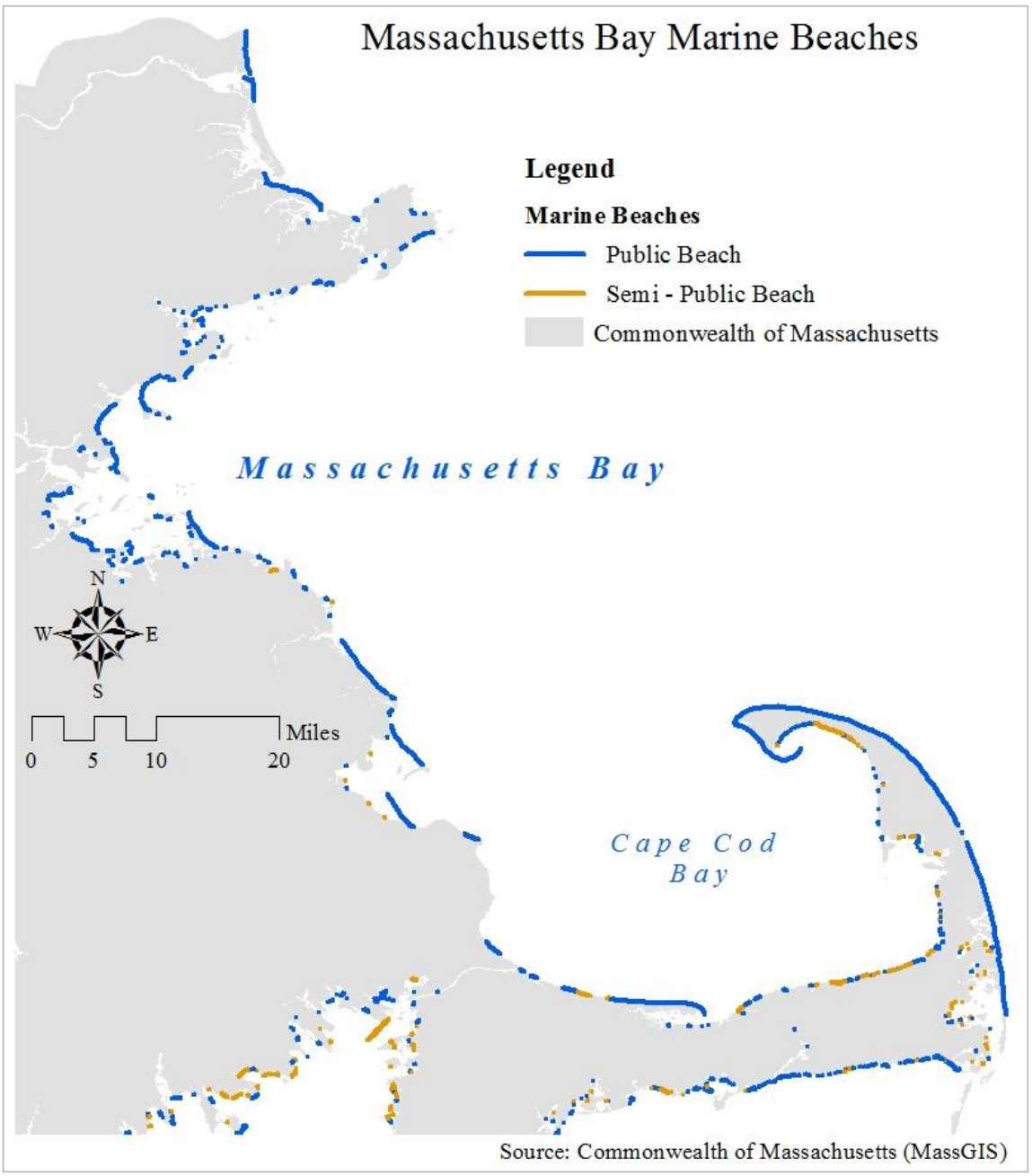

Figure 3-1. Massachusetts Bay marine beaches. Source: MassGIS, map by author.

Studies have estimated the amount of water accidentally ingested during water-based recreational activities. Swimming is associated with the highest amount of water ingested when compared to other surface water-based recreation activities such as boating and wading, estimates range from 16mL to 35mL water ingestion per hour of swimming activity.[19-21] Studies have also found that children ingest more water than adults, sometimes twice as much, and that males ingest more than females. [19-21]

In addition to ingestion while swimming, beach-goers may come into contact with pathogenic organisms present in beach sand. One epidemiological study showed that 'sand contact activities', including digging in sand or being buried in sand, were positively associated with enteric illness.[22] *Enterococcus* bacteria have been isolated from both wet and dry sands at beaches with high and low numbers of human visitors[23], and *Enterococcus* bacteria have been shown to replicate in laboratory experiments using beach sand microcosms that mimic natural conditions.[24] In addition to isolating *Enterococcus* bacteria from beach sand, Methicillin-resistant *Staphylococcus* aureus bacteria (MRSA, a public health threat because of its resistance to antibiotics) has been isolated from beach sand and seawater in southern California and Washington state, fueling speculation that public beaches may be a previously overlooked environmental reservoir for MRSA transmission.[25] To our knowledge beach sands in Massachusetts are not monitored for these types of risks. Current water quality standards for Massachusetts marine beaches are discussed in the next section.

Coastal Recreation: Massachusetts Water Quality Standards. Existing public health protection measures for marine recreational waters are largely based on monitoring for high numbers of *Enterococcus* genus bacteria. The current water quality standard is 104 colony forming units (CFU) of *Enterococcus* per 100 mL of water sampled. [26] There is evidence that traditional fecal coliform indicator bacteria such as *Enterococcus* do not always closely track or accurately predict the absence of other risk factors relevant to public health. [27; 28] Research supports the assertion that commonly monitored *Enterococcus* species are not indicative of the presence of human pathogenic viruses. Specifically this has been shown for the adenovirus, enterovirus, and astrovirus groups in Massachusetts Bay in a study spanning the years 1998 – 2002.[29] Although human enteric viruses were significantly correlated with certain types of coliphages (viruses that infect *E. coli* bacteria), the *Enterococcus* indicator bacteria were not significantly correlated to any of the virus or phage groups studied.[29] Virus levels in seawater have not been ignored by public health authorities, but detection methods have historically been limited and costly. Older viral isolation and culture methods are less sensitive than newer methods; one study showed 23% vs 46% positives respectively for paired samples.[29] If recreational water can be a vector for both harmful bacteria and viruses, and the two categories of organisms do not co-vary in abundance, then sampling programs should be updated to test for viruses using modern techniques. Despite the knowledge that multiple types of biological risks exist, state water quality regulations continue to largely depend on *Enterococcus* sampling and chemical hazard monitoring (e.g., for oil spills).

Raw Shellfish: Consumption Activities and Exposure. Raw shellfish consumption is another potential vector for marine-sourced risks. In 2014 Massachusetts towns had over 1,000 acres under cultivation for aquaculture of multiple shellfish species, including quahogs, oysters, softshell clams, blue mussels, and razor clams.[18] The value of combined aquaculture landings (from all waters of Massachusetts not just Massachusetts Bay) in 2014 was over US$19 million.[18] In addition to aquacultured shellfish there are also wild caught shellfish. Massachusetts inshore and intertidal shellfish landings (both wild caught and aquacultured) were valued at approximately US$30 million in 2014.[18] There were over 30 million American oyster pieces (the unit of measure) landed by Massachusetts aquaculturists in 2014.[18] It is likely that some of these aquacultured and wild caught inshore/intertidal shellfish were consumed locally. Locally harvested shellfish can be sold at any month of the year[30], therefore the potential for consumption of marine-sourced risks exists year-round.

Raw shellfish: Massachusetts Nearshore Harvest Water Quality Standards. In Massachusetts, cities and towns are responsible for managing most shellfisheries within their boundaries that are not closed by the state for public health reasons.[31] Areas officially open to shellfish harvesting are known as "approved" or "open", areas closed to harvesting are referred to as "restricted", "prohibited", or "closed." For a map of these areas in Massachusetts see Figure 3-7. Existing public health protection measures for marine shellfish harvesting waters are based on "1) an evaluation of pollution sources that may affect an area, 2) evaluation of hydrographic and meteorological characteristics that may affect distribution of pollutants, and 3) an assessment of water quality."[32]

However, given the dynamic nature of the coastal environment, the lag time between current sampling and reporting practices, and the variety of risks, it is unlikely that existing monitoring regimes adequately reflect the full suite of risks. An additional consideration is that new risks may be emerging as environmental conditions change. For example, the year 2011 saw the first confirmed case of *Vibrio parahaemolyticus* bacteria food poisoning from shellfish harvested from Massachusetts Bay (specifically Eastern Cape Cod Bay);[33-35] in 2012 there were 9 confirmed cases, in 2013 there were 33 cases, and in 2014 there were 11 cases.[18] *V. parahaemolyticus* has been known to exist in New England coastal waters for over 40 years[36; 37] and Vibriosis is a reportable disease in Massachusetts[38; 39] so the cases starting in 2011 might simply be a result of better detection, not the emergence of a new pathogen. As a result of the confirmed *V. parahaemolyticus* cases the Massachusetts Division of Marine Fisheries (MA-DMF) issued new regulations for commercial oyster harvesting and handling during warmer times of the year.[18] These regulations do not require sampling for *V. parahaemolyticus* in the water by harvesters or town public health boards. [31; 35; 40] However, MA-DMF does collect oyster tissue from major harvesting areas and analyze it to determine the level of *V. parahaemolyticus* present; in 2014 MA-DMF collected 36 samples for this purpose but the test results are not included in their 2014 annual report.[18]

Section Summary. As explained above, there is under-reporting of illnesses associated with marine sources. This reality of limited epidemiological information for multiple causative disease agents is unlikely to change because the cost of definitive diagnostic testing is too high for use in every case of illness and many cases do not come

to the attention of medical providers in the first place. We suggest that a new approach to multiple marine-sourced risk prediction is needed because existing public health measures do not account for the full suite of risks that exist in recreational or shellfish-harvesting waters of Massachusetts Bay. Recreational sea bathing and raw shellfish consumptions are popular pastimes for millions of Massachusetts residents and visitors, yet each carries a set of risks which should be recognized and minimized.

Multiple factors interact to influence environmental conditions in Massachusetts Bay. The Massachusetts Bay environment is part of a larger system that includes both the larger ocean and on-shore areas, especially coastal watersheds which have a close hydrological and human connection. Revealing the factors that influence changes in risk potential requires an understanding of each risk agent and the influences on its abundance. The rest of this chapter is divided into three sections: 1) a physical description of Massachusetts Bay and land use of the adjacent coastal watersheds; 2) estimates of human demographics in Massachusetts Bay coastal watersheds and their potential relation to marine-sourced risk vulnerability, and 3) results of the work that assembled epidemiological background for, and biological information on the factors that influence abundance of, five specific marine-sourced risks in Massachusetts Bay.

Massachusetts Bay and Coastal Watersheds–Characteristics.

Massachusetts Bay is a relatively open temperate bay area along the heavily urbanized Massachusetts coast near Boston, MA at a latitude of 42 degrees North.[41] Massachusetts Bay is connected to the more enclosed Cape Cod Bay to the south, and both bays are part of the larger Gulf of Maine system.[41] Overall circulation and water

properties in Massachusetts and Cape Cod Bay are driven by the Gulf of Maine water flow, but modified by local and regional winds.[41] Seasonal changes in temperature, light, water column mixing, nutrient availability, and large scale ocean processes (e.g., El Niño South Oscillation, North Atlantic Oscillation) contribute to natural variability that can affect marine community composition and phenomenon such as phytoplankton blooms.[41] Ocean-driven environmental influences on Massachusetts Bay interact with land-based environmental influences to create situations that may favor the growth or persistence of multiple marine sourced risks. In the nearshore coastal zone extensive human interaction with the ocean leads to the possibility of exposure to multiple marine-sourced risks.

Massachusetts Bay Coastal Watersheds – Boundaries. Coastal watersheds are the environmentally-relevant unit of analysis for this work. The boundaries for watersheds in eastern Massachusetts are shown below in Figure 2. The six coastal watersheds bordering Massachusetts Bay are labeled as (from North to South) North Coastal, Mystic River, Charles River, Neponset & Weir River, South Coastal, and Cape Cod. These six watersheds and the Bay itself constitute our study area.

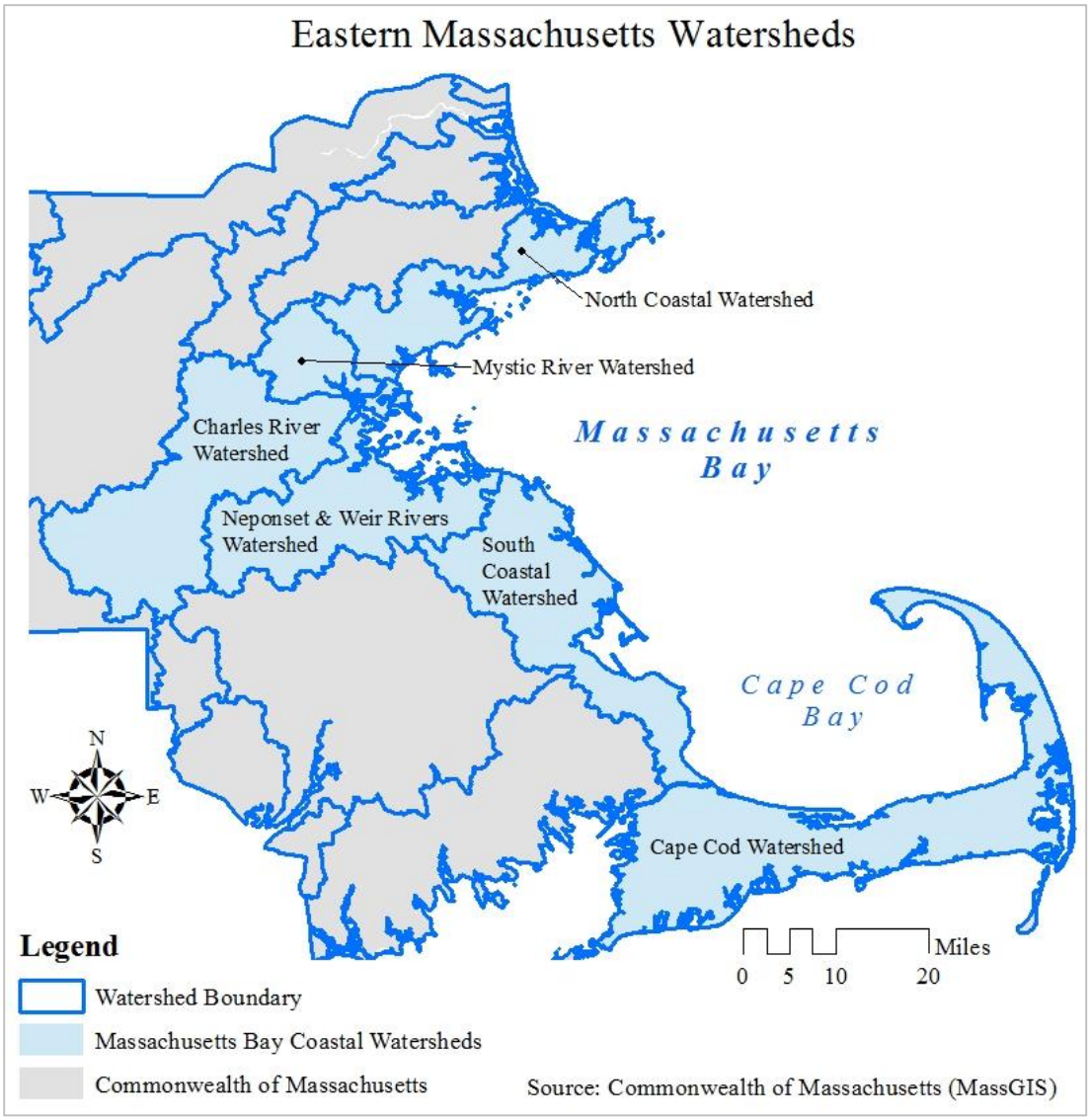

Figure 3-2. Map of Eastern Massachusetts watersheds, the six watersheds bordering Massachusetts Bay are labeled with the names used in this paper. Source: MassGIS; map created by author.

Massachusetts Bay Coastal Watersheds - Land Cover. As a measure of the *state* of coastal environments we can use land-use classification, produced by the Commonwealth of Massachusetts for the year 2005, to estimate the amount of

impervious surface in each watershed. Impervious surface is a proxy for development and levels of surface water runoff. Rain that falls in a watershed with high impervious surface coverage is more likely to run-off quickly into a receiving body of water and carry with it whatever pollutants are on the surface.

The Commonwealth of Massachusetts Office of Geographic Information (MassGIS) has published a statewide land use / land cover database and map file for the year 2005 that identifies 40 different land uses ranging from 'Forest' to 'Junkyard'.[42] Fourteen of these land use categories are associated with low levels of impervious (paved) surfaces: cropland; pasture; forest; non-forested wetland; open land; water; saltwater sandy beach; golf course; cemetery; orchard; nursery; forested wetland; very low density residential; brushland/successional. The area and proportion of each watershed covered in low impervious surface land use types in 2005 is shown below in Table 3-1. The Mystic River watershed has the lowest percent (27%) of land use with low-impervious characteristics. Cape Cod has the highest percent (75%) of land use with low-impervious characteristics. Based on these results we would expect nearshore waters around the Mystic River watershed to receive more pollutants immediately after a rain events than the waters around Cape Cod.

Table 3-1. Area of low-impervious surface land uses by watershed.

Watershed Name	Total watershed area (meters^2)	Area in low-impervious uses (meters^2)	Percent watershed in low-impervious uses
North Coastal	520,056,865	264,974,823	51
Mystic River	261,243,078	70,186,097	27
Charles River	1,095,069,395	660,615,535	60
Neponset & Weir Rivers	737,930,784	418,094,872	57
South Coastal	748,189,075	528,571,804	71
Cape Cod	1,502,298,729	1,131,517,666	75
TOTAL	4,864,787,927	3,073,960,798	63
Low-impervious land use categories: cropland; pasture; forest; non-forested wetland; open land; water; saltwater sandy beach; golf course; cemetery; orchard; nursery; forested wetland; very low density residential; brushland/successional			
Calculations based on "Land Use 2005" and "Major Basins" shapefiles, MassGIS			

Land cover is an important descriptive element because not all human settlements are structured in the same way. Are that include heavy industry might have low resident population but be at higher risk of chemical spills, as opposed to agricultural areas that might have problems with non-point source nutrient loading. The combination of land-use type and human population data is more informative than either element on its own.

Massachusetts Bay – Human Demographics.

Human population is a *pressure* that influences the *state* of the local environment, but not always to the same degree. Human activity can lead to land use changes that affect hydrologic flows and run-off patterns, increased nutrient releases from agriculture or human wastewater, and the direct introduction of microbes from humans as

they physically interact with coastal waters. Understanding these 'upstream' influences on coastal waters is important even though some regional scale changes may happen slowly. We can measure human population in multiple points in time from national census data.

Human demographics and socio-economic factors are relevant to environmental health for four main reasons, 1) human population density is an indicator of multiple types of environmental impacts; 2) human population density can affect disease transmission or pathogen release into the environment; 3) the age structure of a population can influence the population vulnerability to infectious diseases; 4) the wealth of a community influences its access to health resources and overall vulnerability. Therefore, in order to understand marine-sourced risk potential in Massachusetts Bay we must first identify the population most likely to be exposed to these risks. For this paper the residents of coastal watersheds around Massachusetts Bay are the population of interest. The demographics of coastal populations matter because vulnerability to infectious diseases and environmental toxins changes with age. For example, children under 5 years of age (with less developed immune systems) and adults over 65 years of age (with age-related weakening of the immune system) are considered more vulnerable to developing complications from infectious diseases.[43] The presence of pre-existing health problems which may increase vulnerability to environmental pathogens (e.g., immunosuppression to due cancer therapy) is generally higher in older populations. When considering public health monitoring and notification programs for recreational waters it is important to consider frequent beach-goers (likely local residents) as well as

the most vulnerable visitor populations. If the demographics of the resident population in a coastal watershed are changing, so too is the risk profile of the population in that watershed. The next section provides an estimate of the population in the six coastal watersheds around Massachusetts Bay.

Massachusetts Bay Coastal Watersheds – Demographic Estimates. Coastal watersheds are the environmentally-relevant unit of analysis for this work. The boundaries for watersheds in eastern Massachusetts are shown above in Figure 3-2. Towns that do not border the ocean can still be part of a coastal watershed and exert an influence on coastal ocean conditions. Town boundaries may also cross watershed boundaries making them an imperfect unit of analysis for watershed population estimates. However, the U.S. Census Bureau measures population at spatial scales smaller than the town level (e.g., census tract), allowing for a more nuanced spatial analysis. Census tracts cover the entire U.S., providing the potential to apply this method in other places and at varying scales of analysis. The following sections describes available data products from the U.S. Census Bureau, their relevance to this research, and our method for arriving at population estimates for each of the six coastal watersheds bordering Massachusetts Bay.

U.S. Census Data Sources for Demographic Estimates. Demographic trends across the nation are documented by the U.S. Census Bureau through products such as the decadal census and the American Community Survey (an annual survey). [44-46] There are multiple units of spatial analysis used by the Census Bureau for population counting purposes, including census blocks, block groups, and tracts. [47] Blocks are the smallest counting unit used by the Census. A census block is an area "bounded by visible features

such as streets, roads, streams, and railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county limits." [48] Census block groups are clusters of census blocks within the same census tract, and each tract contains at least one block group. [49] Census blocks are a smaller spatial unit than tracts, but in the publicly released datasets the block-level files only contain population and housing unit counts. Census tracts are small subdivisions of a county, usually covering a contiguous area, with a general population size between 1,200 and 8,000 people.[47] Census tract boundaries are lined up with stable government boundaries (such as town or county boundaries) to allow for tract-to-governmental-unit analysis. Census tracts have a larger populations and spatial area than blocks, but each tract record contains a richer set of demographic details. [50] These details include population counts in various categories such as racial groups, males and females, age groups in 5-year blocks from 'age 5 and under' to 'age 85 and over', median age of males and females, household size, information on housing units (total number, number of vacancies, owner-occupied, and renter-occupied units), and median income. [50] These and other demographic characteristics can be used to estimate social vulnerability to various hazards.[51] The value of the greater information available at the tract level outweighs the slight increase in spatial accuracy of population estimates at the block level for the purposes of this research. Notable difference between the tabulation units used by the U.S. Census are summarized below in Table 3-2.

Table 3-2. Characteristics of U.S. Census tabulation units

Characteristics of U.S. Census tabulation units			
	Block	**Block Group**	**Tract**
Population Size	0 – 600	600 – 3,000	1,200 – 8,000 (optimum = 4,000)
Boundaries	Visible features (streets, roads, streams); non-visible features (property lines, city limits)	Visible and non-visible features. Block groups never cross state, county, or census tract boundaries.	Visible and identifiable features, nonvisible legal boundaries. Never cross state or county lines.
Relation to other census units	Smallest census unit, nests within all other census units.	Usually covers area of contiguous census blocks.	Contains at least one block group.
Frequency of change	Most responsive to development.	Can be split in response to growth.	Relatively permanent, designed to be stable.
Types of data	Population count, housing count	Population count, housing count	Population count, housing count, income, age, sex, and others

Massachusetts Bay Coastal Watersheds - Demographic Estimate. The irregular shape of census blocks and tracts makes a visual assignment to a watershed subject to human error. Some census units cross watershed boundaries and fall in two coastal watersheds, and some census units contain areas within both coastal and non-coastal watersheds. Generating a reasonable estimate of the human population within a watershed required assigning the population within a census tabulation area to a specific watershed using spatial analysis software. The software used for all spatial analysis in

the project was ArcMap 10.1 (ESRI, Redlands, CA).[52] Within the spatial analysis software is it possible to locate the centroid point of a polygon feature such as a census tract. It is then possible to 'assign' a polygon to another polygon feature, such as a larger watershed polygon, based on the centroid of the smaller polygon. This centroid assignment method has been used in at least one other study in the northeastern U.S.[53] A map showing the total area of census tracts assigned to each watershed is shown below in Figure 3-3.

In some places the census tract extends outside of the watershed boundary, in other places there are areas within a watershed where the population is not counted because the tract's centroid was outside the watershed. This is most notable for the census tracts in the South Coastal watershed, depicted in pink in Figure 3-3. This paper combined census block and census tract spatial extract data products from the U.S. Census combined with Massachusetts Bay coastal watershed boundaries from MassGIS to estimate the human population in six coastal watersheds.[50; 54] The population estimate results from this spatial analysis method are shown in Table 3-3 below. For the year 2010 there is a 0.28% difference between the two estimates for total population in all watersheds, with the estimate based on census tracts slightly lower (2,924,701 people) than the estimate based on census blocks (2,932,958 people). The largest difference in estimates is for the South Coastal watershed, where the tract-based estimate is 2.6% lower than the block-based estimate. Given the similarity of the results from these two methods, and the greater information associated with tract-level datasets, this paper uses tracts as the population unit of analysis.

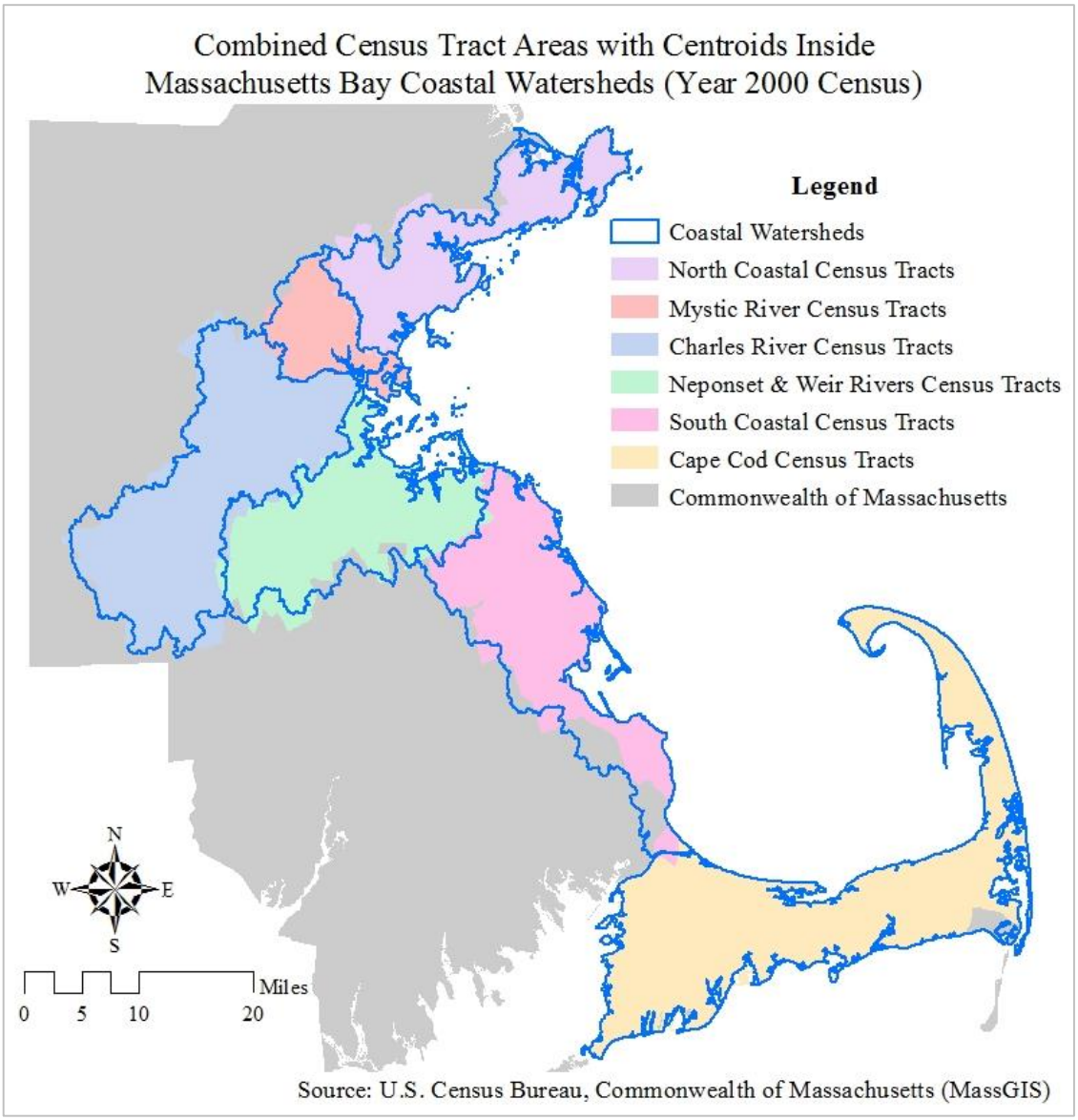

Figure 3-3. Total area of census tracts with centroids inside of each Massachusetts Bay coastal watershed. Note: Boundaries between census tracts removed for clarity. Source: U.S. Census Bureau, MassGIS, map by author.

Table 3-3. Results of centroid assignment method for census blocks and tracts within Massachusetts Bay coastal watersheds. Calculations by author.

Watershed Name	2010 Population estimate based on **census tract** centroid assignment	2010 Population estimate based on **census block** centroid assignment	Percent difference* (rounded) between population estimates ((tract-blocks)/ tract)*100.	Number of **census tracts** with centroid inside watershed	Number of **census blocks** with centroid in watershed
North Coastal	467,244	461,040	1.3	99	5,890
Mystic River	498,657	507,428	-1.8	116	5,614
Charles River	940,948	934,470	0.7	228	9,910
Neponset & Weir Rivers	627,971	635,638	-1.2	143	7,639
South Coastal	183,744	188,415	-2.6	34	2,922
Cape Cod	206,137	205,967	0.1	54	11,789**
TOTAL	**2,924,701**	**2,932,958**	-0.28	680	43,764

**Percent difference is rounded to nearest tenth.*
***Note: Over 4,900 census blocks in Cape Cod have a 2010 population of '0' because most, or all, of the spatial area within the block is water.*

As shown in Table 3-3, there is little difference in the population estimate for each of the six coastal watersheds when using either census blocks or tracts as the unit of analysis. Because of the close agreement in population estimates we use census tract data as it contains both socioeconomic factors and population estimates. Table 3-3 shows that the Charles River watershed contains the highest total population (940,948 people) and Cape Cod the lowest (206,137 people). As described at the start of this section, people ages 5 and under ($\leq$5), or 65 and over (65+) are considered the two most immunologically vulnerable age groups, year 2010 population estimates for these groups are shown in Table 3-4. There is little difference in the percentage of residents age 65+ between most of the Massachusetts Bay watersheds, but Cape Cod stands out for having the largest

percentage (25%) of residents age 65+. As of the 2010 census, the Cape Cod watershed had the highest percentage of residents aged 65+, and the lowest percentage and lowest total number of residents aged $\leq$5. Table 3-5 shows the average median household income of the six coastal watersheds, for which Cape Cod had the lowest (US$60,307) and South Coastal the highest (US$85,832). Median household income is a general indicator of social vulnerability,[51] a higher income implies greater access to resources, including medical care. The combination of these factors, a high percentage of elderly residents and the lowest averaged median income, suggests that of all six watersheds residents of the Cape Cod watershed are the most socially vulnerable. Cape Cod is also notable because despite a resident population that shrank between 2000 and 2010 it is a highly popular tourist destination with over 5 million tourist visits every year[55] and Cape Cod towns have large tracts of active shellfish harvesting areas (see Figure 3-7). Many tourists visit beaches and consume local seafood during the course of their visit.[55] Seafood consumption and sea bathing are not limited to tourists visiting Cape Cod, but are popular activities for residents and visitors all around the Massachusetts Bay area.

Table 3-4. Select demographic characteristics for Massachusetts Bay coastal watersheds in 2010. Data source: U.S. Census; calculations by author.

Watershed Name	Average Median Age*	Number Residents Age 65+	Percent Residents Age 65+	Number Residents Age ≤ 5	Percent Residents Age ≤ 5
North Coastal	40.7	72,146	15	26,881	6
Mystic River	37.3	65,021	13	30,864	6
Charles River	35.4	110,815	12	47,030	5
Neponset & Weir Rivers	39.0	87,393	14	35,973	6
South Coastal	42.7	27,059	15	10,232	6
Cape Cod	51.1	52,371	25	8,441	4
*Average median age = average of 'median age' for all census tracts assigned to a watershed					

Table 3-5. Massachusetts Bay coastal watersheds, average of median household incomes for all 2010 census tracts assigned to watershed. Data source: U.S. Census; calculations by author.

Watershed Name	Average Median Household Income (average of all tracts in watershed, US$)
North Coastal	63,497
Mystic River	67,917
Charles River	75,643
Neponset & Weir Rivers	63,470
South Coastal	85,832
Cape Cod	60,307

Massachusetts Bay Coastal Watersheds – Demographic changes from 2000 to 2010. The 2010 Census reports the total population of Massachusetts as 6,547,629 people; Massachusetts contained approximately 2% of the national population in 2010.[56] The six coastal watersheds around Massachusetts Bay contained approximately 44% of the state population in 2010. According to the U.S. Census Bureau between the years 2000 and 2010 the population of Massachusetts increased by 3.13% from 6,349,097 to 6,547,629 people. Table 3-6, below, shows that although growth from 2000 to 2010 was

not evenly distributed among the six coastal watersheds there was a slight population increase (approximately 2.5%) in the six study watersheds, this parallels the population change of the state as a whole.

Table 3-6. Estimated total population change in Massachusetts Bay coastal watersheds from 2000 to 2010 based on census tracts. Data source: U.S. Census, and Commonwealth of Massachusetts, calculations by author

Watershed Name	2000 Population	2010 Population	Percent change from 2000 to 2010
North Coastal	458,843	467,244	1.8
Mystic River	489,480	498,657	1.9
Charles River	910,286	940,948	3.4
Neponset & Weir Rivers	606,107	627,971	3.6
South Coastal	174,392	183,744	5.4
Cape Cod	213,414	206,137	-3.4
Total	**2,852,522**	**2,924,701**	**2.5**

As shown in Table 3-7, the percent of the population age 65+ changed the most in the South Coastal (3% increase) and Cape Cod watersheds (4% increase) between the 2000 and 2010 censuses. Cape Cod had the highest percentage of residents age 65+ in both 2000 and 2010. This suggests that the overall burden of diseases associated with aging, including susceptibility to infectious disease, is highest in Cape Cod.

Table 3-7. Change in number and percentage of residents age 65+ between 2000 and 2010 in Massachusetts Bay coastal watersheds.

Watershed Name	2000 Census Number Residents Age 65+	2000 Census Percent Residents Age 65+	2010 Census Number Residents Age 65+	2010 Census Percent Residents Age 65+	Change in Percent Residents Age 65+ from 2000 to 2010
North Coastal	72,619	16	72,146	15	1%
Mystic River	69,699	14	65,021	13	-1%
Charles River	113,160	12	110,815	12	no change
Neponset & Weir Rivers	77,666	13	87,393	14	1%
South Coastal	20,744	12	27,059	15	3%
Cape Cod	44,648	21	52,371	25	4%

Total population change between the years 2000 and 2010 at the census tract level for all census tracts in the six coastal watersheds is shown on the map in Figure 3-5, below. Figure 3-5 shows that there are pockets of population increase in each watershed. Noticeable growth (shown in orange-red tones) occurred in a few tracts scattered across all watersheds except Cape Cod.

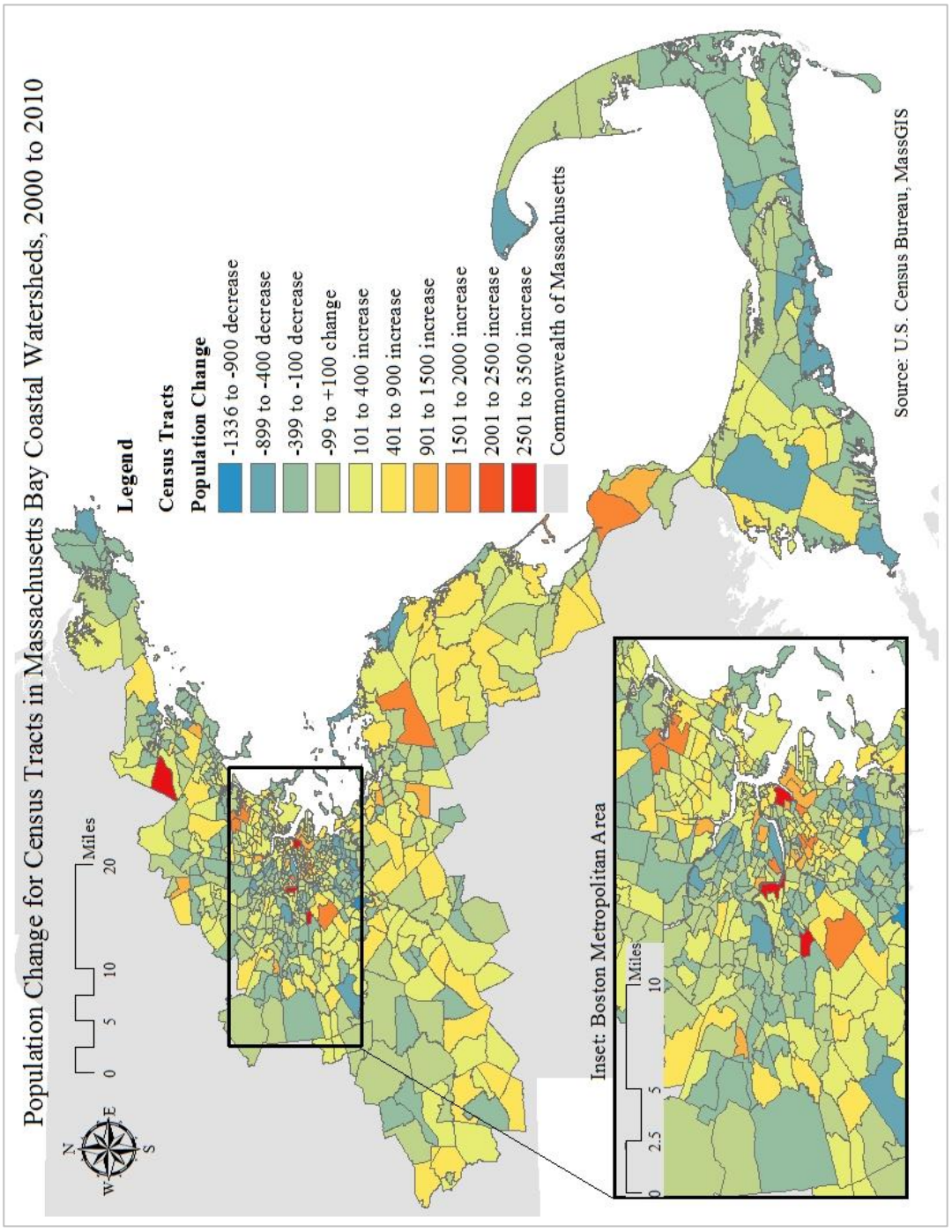

Figure 3-4. Population change for census tract within Massachusetts Bay coastal watersheds, 2000 to 2010. Data source: U.S. Census Bureau, Commonwealth of Massachusetts (MassGIS). Map by author.

Although there were differences in growth between watersheds, overall there was low population growth (in absolute numbers) from 2000 to 2010 around Massachusetts Bay. This suggests that overall human *pressure* on the marine environment was fairly stable during that decade. However, one notable aspect of the human *pressure* on the nearshore marine environment that changed during this decade was the opening of the Deer Island Wastewater Treatment Plant in September 2000, a wastewater treatment facility operated by the Massachusetts Water Resources Authority (MWRA). This changed the flow output of millions of gallons of wastewater effluent (sourced from many towns in the Boston metropolitan area) from being released with minimal treatment into outer Boston Harbor to receiving a higher level of treatment, and then later being discharged 9 miles offshore into Massachusetts Bay after the construction of the outfall pipe.

The areas served by the MWRA sewerage system are shown below in Figure 3-5. The MWRA has conducted extensive biological and chemical monitoring of Boston Harbor and Massachusetts Bay since 1992,[57-61] relevant details of that work will be discussed later. At this stage the important point to note is that since the opening of the Deer Island Wastewater Treatment Plant the levels of nutrients released in Boston Harbor have decreased significantly.[61] So in this area there has been a reduction in one form of human-associated *pressure* on the marine environment *state* despite an overall increase in human population during the same time period.

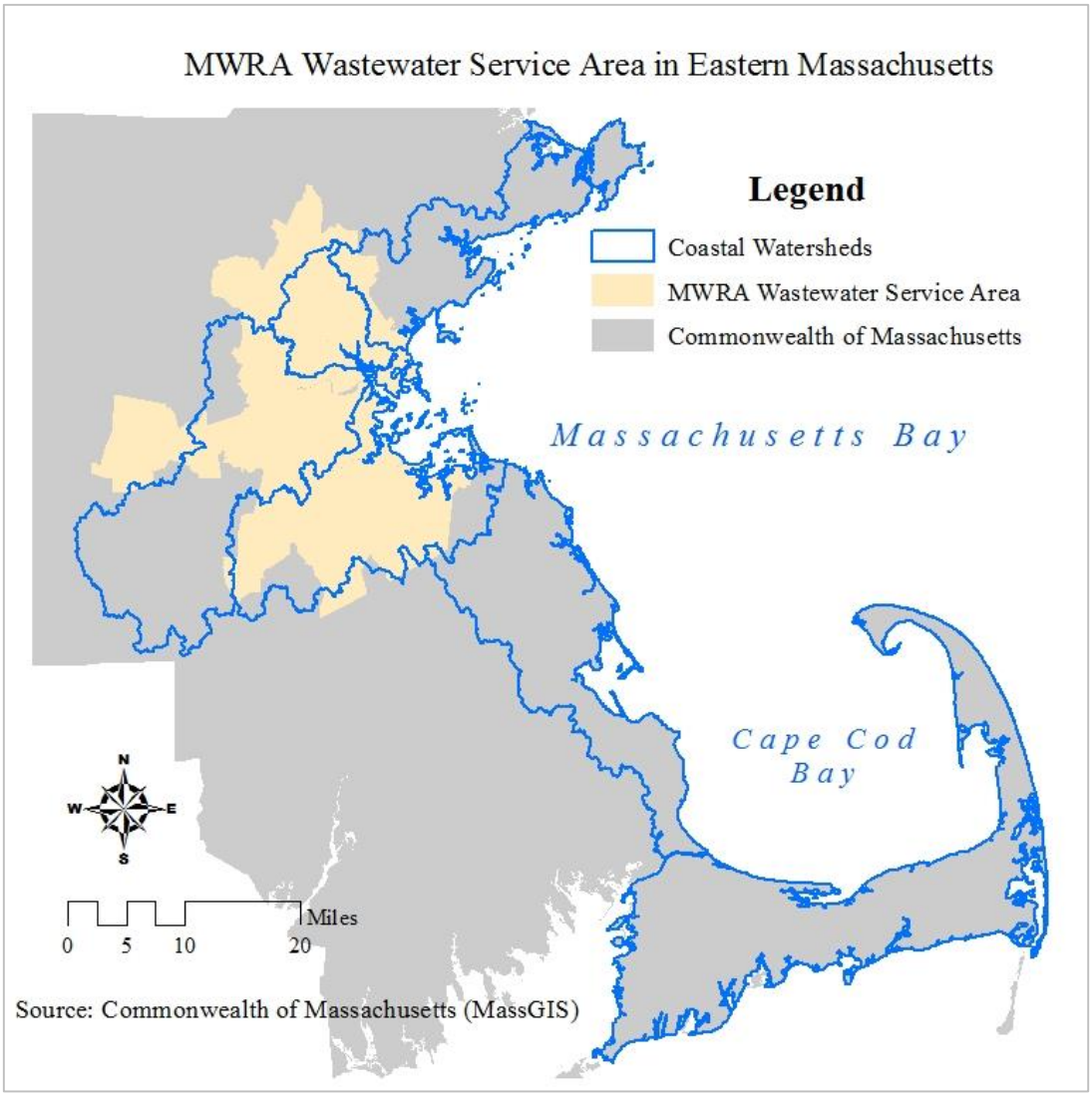

Figure 3-5. Massachusetts Water Resources Authority wastewater services areas in eastern Massachusetts. Source: Commonwealth of Massachusetts (MassGIS)

Figure 3-5, above, shows that there are many areas within these coastal watersheds that rely on non-MWRA wastewater treatment providers, including private septic systems. Non-MWRA, and non-private-residential, groundwater discharge permits are shown as maroon circles in Figure 3-6 below. These discharge permits include facilities such as laundries, car washes, sanitary sewers, and wastewater treatment plants.[62] As shown in Figure 6 there are numerous discharge points within the coastal

watersheds around Massachusetts Bay. Many of these discharge points are within close proximity to bathing beaches (represented as blue lines on Figure 3-6). Figure 3-6 also shows designated shellfish growing areas (open areas in green, closed areas in red), some of which are in close proximity to groundwater discharge permit locations. Pathogens passing through these discharge points that were not inactivated by either publicly-owned or private wastewater treatment facilities are released into the Bay or its freshwater tributaries, carried by various effluent flows or rainwater runoff travelling through sanitary sewers. To clarify, these groundwater discharge points represent some, but not all, conduits through which human pathogens may be introduced in the nearshore coastal environment. The abundance of these points, along with unmapped private septic tank leech fields, suggests that human pathogens can potentially be re-introduced into the nearshore marine environment through wastewater flows, presenting an un-quantified risk for people who use these waters for recreation and shellfish harvesting.

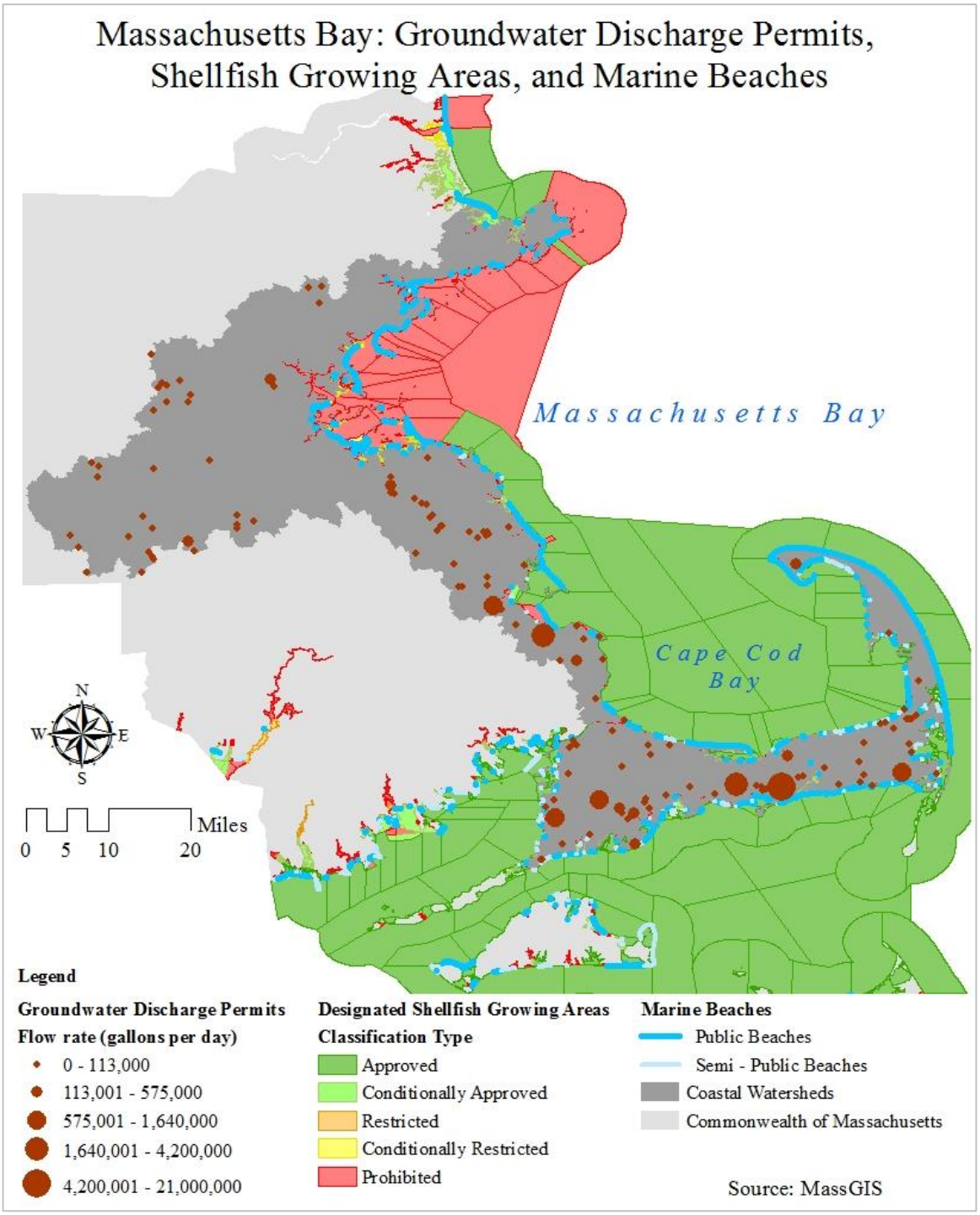

Figure 3-6. Marine beaches, groundwater discharge permit locations, and designated shellfish growing areas (as of June 2014). Source: MassGIS[62-64], map by author.

Nearshore Wastewater Releases in Massachusetts. Some municipalities have antiquated wastewater infrastructure that includes combined sewer overflows (CSOs), which combine storm sewers that collect surface runoff and sanitary sewers that collect residential and commercial wastewater. High rainfall events can cause CSOs (or their destination WWTP) to overflow and discharge raw sewage. The majority of cities with CSOs are found in northeastern and industrial Midwestern states, including Massachusetts.[65] Despite the progress made in reducing untreated wastewater releases into Boston Harbor through the creation and operation of the Deer Island Wastewater Treatment Plant, other sources of raw sewage discharge into Massachusetts Bay exist. These include direct discharges into the Bay, or discharges into rivers that empty in to Massachusetts Bay.[66; 67] The city of Lynn, M.A. is one municipality where raw sewage discharges are permitted during high rainfall events. Residents of nearby Revere, M.A. claim that the raw sewage discharges from Lynn move downstream and end up in the coastal waters off of Revere.[68] Upstream inputs and associated downstream impacts are one reason why cross-boundary water pollution issues can persist for so long-- those who pay for cleanup might not see the benefits retained in their community.

Raw sewage releases associated with high rainfall events are not the only source of untreated wastewater. Septic tanks can contribute to poor water quality and increased viral concentrations when they are located near coastal waters.[69] Some parts of eastern Massachusetts have exceptionally high rates of households with septic systems.[70] Depending on their age and efficacy septic systems can slowly leach nutrients into

groundwater, in addition they can serve as point sources for the introduction of enteric pathogens into groundwater.

The potential for less frequent, but more severe, rain storm events associated with a changing climate could overwhelm what have historically been considered adequate water handling systems. "Enhanced loading of fecal indicators and pathogens is influenced by wet weather events which overwhelm wastewater treatment plants, saturate soils (decreasing the efficiency of septic system drainfields), and result in direct runoff or groundwater base flow from urban and rural areas. Lastly, resuspension associated with storms or bioturbation may consequently reintroduce sediment-associated pathogens into surface waters."[71] The severity of pathogen loading from a wet weather event will be influenced by existing infrastructure, impervious surfaces (which allow for pollutants and animal waste to be flushed into local waterbodies), and absorptive capacity of the soil, leading to local differences within the same watershed. In urban Boston for example, the amount of impervious surface is unlikely to significantly increase because the city is already heavily developed. In fact, there are plans to reduce the amount of impervious surfaces in certain areas to improve soil absorption and flood control capacity.[72] Other coastal watersheds around Massachusetts Bay have lower levels of impervious surfaces, and thus a greater potential for an increase in the area covered by impervious surfaces. Rainfall runoff and nutrient releases through wastewater represent two important pathways for nutrients and pollutants to reach coastal waters. Underlying geology varies across the six coastal watersheds, with Cape Cod being notable for a high amount of permeable sediments that allow for groundwater flow across the peninsula.[70]

Development is not evenly distributed across Cape Cod. For example, residential parcels around Waquoit Bay increased by approximately 15-fold in the years 1940-1989, and local development practices led to the area being both heavily populated and largely unsewered with most homes relying upon septic systems of varying age and efficacy.[70] These septic system outflows could mix with marine waters at numerous locations because of Cape Cod's permeable sediments.

The combination of WWTP releases, nutrient loading from surface runoff or septic system releases, direct bather shedding, and natural organismal population variability paints a complex picture of the possible marine-sourced risk environment in nearshore coastal waters. At present there is little *in situ* direct real-time monitoring in place to give an accurate picture of the full suite of risks faced by those recreating or harvesting shellfish in the coastal zone. There is a clear need to employ other approaches to assist public health authorities in evaluating and responding to changing environmental risks.

Section Summary. Through shellfish harvesting and nearshore water-based recreation activities people may be exposed to multiple types of marine-sourced risks. These risks can be indigenous or introduced agents that comingle in the nearshore environment, however this is not fully reflected in existing water quality monitoring practices. Currently, *Enterococcus* bacteria are the most widely collected and utilized indicator of marine recreational water quality because they are abundantly, but not exclusively, associated with human waste.[73] Not all pathogens or toxins will behave in the same way as a single group of bacteria. Other potentially marine-sourced risks are

recognized as important by public health authorities in Massachusetts, and they are included on the list of reportable diseases.[38; 39] That list includes the following:

- Any case of an unusual illness thought to have public health implications [requires immediate reporting]
- Any cluster/outbreak of illness, including but not limited to foodborne illness [requires immediate reporting]
- Foodborne illness due to toxins (including mushroom toxins, ciguatera toxins, scombrotoxin, tetrodotoxin, paralytic shellfish toxin and amnesic shellfish toxin, and others) [requires immediate reporting]
- Hepatitis A / Hepatitis A virus [requires immediate reporting]
- Vibriosis / *Vibrio* species [requires reporting within 1-2 business days, isolates must be sent to the State laboratory]

These risks however, are not routinely considered during water quality monitoring. Only recently has *Vibrio parahaemolyticus* been officially recognized as a risk requiring a specific control plan for shellfish harvesting activities, that plan does not include direct sampling for *V. parahaemolyticus* by local authorities but limited sampling has been initiated by MA-DPH.[35] If we are to more fully understand the human health risks in the nearshore coastal environment we must consider multiple risks at the same time. Monitoring or separately sampling for each risk individually is likely to remain too costly to implement, we propose a method to organize these multiple risks by type, and then estimate multiple risk potentials together through modeling.

The next phase of this exercise is to identify the risks of interest, their known or suspected epidemiology, and which factors may influence their presence and abundance. This exercise will allow us to identify which factors have the greatest informational value, and which should therefore be the highest priority for data acquisition when attempting to build a predictive model. As a reminder, for this paper we have chosen the following risks: *Enterococcus* bacteria (because they are the current water quality monitoring standard), *Vibrio parahaemolyticus* bacteria (because illnesses causes by *Vibrio* species are reportable in Massachusetts and *V. parahaemolyticus* is native to New England waters), Hepatitis A Virus (because may be transmitted via contaminated food or water and is a reportable disease in Massachusetts), *Pseudo-nitzschia* genus diatoms (because they may produce Domoic Acid, a type of toxin which may cause foodborne illness), and anthropogenic antibiotics (because they are known to be released in the effluent of wastewater treatment plants and influence the development of antibiotic resistance in bacteria).

Description of Five Marine-Sourced Risks Known to Exist in Massachusetts Bay.

In this section we review the background and known epidemiology for five marine-sourced risks in Massachusetts Bay. Those risks are the enteric bacteria genus Enterococcus, the indigenous marine bacteria Vibrio parahaemolyticus, the enteric virus Hepatitis A Virus, the potentially toxigenic diatom genus Pseudo-nitzschia, and anthropogenic antibiotics which may be released through wastewater discharges.

***Enterococcus* species – Bacteria Associated with Mammalian Feces.**

***Enterococcus* – Background**. The genus *Enterococcus* contains 28 species of bacteria, collectively known as Enterococci.[74] Closely related to the *Streptococcus* genus, Enterococci are essential residents of human and animal digestive tracts. Although Enterococci are beneficial residents of the intestinal tract they can cause illness when introduced to other parts of the body such as the urinary tract or surface wounds. Enterococci resistant to the antibiotic vancomycin, known as vancomycin-resistant enterococci (VRE), have been found in clinical settings as well as in the food system where their presence is linked to the use of antibiotics in animal feed and resulting selective pressure for antibiotic resistance.[74] Environmental exposure to *Enterococcus* is possible through multiple pathways, including recreational beach-going activities.

***Enterococcus* - Human Epidemiological Considerations.** Every year, bathing in coastal waters polluted with fecal contamination is estimated to cause more than 120 million cases of gastrointestinal illness and 50 million cases of respiratory disease around the world.[5] As the mix of pollution sources and environmental characteristics of receiving water varies around the globe, finding a single universally-applicable indicator of recreational water quality has proved challenging. In 1986 the U.S. Environmental Protection Agency recommended that Enterococci be used as the sole indicator for ocean water bacterial monitoring.[27] Indicator bacteria are not necessarily pathogenic, but the *Enterococcus* genus of bacteria is associated with human waste which could contain other pathogens.[73] Due to the complexities of multiple sources of pollution interacting in the

environment, there is a continued interest in developing improved indicators or forecasts of water quality.

Water column bacterial counts are one of the most widely collected biological indicators of water quality. These data sets are not without utility. In fact, a "meta-analysis of twenty-two epidemiological studies conducted from 1953 - 1996 at beaches around the world suggests a causal dose-related relationship between gastrointestinal symptoms and recreational water quality as measured by bacterial indicator counts. Among these studies, *Enterococcus* spp. emerged as the indicator bacteria best correlated with health outcomes in marine systems."[25; 75] However, for over fifteen years experts have been questioning the widespread use of fecal indicator organisms (including *Enterococcus*) as the **main** recreational water quality standard. The weakness of using fecal indicators was summarized in a World Health Organization publication of experts in recreational water quality in 1998 (known as the Annapolis Protocol), illustrated by the following excerpt:

> "Present regulatory schemes for the microbiological quality of recreational water are primarily or exclusively based on percentage compliance with fecal indicator counts … A number of constraints are evident in the current standards and guidelines:
>
> · management actions are retrospective and can only be deployed after human exposure to the hazard;
>
> · the risk to health is primarily from human excreta, the traditional indicators of which may also derive from other sources;

- there is poor inter-laboratory and international comparability of microbiological analytical data; and
- while beaches are classified as safe or unsafe, there is a gradient of increasing severity, variety, and frequency of health effects with increasing sewage pollution and it is desirable to promote incremental improvements prioritizing 'worst failures."[76]

Despite concern expressed by researchers, the status quo persists.

It is worth nothing that the choice of indicator used for water quality monitoring can have economic ramifications. Closed beaches are not good for business, this may result in social pressure to select the least restrictive standard or to schedule sampling at the time most likely to provide favorable results. A Southern California study compared the use of total coliform (TC), fecal coliforms (FC), and Enterococci (EC) standards when determining water quality failures.[27] They predicted that replacing the pre-1999 TC-only standard with an EC-alone standard would lead to a five-fold increase in failures during dry weather, and a doubling of failures during wet weather. The switch to a standard based on all three indicators was predicted to lead to an eight-fold increase in failures and have significant implications for beach closures and restrictions.[27] Increased beach closures, due to more accurate monitoring, could reduce the risk potential for exposure and provide improved public health benefits if the closure notices are heeded. Here, we argue and support with best available data that significant underreporting of bathing beach risk exposure exists and is important both because of the public health costs and concerns but also because of the substantial economic and social benefit costs

associated with lost recreational opportunities. This is a significant point because the healthcare and lost wage costs associated with 'contaminated beach water' have been estimated at US$286 million annually for the U.S.[6]

***Enterococcus* - In the Environment**. Enterococci can be released into the environment from the feces of livestock, domestic birds, wild birds, and they have been found to exist naturally in soil and in association with plants, zooplankton, algae, and marine detritus.[77] *Enterococcus faecalis* and *Enterococcus faecium* are the most common *Enterococcus* species found in human feces but they have also been isolated from livestock.[74] *Enterococcus* species can grow in a wide range of temperatures (5 to 50°C), pH (4.6-9.9), and salt (6.5% NaCl) concentrations. Factors associated with the presence or persistence of *Enterococcus* in coastal recreational waters are shown in Table 3-8. A graphical depiction of the same information is shown in Figure 3-7.

Table 3-8. Known Influences of *Enterococcus* bacteria in coastal bathing areas

Influences of *Enterococcus* bacteria in coastal recreational waters	**Evidence strength**	**Reference**
Salinity (weakly negatively associated)	Medium	76; 78
Sea water temperature at surface (SST) (optimum 42.7°C; range 6.5 to 47.8°C) (weakly associated)	Strong	74; 76
Water pH (optimum 7.5: range 4.6 to 9.9 pH) (weakly associated)	Strong	76
Wind speed, to distribute existing plume (weakly associated)	Low	76
Wind direction (weakly associated)	Low	76
Rainfall (may wash bacteria from land to sea) (positively associated)	Strong	7673; 74; 79
Combined Sewer Overflows (presence, volume) (positively associated)	Strong	76; 80
Riverine discharge to area (positively associated)	Medium	76

Table 3-8. Known Influences of *Enterococcus* bacteria in coastal bathing areas

Influences of *Enterococcus* bacteria in coastal recreational waters	**Evidence strength**	**Reference**
Storm drains (presence, abundance) (positively associated)	Strong	[76]
Turbidity (matter in suspension) (strongly positively associated with survival)	Strong	[76; 77; 80]
Plankton in water (positively associated)	Strong	[76; 78]
Air temperature (as it relates to sea water temperature)	Medium	[76]
Wave height (may release *Enterococcus* from sediments into water column)	Low	[76; 80]
Total light or radiation (higher radiation increases mortality) (strongly negatively associated)	Strong	[76; 80]
Tidal state and magnitude (wetted beach sands may release *Enterococcus* into water column) (positively associated)	Medium	[76; 81]
Bather population at beach (direct shedding) (positively associated)	Strong	[21; 76; 82; 83]
Animal population, presence of horses, donkeys, dogs, shore birds (recommend hourly observation) (associated with higher levels)	Strong	[76; 78]
Boats anchored or moored within 1 km of beach	Weak	[76]
Beach debris and sanitation: sanitary plastics, visible grease balls, algae (recommend daily observation)	Medium	[76; 78]
Location of bather facilities (showers, lavatories) and relevance of input from these sources to beach	Weak	[76]
Release of bacteria from beach sand 'reservoir', including seaweed wrack on beaches (positively associated)	Medium	[78; 82-85]

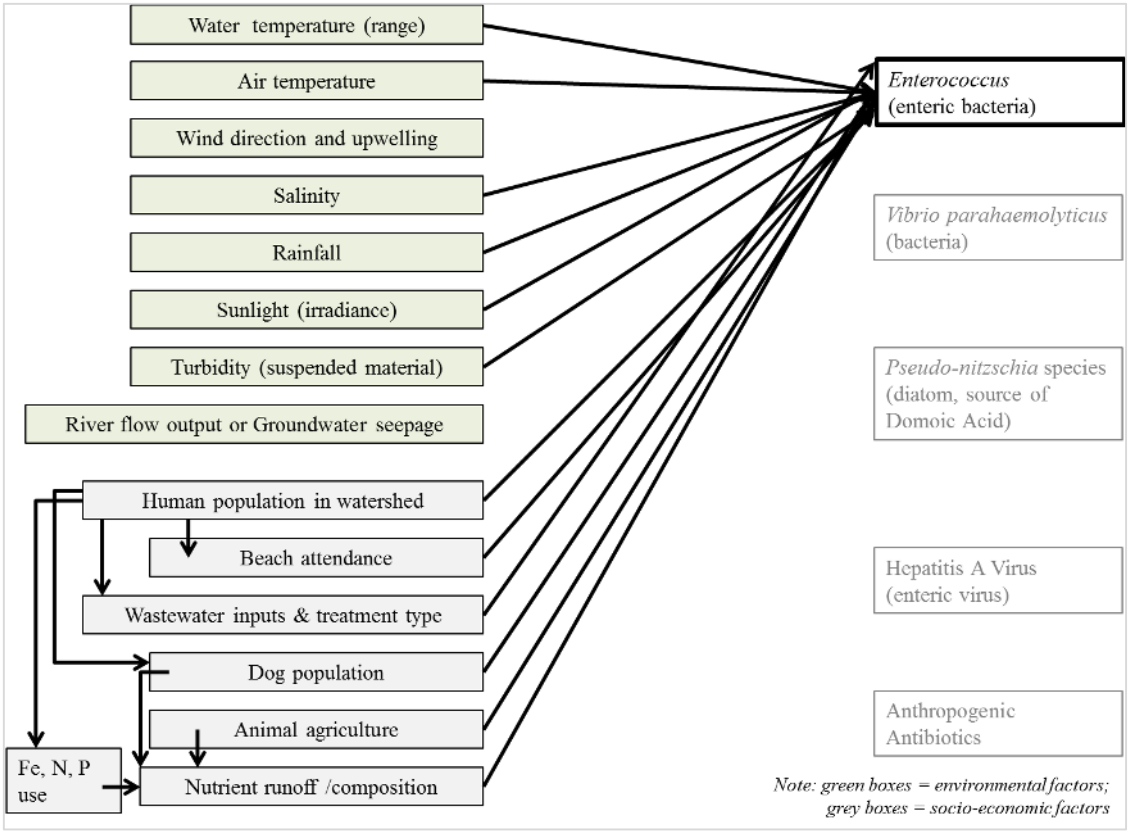

Figure 3-7. Graphical representation of known influences on *Enterococcus* population levels. Lines between influences (left side of figure) indicate interactions between influences.

***Enterococcus* – In the Massachusetts Bay Area**. In Massachusetts *Enterococcus* infections are not a reportable disease unless associated with an usual outbreak or unusual illness. [38; 39] In Massachusetts, the Department of Public Health (MA-DPH) publishes annual reports documenting the results of recreational water quality testing. The standard for marine recreational waters is 104 colony forming units (CFU)/100 mL seawater, samples with counts above 104 CFU/100mL are classified as exceedances. Despite decades of attention, exceedances of *Enterococcus* counts continue to occur at marine beaches in the state. Table 3-9, below, shows total annual water quality exceedances at

marine beaches from 2001 to 2014.[26] Number of exceedances vs. number of samples analyzed per year is plotted in Figure 3-8, below. The lowest percentage of exceedances was recorded in 2002 (2.8% of samples tested), that year also had the lowest number of total samples analyzed (6686). The highest percentage of exceedances (7%) occurred in 2009, the year with the third highest number of samples analyzed (8119).

Table 3-9. Number of samples for which *Enterococcus* concentrations exceeded water quality criterion at public and semi-public marine bathing beaches, 2001-2014.

Adapted from Table 6 in *Marine and Freshwater Beach Testing in Massachusetts Annual Report: 2014 Season*[26]

Year	Number of Exceedances[1]	Total Number of Samples Analyzed	Percent Sample Exceedances (%)
2001	444	7200	6.2
2002	185	6686	2.8
2003	320	7439	4.3
2004	337	7873	4.3
2005	358	8064	4.6
2006	405	8367	4.8
2007	253	7693	3.3
2008	433	7639	5.7
2009	571	8119	7.0
2010	490	7919	6.2
2011	481	8140	5.9
2012	343	8006	4.3
2013	475	8132	5.8
2014	329	7516	4.4
Average	388	7771	5.0

1. For marine beaches *Enterococcus* is the indicator species. A sample is in exceedance if the number of colony forming units (CFU) / 100 mL is greater than 104.

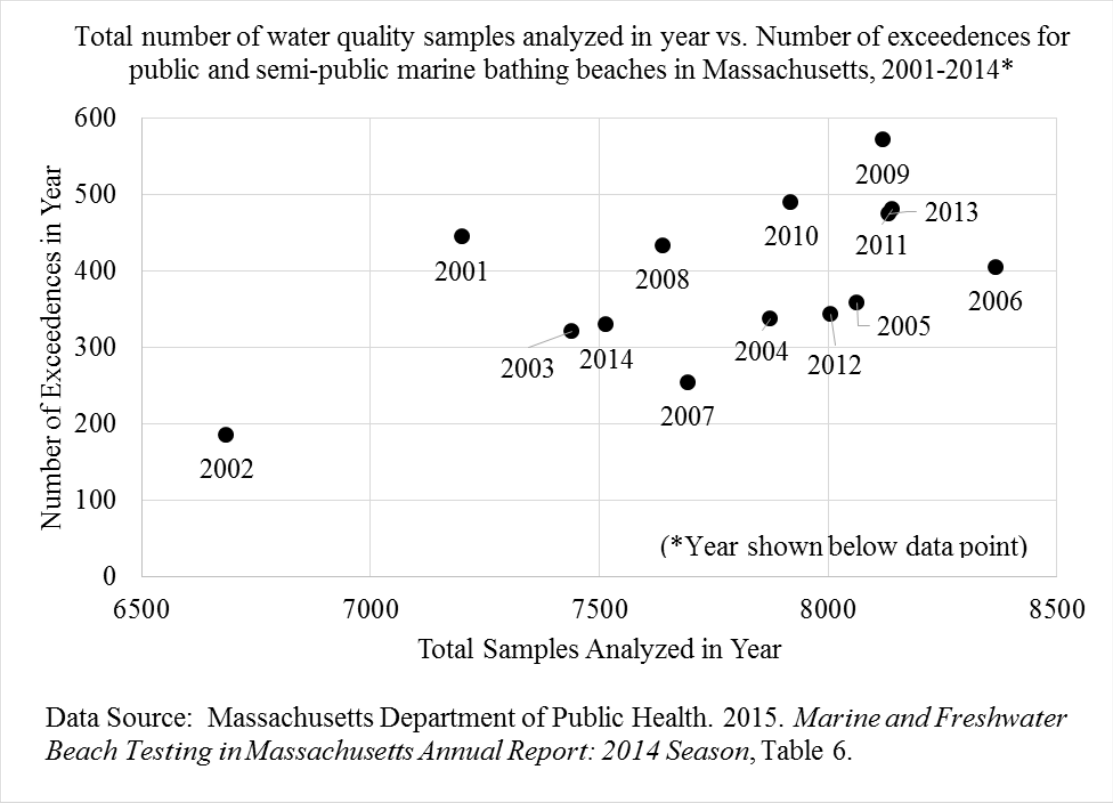

Figure 3-8. Number of water quality samples analyzed vs. Number of exceedances for public and semi-public marine bathing beaches in Massachusetts, 2001-2014. Data Source: Commonwealth of Massachusetts.

In the 2014 Annual Report on beach water quality testing, the MA-DPH noted that bacterial exceedances at marine beaches are closely tied to rainfall events as shown in Table 3-10, below.[26] However, the strength of this relationship appears to vary by location, time of year, and probably other factors. For example, in the Boston area the month of August 2014 had the lowest total rainfall of the three summer months (June 2.62 inches; July 4.57 inches; August 1.75 inches), but the highest percentages of samples that exceeded water quality standards for beaches.[26] The reason for this outcome is unclear, but it suggests that other factors besides rainfall influence *Enterococcus*

sampling results. Research has shown that *Enterococcus* may persist in the environment in association with soil, plants, suspended particulates, and in beach sands.[24; 77; 85] These environmental reservoirs, along with *Enterococcus* releases from local animals, may contribute to sample exceedances in the absence of rainfall events.

Table 3-10. Water quality exceedances reported based on the number of days since last rainfall at public and semi-public marine bathing beaches in Massachusetts, 2014 bathing season.

Adapted from Table 15 in *Marine and Freshwater Beach Testing in Massachusetts Annual Report: 2014 Season*[26]

Number of Days Since Rainfall Events	**Number of Exceedances**	**Percentage (%)**
0	167	60.3%
1	17	6.1%
2	20	7.2%
3	18	6.5%
4	34	12.3%
5	18	6.5%
6	0	0.0%
7	1	0.4%
8	0	0.0%
9	2	0.7%
10	0	0.0%
10+	0	0.0%
Total	277*	100.0%

*Out of 329 bacterial exceedances. Fifty two exceedances had no corresponding rainfall information.

Massachusetts marine beaches are divided into three tiers according to the historical pollution severity, these tiers determine the current monitoring schedule.[26] Tier One includes heavily used beaches which have pollution problems, they are generally sampled sub-weekly, the five Tier One beaches in Massachusetts are tested daily during the recreational bathing season. [26] Tier Two beaches are higher-use beaches with some

pollution and must be tested one per week, 425 of the 530 marine beaches in Massachusetts are in this category. [26] Tier Three beaches have no known pollution problems and only require testing every two weeks or less (if a variance is granted by the local board of health or MA-DPH), there are 100 marine beaches in this category.[26] The infrequent sampling at Tier Two and Tier Three beaches reduces the likelihood of identifying an exceedance if one were to occur.

As described above, it is not always apparent which source(s) contributes the greatest amount of *Enterococcus* bacteria to coastal waters and if the presence of high levels of *Enterococcus* in water samples indicate contamination by human fecal wastes. Fluctuations of human and animal populations may change non-point-source loading within a watershed. Bathers themselves may re-suspend bottom sediments and subsequently cause elevated levels of Enterococci and other microbes in bathing waters, however this is not likely to be a significant factor for Massachusetts marine beaches. As shown in Table 3-11 below, for 2014 the vast majority of water quality samples were taken when there were between 0 and 10 bathers present at a beach, yet this category included the greatest total number of exceedances (260 out of 329 total for the season). However, when samples were taken at beaches with 50 or more people present, they were more likely to result in an exceedance, almost 13% of these samples were associated with an exceedance. This suggests that sub-daily water sampling and bather counts might reveal bather-induced changes in *Enterococcus* levels at bathing beaches.

Table 3-11. Exceedances grouped by bather density at time of water sample collection for Massachusetts marine beaches in 2014.

Adapted from Table 16 in *Marine and Freshwater Beach Testing in Massachusetts Annual Report: 2014 Season*[26]			
Bather Density (Number of people present at time of sample collection)	Number of samples	Number of exceedances	Percent of samples that resulted in an exceedance
0-10	6,211	260	4.2%
10-20	261	2	0.8%
20-50	95	0	0.0%
>50	62	8	12.9%
Not indicated	887	59	6.7%
Total	7,516	329	4.4%

Section Summary. Traditional water quality testing methods using microbiological indicator organisms can reveal the presence of fecal contamination in coastal waters, but the most widely-used indicator does not distinguish the source of pollution (human or animal).[86] Current testing methods for *Enterococcus* take 24 hours to provide results, making it difficult for investigators to track any contamination back to the source. If efforts to improve bathing water quality (a DPSIR *response* level change) are to be properly targeted at human or animal sources (DPSIR *pressures* that influence the *state*) it is essential to know where the greatest cause for concern lies and which responses have the potential to be successful. For this work we include *Enterococcus* as one of the 5 marine-sourced risks because it is the current standard and can serve as an indicator of degraded water quality. However, because of the limitations described in this section we do not believe it should be the sole biological criterion for water quality.

***Vibrio parahaemolyticus* – An Indigenous Bacteria Species.**

***Vibrio parahaemolyticus* – Background**. *Vibrio* species bacteria exist naturally in the marine environment, they are considered indigenous in many parts of the world. Humans can be exposed to, and potentially infected by, *Vibrio parahaemolyticus* through consumption of contaminated seafood or through direct skin contact. *V. parahaemolyticus* bacteria exposure can cause gastroenteritis and diarrhea, but is rarely fatal.

By 1982 *V. parahaemolyticus* had been found in waters from Madagascar to Alaska (including Australia, Vietnam, China, India, Iran, Russia, Western Europe, Togo, Panama, and Canada).[87] Identifying and enumerating *V. parahaemolyticus* requires microbiological techniques including biochemical profiling or genetic analysis[88] and is more technical challenging and costly than the *Enterococcus* bacterial counts commonly used for recreational- and shellfish harvesting-water quality testing. A 2012 study noted that standard microbial approaches for determining the opening or closing of shellfish harvest areas are still not useful for control of exposure to pathogenic *Vibrio* species and that despite over 30 years of accumulated evidence[89] these approaches continue to be used and are generally accepted in the U.S.[28]

***Vibrio parahaemolyticus* – Human Clinical and Epidemiological Information.**

V. parahaemolyticus bacteria are a major cause of seafood-associated foodborne illness globally.[87; 90] *V. parahaemolyticus* is one of the three most important *Vibrio* species associated with human illness in the United States[91], and has been recognized as a major cause of seafood-associated food poisoning for well over 30 years.[87] The incubation

period for *V. parahaemolyticus* infection is usually 12-72 hours, but can be as long as 1 week[92], limiting the ability to investigate and definitively identify the bacterial vector.

The twin realities of illness underreporting[1] and unidentified causative agents make it difficult to present an accurate picture of the true burden of illness due to *V. parahaemolyticus* infections; however between 1973 and 2006 there were 45 recorded seafood-associated outbreaks of *V. parahaemolyticus* infection in the U.S. resulting in 1,393 documented cases, of which 24 required hospitalization.[90] For all seafood-associated outbreaks in the United States between 1973 and 2006, *V. parahaemolyticus* was responsible for 35% of illnesses with an identified causative agent, the highest percentage attributed to any single species when considering known bacteria, virus, and parasitic species[90]; it should be noted that 80% of all foodborne illness cases in the U.S. are attributed to unknown or unidentified agents.[2] The CDC estimates that in the U.S. for every reported case of *V. parahaemolyticus* foodborne illness there are 142 cases not diagnosed.[93] The CDC also estimates that in the U.S. on average there are 215 culture-confirmed cases, 30 hospitalizations, and 1-2 reported deaths from *V. parahaemolyticus* infections annually.[94] For the entire U.S. a separate study estimated the annual cost of *V. parahaemolyticus* illnesses (based on lost wages, physician and hospital services, and statistical cost of a premature death) to be US$20.63 million, not including any cost of pain and suffering.[6]

V. parahaemolyticus has been established as widely present in products harvested from coastal waters around the U.S., but the majority of *V. parahaemolyticus* cases in the U.S. are reported from the Pacific Northwest despite reports of lower overall levels of

pathogenic *V. parahaemolyticus* in that region.[91] This may be due to increased virulence of *V. parahaemolyticus* infections in that region, increased awareness of the disease among the public, increased awareness of the disease among medical personnel and thus more complete reporting, or a historically higher percentage of the population with health insurance and greater access to medical care.[91; 95] Anecdotal evidence suggests that greater physician awareness plays a role. A 3-year prospective study in British Columbia, Canada recruited 13 people after they arrived for outpatient visits to physicians' offices with gastrointestinal illness and tested them for V. parahaemolyticus infection.[96] Investigators found that although some of the *V. parahaemolyticus* infections demonstrated substantial morbidity, none of the patients required hospitalization, and the specific infectious agent would not have been detected if only hospital laboratory-identified specimens had been included in the study.[96] The reason that these moderate illness would not have been identified is because the standard of care for moderate gastrointestinal upset is typically a course of broad spectrum antibiotics. The results from that 1988 British Columbia study[96] support the argument that *V. parahaemolyticus* infection cases are underreported even when illness is severe enough to seek outpatient medical treatment.

***Vibrio parahaemolyticus* – In The Environment.** In addition to its environmental persistence across a wide variety of temperatures, *V. parahaemolyticus* has one of the shortest generation (reproduction) times of any bacterium (less than 10 minutes), and an optimum growth temperature around 37ºC.[97] All of these factors contribute to the persistence of *V. parahaemolyticus* in the environment. Microbiological studies have

been exploring the influence of local environmental factors on *V. parahaemolyticus* abundance since at least the 1970s, with mixed results.[87; 89; 98; 99] Thompson and Vanderzant (1976) took water, sediment, and oyster samples from Galveston Bay, TX and found no significant relationship between *V. parahaemolyticus* culture counts and multiple abiotic environmental factors.[89] However, in the 1970's Kaneko and Colwell[100; 101] observed a seasonal pattern of *V. parahaemolyticus* presence in the water column of a tributary river of the Chesapeake Bay where surface water temperature ranged from -2°C to 31.2°C.[87; 89]

Early environmental models of *V. parahaemolyticus* abundance were based almost entirely on water temperature and were insufficient to explain the inter-annual variation in pathogen populations.[97] Recent work by Johnson et al. (2012) examined the relationship between *V. parahaemolyticus* abundance and parameters that can be monitored via satellite such as salinity, turbidity, and chlorophyll at three study sites in the Pacific Northwest, Gulf Coast, and mid-Atlantic.[102] The finding that sea surface temperature and suspended particulate matter are good predictors of *V. parahaemolyticus* total abundance at ecologically distinct sites may help inform future seafood safety efforts.[102] Such efforts are still needed, a 2007 survey of market oysters from around the United States isolated *V. parahaemolyticus* from oysters harvested in North Atlantic waters (Connecticut, Maine, Massachusetts, New York, Rhode Island), Mid-Atlantic waters (Delaware, Maryland, New Jersey, South Carolina, Virginia), Gulf Coast waters (Alabama, Florida, Louisiana, Texas), and Pacific Coast waters (Washington).[91]

As a response to the expected gap between traditional indicator bacteria population levels (as reflected in water quality samples) and concurrent *V. parahaemolyticus* population levels, environmental forecasting models for *V. parahaemolyticus* need to be developed further. Table 3-12 presents the results of a literature review for known influences on *V. parahaemolyticus* population levels in the environment.

Table 3-12. Environmental influences on *Vibrio parahaemolyticus* abundance

Environmental influences of *Vibrio parahaemolyticus* in the environment	**Evidence strength**	**Reference**
Water temperature (optimum growth temp ~37°C) Warmer water positively associated with growth, *V. parahaemolyticus* can survive winter water temperatures below 0°C at the surface by associating with sediments and larger animals such as plankton and shellfish.[87]	Very strong	87; 97; 99; 103;102
Salinity (optimum ~23ppt, 10-34 ppt reported to support populations) Higher concentrations of *V. parahaemolyticus* are found in estuarine environments.[87] mixed reports about *Vibrio* relationship with salinity, might depend on effect of other factors.	Mixed	87; 97; 99; 102
Nutrient concentrations	Medium	87
Calcium (Ca) availability (high Ca can increase cytotoxicity to host cells)	Weak	104
Iron (Fe) availability (low Fe can increase virulence)	Weak	104
Presence/abundance of host zooplankton	Weak	4; 87; 97
Turbidity (suspended particulate matter) (strongly positively associated)	Strong	4; 87; 102; 103
Chlorophyll *a* (weakly positively associated)	Medium	102
Dissolved oxygen (positively associated)	Strong	103
Dissolved organic Carbon (DOC) (weakly positively associated)	Medium	102

A graphical representation of environmental and socio-economically linked influences on the growth and persistence of *V. parahaemolyticus* in the environment is shown below in Figure 3-9. As shown in Figure 3-9, most of the known influences on *V. parahaemolyticus* abundance are environmental factors, with nutrient input being the only clear socio-economically-linked factor.

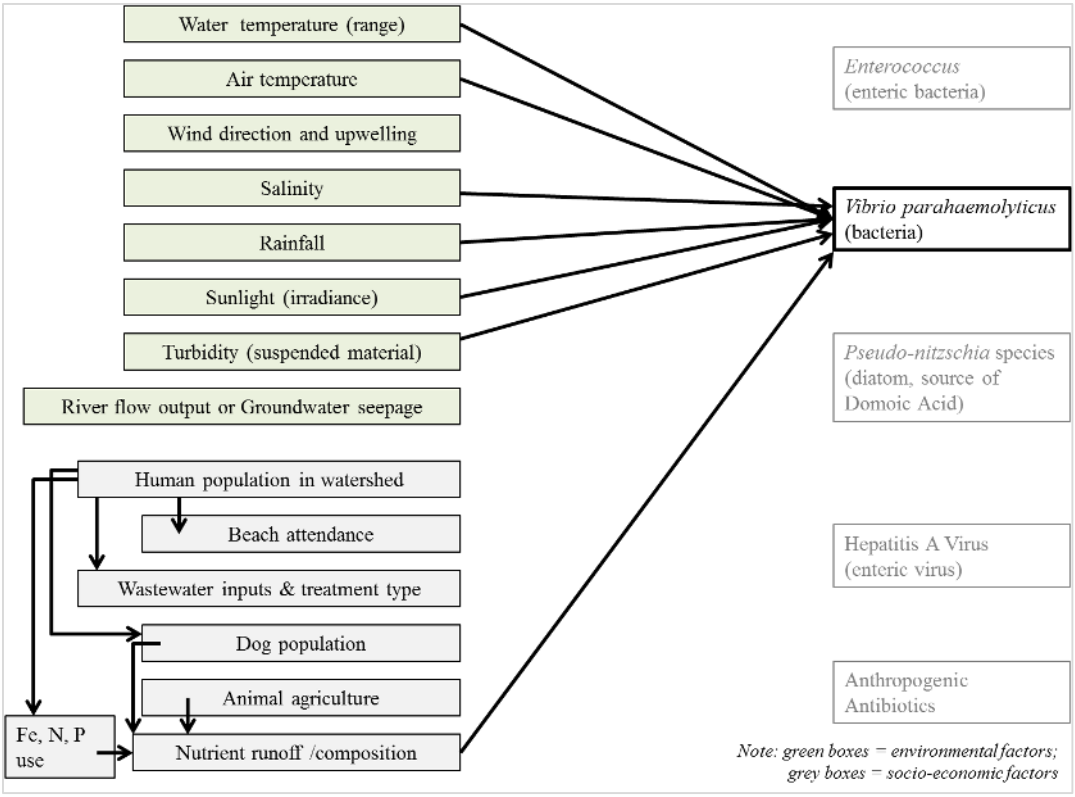

Figure 3-9. Graphical representation of known influences on *Vibrio parahaemolyticus* population levels in coastal waters. Lines between influences (left side of figure) indicate interactions between influences.

***Vibrio parahaemolyticus* – In the Massachusetts Bay Area.** In coastal New England waters *V. parahaemolyticus* is considered an indigenous bacteria, naturally present in the environment.[36; 37] In the United States, risk of *Vibrio* species consumption has traditionally been associated with consumption of raw shellfish harvested from warm water areas such as the Gulf of Mexico and Florida.[90; 105-108] Over 40 years ago scientists showed that *V. parahaemolyticus* inhabits New England waters[37] and can be isolated from seafood such as Cape Cod soft-shell clams.[36] However, it was not until 2011 that cases of *V. parahaemolyticus* infection were officially attributed to raw shellfish commercially harvested in Massachusetts waters.[40] The cases in 2011 were not the first cases of Vibriosis (illness due to *Vibrio* species bacteria) recorded in Massachusetts. As shown in Figure 11, below, Vibriosis have been reported in Massachusetts since at least 1999, but the MA-DPH data do not indicate the source of exposure or the *Vibrio* species. While Massachusetts Bay waters may not reach temperatures of 81ºF (27.2ºC), considered the threshold for stronger shellfish harvest control actions, the tidal cycle can leave harvest areas exposed to warm air for hours at a time, potentially leading to unsafe bacterial counts in seafood.[40] There is now an official Vibrio Control Plan in Massachusetts with a focus on harvesting and transport practices, not on environmental sampling.[35; 40; 109]

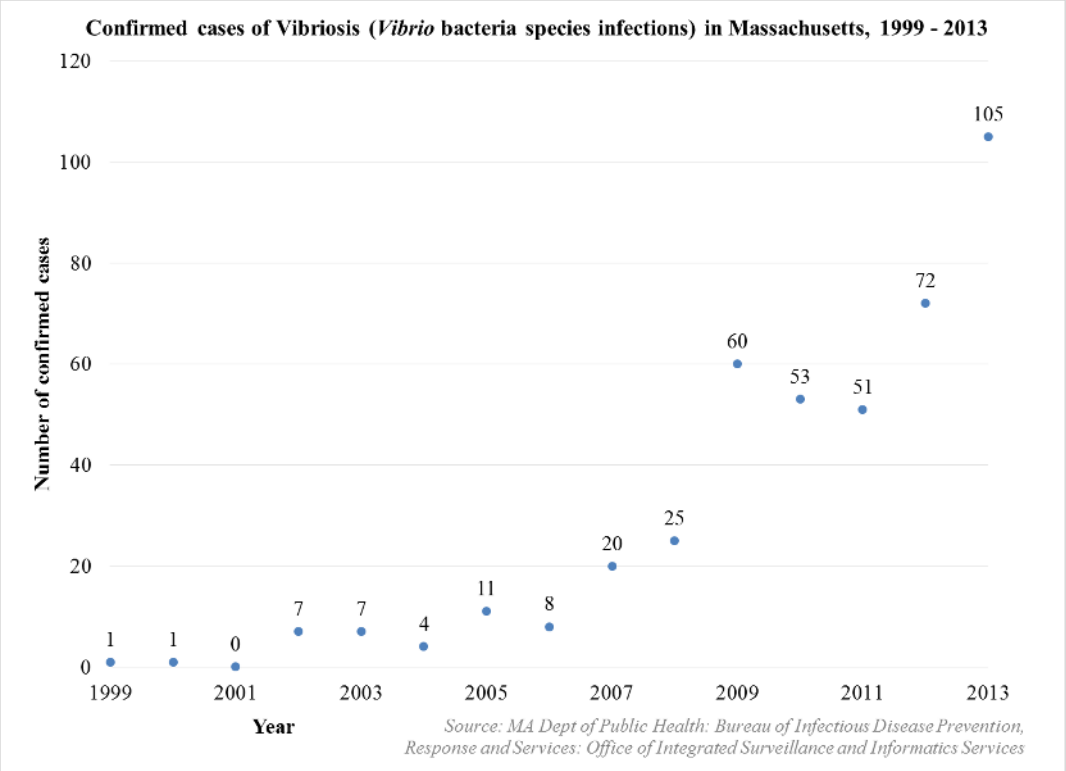

Figure 3-10. Confirmed cases of Vibriosis in Massachusetts, 1999-2013. Vibriosis refers to any illness causes by bacteria in the genus *Vibrio*, including *V. parahaemolyticus*, *V. cholera*, and *V. vulnificus*. Infection by *V. parahaemolyticus* is a more commonly identified than infections with other *Vibrio* species. Source: Commonwealth of Massachusetts, graph by author.

Section Summary. The limitations to using traditional indicator bacteria as the main measure of water quality exist nationally and are clearly applicable to the recreational and shellfish harvesting waters of Massachusetts Bay. Although *V. parahaemolyticus* environmental influences are an active area of study, much of the historical research focus has been on Gulf Coast and mid-Atlantic waters.[102; 110] The reality of *V. parahaemolyticus* presence in the water column and confirmed illness traced to shellfish harvested in Massachusetts Bay[33; 34] supports the argument that current

shellfish harvest area water quality monitoring practices should be revised to reflect the risk of *V. parahaemolyticus* in Massachusetts waters.

Pseudo-nitzschia Diatoms –Species That Can Produce the Toxin Domoic Acid.

***Pseudo-nitzschia* and Domoic Acid – Background**. Domoic acid (DA) is a neurotoxin produced by *Pseudo-nitzschia* genus diatoms, these diatoms are found in estuarine and ocean habitats from tropical to polar waters along every continent.[111] Human exposure to DA via seafood can cause gastrointestinal distress, cardiovascular problems, and memory loss or other neurological effects.[112] DA intoxication is known as Amnesic Shellfish Poisoning because memory loss is one of the most prominent symptoms in human victims. [112] DA has a high binding affinity for nerve cell surface receptors – up to 100 times more powerful than endogenous neurotransmitters.[113; 114] By binding powerfully and not being released from the cell surface DA stimulates nerve cell activation so that, in general terms, DA will stimulate a nerve cell to death.[114] Multiple animal species exhibit negative effects after DA intoxication including anchovies, krill, cephalopods, wild seabirds, sea lions, northern fur seals, and sea otters.[112; 114] There is evidence suggesting that repeated low-level DA consumption, below existing regulatory limits, can have long-lasting negative effects in vertebrates.[115116] The following sections discuss human health risks from DA and reported environmental influences on *Pseudo-nitzschia* spp. population levels.

Domoic Acid - Human Clinical and Epidemiological Considerations. DA was first identified after a 1987 Canadian food poisoning event associated with consumption of contaminated blue mussels (*Mytilus edulis*).[112] This event resulted in 107 confirmed

cases and 38 probable cases, including 19 hospitalizations and 3 deaths within 18 days.[112] After the 1987 outbreak research led to an internationally accepted regulatory limit for DA in food of 20µg DA/g wet weight of tissue.[114; 117] While there have been no known human deaths from DA intoxication since the 1987 Canadian outbreak[114] the true extent of human health impacts from DA consumption is open to question given the limitations on epidemiological data and the possibility of sub-clinical cumulative effects. Since 1987 DA contamination has been regularly documented in multiple marine animal species in at least 62 unique events around the world,[114] indicating potentially widespread presence.

DA is water soluble and exposure occurs by consuming an organism with DA in its system, a process known as biotransfer.[116] DA concentrations are highest in planktivorous species that feed directly on DA-producing diatoms.[116] Human exposure to DA usually occurs through biotransfer via consumption of raw molluscan shellfish that have ingested DA-producing *Pseudo-nitzschia* while filter feeding.[110; 118-123] Being water-soluble, DA is normally cleared from general circulation by the kidneys then excreted in urine within 24 hours.[114] Data from clinical cases indicates that the elderly (age 65+ years), those with impaired renal function, a compromised blood-brain barrier, pregnant women, infants and young children are more sensitive to DA intoxication.[112; 114] DA transfer via maternal milk has been demonstrated in rats, with longer DA retention in milk than in the maternal blood plasma after DA ingestion.[112; 124-126] At present no antidote to DA exists.[114] Exposure to low concentrations of DA can result in significant and permanent effects to the central nervous system, particularly when repeated over long

period of time, raising questions about the safety afforded by current regulatory levels for DA in food.[116] The issue of regulatory safety levels for DA in seafood is discussed at length in Angus (2015)[116] and is not the focus of this work. However, it is worth noting that Washington State has grouped *Pseudo-nitzschia* species into three categories, each with its own threshold abundance level for triggering DA testing in seafood (the lowest of which is 30 cells per milliliter for *P. australis/heimii/fraudulenta*).[116; 127] Similarly, Great Britain has a threshold abundance level of 50 cells per milliliter for total *Pseudo-nitzschia* (not distinguishing between any sub categories) above which shellfish samples are tested for DA.[116; 128-130] To our knowledge Massachusetts has no equivalent official limits.

Direct healthcare costs and ancillary costs from DA exposure or Amnesic Shellfish Poisoning have not been published. However, Ralston, Kite-Powell, and Beet (2011) generated a cost estimate for Paralytic Shellfish Poisoning (PSP) which has a similar route of exposure but a different causative organism (usually the dinoflagellate *Alexandrium fundyense*). The authors estimated the costs of PSP to be US$12.58 million per year.[6] Healthcare costs are just one estimate of the significance of a disease, another measure might be the amount spent on government monitoring programs, however these estimates are beyond the scope of this work.

***Pseudo-nitzschia* species– In the Environment.** Nomenclature and taxonomy have shifted over the years as electron microscopy and molecular approaches have allowed for finer distinctions among *Pseudo-nitzschia* species.[115] At least 38 species of *Pseudo-nitzschia* have been identified[114; 115], but not all have been tested for their ability to produce DA. DA production has been documented for at least 13 species of *Pseudo-*

nitzschia as well as the related diatom species *Nitzschia navis-varingica*.[115; 116] Laboratory testing has shown that for multiple *Pseudo-nitzschia* species DA production increases with increasing environmental stressors.[114] Some researchers believe that all *Pseudo-nitzschia* species can be toxigenic under the right growth conditions.[114] Field observations suggest that large-scale environmental drivers are important influences on *Pseudo-nitzschia* blooms as there are seasonal patterns to blooms around the globe.[114] *Pseudo-nitzschia* blooms, like those of other diatoms, tend to occur in upwelling zones, coastal bays, or in response to controlled nutrient pulses (such as Fe-enrichment).[111; 114] European waters historically experience blooms of *Pseudo-nitzschia* from January-May, eastern North American in the autumn, Washington State in early autumn, and the Pacific Mexican coast in late spring.[114] Off the coast of central California the concentration of DA-producing diatoms is usually highest between late summer and fall, a time associated with the end of seasonal coastal upwelling and nutrient depletion in the water column.[116] There is no set abundance that defines a bloom event, different thresholds have been used by researchers in different locations.[110; 131] *Pseudo-nitzschia* blooms vary in their toxicity and the relationship between environmental factors, *Pseudo-nitzschia* abundance, DA production, and DA bioaccumulation in filter feeders is poorly understood.[131] It appears that the relative proportion of nutrients in the water column may be the limiting factor for *Pseudo-nitzschia* diatom growth.[116] Upwelling zones and coastal bays are areas naturally high in nutrients and trace metals, factors which have been linked to DA production in multiple *Pseudo-nitzschia* species.

The state of one environmental factor can influence the response of *Pseudo-nitzschia* to another environmental factor. Species in the *Pseudo-nitzschia* genus tolerate water temperatures ranging from -1.5°C to 30°C, and laboratory studies have demonstrated that multiple species can have ~10°C overlap in their temperature tolerance – making this environmental variable minimally suited for predicting species succession or niche occupation.[114] For example, in laboratory studies *P. cuspidata* can tolerate a wider range of temperatures when grown at its optimum salinity (30 psu), and *P. pseudodelicatissima* achieved its highest growth rate at 25°C when also grown in its optimum salinity (other temperatures not tested).[114] Field sampling has documented the presence of *Pseudo-nitzschia* species in Massachusetts Bay at water temperatures ranging from approximately 2 to 22°C.[132; 133] Division rates for *P. multiseries* asexual reproduction in culture have been shown to vary between 0.21 and 1.2 divisions per day depending on temperature and light intensity conditions.[134] All species of *Pseudo-nitzschia* display phenotypic variety in their size (width), possibly due to local environmental conditions, complicating species-level identification efforts. Waters around continental margins and near-coastal regions commonly contain larger species such as *P. australis* and *P. multiseries,* larger species are reported to be able to produce more DA per cell.[114;111] Previous research has attempted to quantify the relationship between either *Pseudo-nitzschia* growth or DA production, and environmental factors. A list of commonly investigated variables is shown in Table 3-13, below. Major nutrients (e.g., Nitrogen, Carbon, and Phosphorous in different forms) and physical parameters (e.g., temperature, sunlight, turbidity, and salinity) have been investigated in field studies

and laboratory experiments, but the variety of species and environmental conditions limits our ability to generalize across the entire *Pseudo-nitzschia* genus.

Table 3-13**.** Environmental influences of *Pseudo-nitzschia* species growth

Potential environmental influences of *Pseudo-nitzschia* species growth	**Evidence strength**	**Reference**
Total inorganic carbon (bicarbonate and carbonate) (positively associated)	Weak	114
Nitrogen in the form on ammonium, nitrate, nitrogen dioxide, urea, ammonia, or glutamine (positively associated)	Strong	110; 114
Salinity (positively associated)	Strong	110
Freshwater discharge (negatively associated)	Strong	110; 131
Temperature (higher temperature is negatively associated)	Strong	110; 131
Inorganic Phosphate (PO_4) (slightly negatively associated)	Medium	110
Ln of Silicic Acid ($Si(OH)_4$) (negatively associated)	Strong	110; 131
Ratio of Silicon to Phosphorous or Orthophosphate (Si:P) (slightly positive)	Weak	110
Turbidity or Secchi depth (slightly positive)	Strong	110; 114
Dissolved Organic Carbon (DOC) (positively associated)	Weak	110
Chlorophyll a (not associated in Chesapeake Bay, but positively associated in Monterey Bay, CA)	Mixed	110; 131
Trace metal: Iron (strongly positively associated)	Strong	111; 114
Trace metal: Copper (negatively associated with cell growth; positively associated with DA production)	Weak	114
Trace metal: Lithium	Weak	114
Upwelling (positively associated)	Medium	131

A graphical representation of environmental factors that seem to generally influence *Pseudo-nitzschia* abundance is presented in Figure 3-11, below. Nitrogen, salinity, water temperature, iron, carbon and silicic acid appear to influence positively *Pseudo-nitzschia* species abundance under multiple conditions. However, as noted by

Downes-Tettmar et al. (2013), *Pseudo-nitzschia* species vary in their response to the environment and thus may form corresponding groups that are not based on morphology.[130] Studies that group multiple species of *Pseudo-nitzschia* together based on size (a commonly used morphology distinction) may lose important ecological information in their analysis, concurrent species-level and group-level analysis is likely to be more revealing.[130]

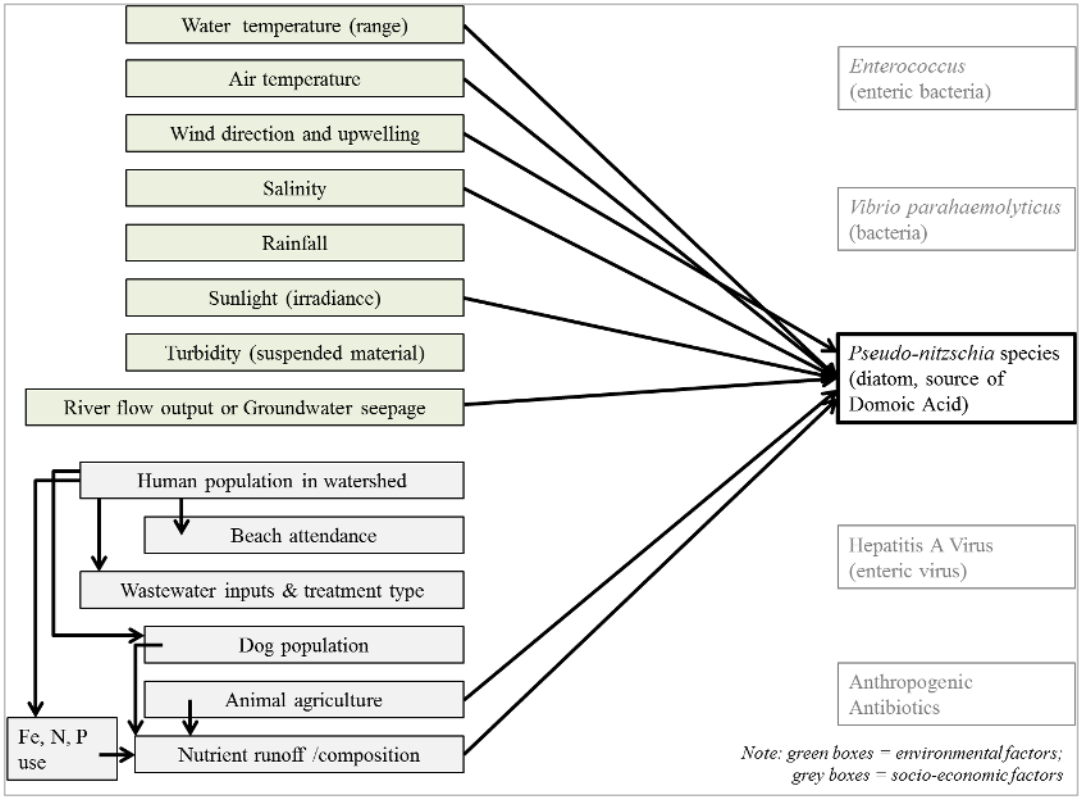

Figure 3-11. Graphical representation of influences of *Pseudo-nitzschia* species abundance. Although there is variety between species. Lines between influences (left side of figure) indicate interactions between influences.

***Pseudo-nitzschia* species. – In the Massachusetts Bay Area.** *Pseudo-nitzschia* spp. were originally identified as part of Gulf of Maine phytoplankton assemblages in the 1920s.[115; 135] Although not every species of *Pseudo-nitzschia* has been documented in Massachusetts Bay, multiple species that have been confirmed in the western North Atlantic Ocean.[114] Multiple *Pseudo-nitzschia* species could be present year-round, or be introduced seasonally, in Massachusetts Bay. DA was detected in shellfish off Nantucket, MA as early as 1991 and the presence of at least one DA-producing diatom species in Massachusetts Bay was confirmed in 1992.[136] Public, but unpublished, data from the Massachusetts Water Resources Agency (MWRA) demonstrates that ongoing

sampling for *Pseudo-nitzschia* species shows they have been regularly present in Massachusetts Bay since sampling started in 1992.[132] We have acquired the *Pseudo-nitzschia* sampling dataset from the MWRA for the years 1992-2014, the details of which will be discussed in Chapters 3 and 4. A list of *Pseudo-nitzschia* species whose presence has been confirmed in western North Atlantic waters, including the Gulf of Maine and Massachusetts Bay, is presented in Table 3-14, below.

Table 3-14. Presence of *Pseudo-nitzschia* species in western North Atlantic waters.

Species	**Group**	**Cell Width (µm)**	**Toxicity**	**Documented distribution**
Nitzschia closterium[+]	Pre-taxonomic reorganization	Not reported	Not tested	Vineyard Sound, Massachusetts [135]
Nitzschia longissima[+]	Pre-taxonomic reorganization	Not reported	Not tested	Vineyard Sound, Massachusetts [135]
Nitzschia seriata[+]	Pre-taxonomic reorganization	Not reported	Not tested	Vineyard Sound, Massachusetts [135]
Nitzschia pungens f. multiseries[+]	Pre-taxonomic reorganization	Not reported	Yes	Massachusetts Bay, USA[136]
N. pseudodelicatissima	Pre-taxonomic reorganization	Not reported	Yes	Massachusetts Bay, USA[136]
P. americana	neither	2.5 – 4.5	No	Bay of Fundy Narragansett Bay
P. americana	*Americana*	<2.0	Not tested	Gulf of Maine[*]
P. calliantha	*delicatissima*	1.1-2.6	Some	Gulf of St. Lawrence
P. delicatissima	*delicatissima*	1.0-2.4	Some Yes[§]	Gulf of St. Lawrence, US northeast coast Gulf of Maine[*]
P. fraudulenta	*seriata*	4.0-8.0	Some Yes[§]	Bay of Fundy Gulf of St. Lawrence US northeast coast Gulf of Maine[*]

Table 3-14. Presence of *Pseudo-nitzschia* species in western North Atlantic waters.

Species	Group	Cell Width (µm)	Toxicity	Documented distribution
P. linea	*delicatissima*	1.8-2.2	Not tested	Narragansett Bay
P. multiseries	*seriata*	3.5-4.8	Yes	Bay of Fundy Gulf of Maine Gulf of St. Lawrence US Northeast
P. obtuse	*seriata*	2.9-5.0	No	Gulf of St. Lawrence Hudson strait Newfoundland
P. pseudodelicatissima	*delicatissima*	1.1-2.1	Some Yes[§]	Bay of Fundy Gulf of Maine[*]
P. pungens	*seriata / pungens*	2.2-5.4	Some Yes[§]	Bay of Fundy Gulf of St. Lawrence US east coast Gulf of Maine[*]
P. seriata	*seriata*	4.6-8.0	Some Yes[§]	Bay of Fundy Gulf of St. Lawrence US northeast coast Gulf of Maine[*]
P. subpacifica	*seriata*	5.0-7.0	No Yes[§]	Bay of Fundy Gulf of Maine[*]
P. turgidula	*pungens*	2.5-5.0	Yes[§]	Bay of Fundy Gulf of Maine[*]
P. heimii	*Seriata*	>4.0	Not tested	Gulf of Maine[*]
P. species Gulf of Maine (novel form)	*pseudo-delicatissima/ delicatissima*		Yes[§]	Gulf of Maine[*]

Adapted and expanded from Lelong et al. (2012). [114] Toxicity refers to toxin analyses: Yes = species produces DA; No = values below the limit of detection; Some = not all strains show toxicity.[114] Yes[§] = DA production confirmed in Gulf of Maine

[+] Reports published before adoption of the currently used nomenclature and species identification criterion for *Pseudo-nitzschia* species. DA-producing diatoms were originally reported as *Nitzschia pungens* f. *multiseries*, more recent attribution is to *Pseudo-nitzschia multiseries*.

[*] Identified in Fernandes et al. (2014)[115]

Due to ocean circulation patterns in the Gulf of Maine there is the possibility for movement and mixing of multiple *Pseudo-nitzschia* species. Lack of published documentation about the presence of a single species in a single place should not be taken as an indication that a species is not sometimes present in a water body, especially given the large spatial scale of these environments and limited sampling coverage.

Section Summary. *Pseudo-nitzschia* species diatoms have been found in Massachusetts Bay, the Gulf of Maine, and along the Eastern coast of North America. These diatoms are a public health concern because of their capacity to produce the neurotoxin Domoic Acid, which may be transmitted to humans via shellfish consumption. The presence of *Pseudo-nitzschia* in the water column is considered necessary, but not sufficient, for DA production. In Washington State and Great Britain public authorities have set threshold abundance levels for *Pseudo-nitzschia* which can automatically trigger testing for DA in shellfish. Chapter 4 of this dissertation will attempt to identify environmental variables associated with *Pseudo-nitzschia* bloom levels in Massachusetts Bay – where shellfish harvesting is culturally and commercially important.

Hepatitis A Virus (HAV) – A Virus That Damages the Human Liver.

HAV – Background. Hepatitis A Virus (HAV) is a picornavirus transmitted through contaminated food and water and direct person-to-person contact; it is estimated to infect tens of millions of individuals worldwide every year.[137; 138] Humans are the only known reservoir of HAV, there is no other animal host, but the virus can survive outside

of humans for varying amounts of time depending on environmental conditions and be conveyed on food, including seafood.[139-141] There is no treatment for acute HAV infection, only supportive care in response to symptoms which commonly include fever, malaise, anorexia, nausea, abdominal discomfort, jaundice, necrosis and inflammation of the liver.[69; 137] HAV is extremely infectious, the minimal infectious dose is extremely low, possibly just a single virus.[139] A complicating factor in epidemiological investigations of HAV is the lag time between exposure and the development of symptoms, typically 21 to 35 days.[142]

HAV - Human Epidemiological Considerations. Since 1995 a vaccine for HAV has been available in the U.S. and many other countries; in 2005 the allowable age for the first dose was lowered to 12 months for children in the U.S.[137] In unvaccinated populations the age of first exposure to HAV is important, because symptoms typically worsen with increasing age of first exposure.[142] Endemicity of HAV infection is low in the U.S. because of the wide access to treated drinking water, however HAV remains a public health concern because of its highly infectious nature and the ability for patients with active infections (including asymptomatic children) to release large numbers of HAV into the environment through feces[139] and infect HAV-naïve individuals.

Globally, the incidence rates (number of cases per population) of HAV are strongly correlated with socioeconomic status and access to clean drinking water.[138] Improved sanitation and hygiene means that fewer countries are considered highly HAV-endemic, but instead are considered to have intermediate or low endemicity. [139; 143] The CDC notes that countries with decreasing prevalence of HAV infection have increasing

numbers of susceptible people, and subsequently there is a risk of large outbreaks of HAV among these susceptible populations.[137] European countries with low endemicity and susceptible populations have seen an increased number of HAV outbreaks among adults, including a 2008 outbreak in the Czech Republic that resulted in over 1600 cases, and a 2007-2009 outbreak in Latvia resulting in over 3200 reported cases.[143] In Argentina, before the HAV vaccine was introduced into the national immunization program for young children most HAV cases were in children age 5-9 years, but by 2010 that had shifted to adults age 15-44 years (there are no reports of outbreaks like those in Latvia or the Czech Republic).[144] The most prevalent reported risk factors for HAV in the U.S. is international travel.[145] U.S. residents who travel outside of the country may be exposed to HAV through contaminated food, drinking water, or recreational water.[144; 146] This is due to higher levels of HAV endemicity in most other parts of the world except Northern Europe and Japan.[139]

HAV – Outbreaks Associated With Seafood. Seafood is a known potential vector for HAV. As early as in 1961 raw clams and raw oysters were implicated in outbreaks of 'infectious hepatitis' in the U.S.[147; 148] This was before infectious agents for the different strains of viral hepatitis, including HAV, had been isolated and identified. A more recent outbreak of seafood associated HAV happened in 2005. The 2005 outbreak involved twelve restaurants in four states (Alabama, Florida, South Carolina, Tennessee) that received oysters harvested from approved harvest areas in the waters of eastern Louisiana.[142] Based on an epidemiological investigation and successful traceback of the suspected oysters, researchers were able to identify HAV in shellstock oysters from the

same harvest batch. Investigators concluded that the contamination of oysters with HAV most likely happened in the harvest area, rather than during the processing stages (shucking, packing) and that the most probable sources of contamination were either 1) illegal wastewater discharges from harvest vessels or recreational boats in the harvest areas, or 2) illegal harvesting in closed areas.[142] No confirmed cases of HAV originating in Massachusetts-harvested shellfish have been reported, but the link between 'viral hepatitis' and shellfish consumption in New England was observed by medical professionals over 40 years ago.[147; 148] A 1967 study carried out at 10 Boston hospitals found that ingestion of raw shellfish (oysters and clams) and steamed clams was significantly more common in viral hepatitis patients than in matched controls during a prospective epidemiological study of 270 patients (258 of whom were New England residents).[147; 148] This same study noted that "ingestion of raw shellfish and steamed clams seems to be as common a source of infection as contact with jaundiced persons."[147; 148] This is perhaps not surprising since we know that HAV is transmitted via the fecal-oral route and contact with infected individuals (such as 'jaundiced persons') is a risk factor for HAV transmission. The 1967 Boston study attempted to trace the source of implicated shellfish during the study period, but results were inconclusive.[147; 148] A slightly later publication from 1968 noted that "present shellfish cleansing technics (depuration) may not effectively remove the virus."[149] There is currently no monitoring or sampling program for HAV in Massachusetts coastal waters.

HAV – In the Environment. Humans are the only known carriers for HAV, but the virus can survive outside of the human body for extended periods of time depending

on environmental conditions.[139] HAV survival has been documented in seawater [150] and groundwater[151], including groundwater serving as drinking water via private wells. Viruses that infect the gastrointestinal tract (known as enteric viruses) tend to have prolonged survival in the environment; HAV falls into this category.[152] HAV is described as being 'thermally resistant' due to its ability to survive at low temperatures (0-10°C) for long periods of time.[152; 153] Other characteristics of enteric viruses that favor their survival outside of the human body include an adaptation to waterborne route of transmission[152] and small size (25-100 nm). [151] The successful spread of HAV via water can depend on physical parameters such as soil structure, organic carbon contact, soil pore water pH, environmental factors such as rainfall and temperature, and virus-specific characteristics such as size and electrical charge (which affects solubility in water).[151; 154] Since humans are the only known carriers for HAV any waterborne transmission must be preceded by human fecal contamination.[139] Fecal contamination may be spatially and temporally distant from the site of eventual exposure since HAV can survive in water or on contaminated food at length, studies have demonstrated HAV survival on fresh produce for 90 days when stored at 20°C, and in oysters for >21 days in a depuration tank at 20-25°C.[140; 141] Potential sources of fecal contamination into the environment include leaking or ineffective septic tanks, leaking sewer lines, unlined landfills, irrigation with wastewater, or subsurface injection of wastewater.[152]

This research does not attempt to quantify the amount of HAV released into the Massachusetts coastal environment. Rather, we assert that introduction of HAV into the coastal environment may occur and that HAV may be present in recreational and

shellfish harvesting waters in Massachusetts Bay. These waters are not currently monitored for the presence of HAV. If water samples were tested for HAV, and it were found to be present, environmental modeling efforts would consider the known potential influences on the survival of any human virus once released into the environment. Environmental influences on the release and survival of any human enteric virus are shown in Table 15 below (adapted from Gerba 2007 and expanded).[152] The same information is presented graphically in Figure 3-12, both Figure 3-12 and Table 3-15 indicate that unintentional release of enteric viruses through fecal contamination can happen through multiple pathways.

Table 3-15. Influences on Virus Presence and Survival in Groundwater and Surface Water. *Adapted from Gerba (2007)*[152]

Influences on viral persistence in the environment	**Evidence Strength**	**Reference**
Sewage discharge (positively associated)	Strong	139; 152
Sewage Treatment (Disinfection of wastewater reduced the number of viruses found in surface waters.) (negatively associated)	Medium	152
Ultraviolet light (Viruses vary in sensitivity) (negatively associated)	Medium	152; 155
Time of year (Longer survival of viruses in colder winter waters. Higher concentrations of enteroviruses in the summer than in winter in temperate climates.)	Strong	152
Rainfall (During high rainfall events sewage treatment plants may bypass certain treatment steps or reduce treatment times, introducing nutrients [66])	Medium	152
Direct shedding from infected individuals (may re-introduce virus into waters)	Strong	139; 152
Temperature, cooler temperatures associated with longer survival time. (Infectious HAV survived >21 days in oysters 4^{o}C seawater)	Strong	140; 141

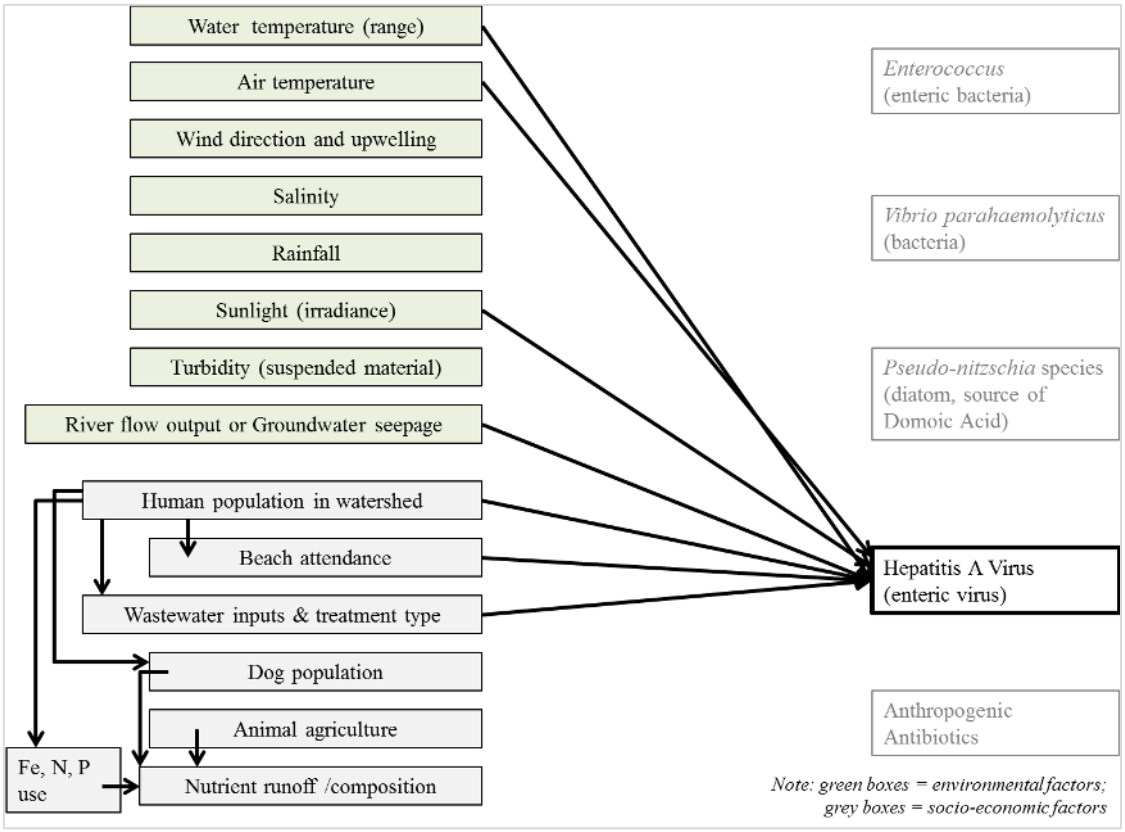

Figure 3-12. Graphical representation of influences for HAV presence in coastal waters. Lines between influences (left side of figure) indicate interactions between influences.

Areas with sewage contamination of seawater are at risk of containing HAV, as are areas with large congregations of recreational bathers who might be shedding enteric viruses into the water.[152] Although Table 3-15 above treats groundwater and surface water separately, nearshore coastal ocean waters may receive freshwater groundwater inputs. Depending on local hydrology viruses may travel via groundwater into nearshore coastal waters. The large human population in the coastal watersheds around Massachusetts Bay, combined with a variety of sewage treatment methods releasing

wastewater into the environment, suggest that if HAV is present in the local population is may be released into coastal waters through multiple routes.

HAV – In Massachusetts. Vaccination for HAV is recommended by the CDC as part of the regular schedule of vaccinations for children,[156] but the HAV vaccine is not required for public school attendance in Massachusetts.[157] HAV vaccination rates have changed over time in Massachusetts, they appear to be increasing for young children, but immunity rates remain low for adults. In 2003 the estimated HAV vaccination rate for children aged 24-35 months was less than 1% for Massachusetts, and just under 4% for the Boston area.[158] However, in 2014 the estimated vaccination coverage of $\geq$2 doses among children aged 19-35 months had risen to 64% in Massachusetts, with a national average of 57%.[159] Encouragingly, reported cases of acute HAV infection in Massachusetts decreased by 61% between 1999 and 2008 and remained low through 2013, shown below in Figure 3-13.[160; 161]

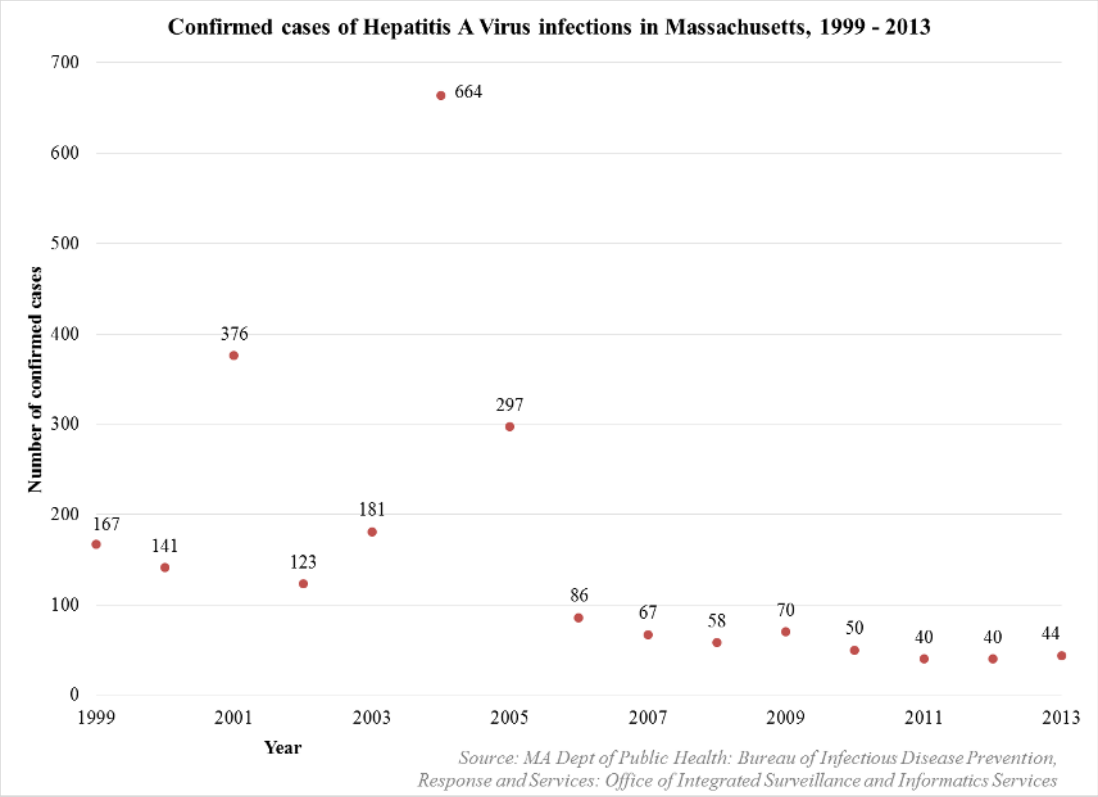

Figure 3-13. Confirmed cases of Hepatitis A Virus infection in Massachusetts, 1999-2013.[109] Data Source: Commonwealth of Massachusetts, graph by author.

A 2010 national survey of U.S. children ages 6-19 years old reported 37.6% prevalence of immunity to HAV, with the lowest prevalence among white, non-Hispanic children.[145] Immunity can be acquired through vaccination or environmental exposure, so immunity prevalence is not wholly indicative of vaccination coverage. A 2009 study of HAV vaccine completion among children aged 13-17 estimated that in Massachusetts the coverage from completing 1 dose of the vaccine was 9% and coverage for having completed 2 doses was almost 7%, much lower than the national 1-dose coverage estimate of 42% for the same age group[162] (note that this age cohort would be 19-23 years old in 2015). In 2013 a national survey for adults aged 19 or older estimated that the

proportion who had ever received 2 or more doses of HAV vaccine was 9%, but among those who had traveled outside of the U.S. (to countries other than Europe, Japan, Australia, Canada, or New Zealand since 1995) the estimated vaccine coverage was almost 16%.[163] National and state level trends suggest that HAV vaccination coverage for young children is improving, but that for many age groups coverage is still well below the CDC target of 85%.[159] Despite CDC recommendations and vaccine accessibility, immunity to HAV is not universal among children or adults in the U.S. This creates the possibility of increasing age of first exposure to HAV and resulting increased severity of individual illness or larger outbreaks.

Section Summary. Seafood, including raw shellfish, can be a vector for HAV. HAV can survive for extended period of time outside of a human host. HAV can be introduced into the marine environment through human fecal contamination. Although HAV vaccination rates are increasing in the U.S., there is a large population without immunity who are at risk of a more severe acute illness if exposure to HAV first occurs in adulthood.

Anthropogenic Antibiotics- Manufactured and Released by Humans.

Anthropogenic Antibiotics – Background. Anthropogenic antibiotics are antibiotic compounds manufactured by humans, as opposed to the antibiotics produced by bacteria species in the wild. The presence of antibiotics exerts selection pressure upon bacteria and favors the survival of antibiotic resistant bacteria (ARB). ARB are globally recognized as a serious public health problem.[164-166] Three major factors contribute to the development and spread of ARB: 1) the overuse, or incorrect use, of antibiotics in human medicine, 2) the medically unnecessary use of antibiotics for livestock growth promotion and the subsequent entry of antibiotic resistant bacteria into the food supply, and 3) the spread of ARB between people or from environmental exposure to non-human sources of such bacteria.[167]

Anthropogenic Antibiotics – Human Epidemiological Considerations. The CDC conducts surveillance for antimicrobial resistance among **enteric** bacteria isolated from humans through the National Antimicrobial Resistance Monitoring System (NARMS).[168] In 2009 NARMS started testing isolates of *Vibrio* species other than *V. cholera* for antibiotic resistance, and public health laboratories were asked to forward every isolate of *Vibrio* species that they receive to the CDC, note that the number of isolates is not the same as the number of culture-confirmed *Vibrio* infection cases.[168] In 2013 the CDC received 607 (non-*V. cholera*) *Vibrio* isolates to test for antibiotic resistance in the NARMS program, 47 of these were from Massachusetts.[168] Only two states sent more non-*V. cholera Vibrio* isolates to the CDC for testing in 2013, Florida (124) and Washington State (62).[168] Of the 607 *Vibrio* isolates received 317 were *V.*

parahaemolyticus, and 40% of those samples demonstrated resistance to the penicillin-class antibiotic Ampicillin, but not to the other antimicrobial agents tested. *Vibrio* is only one genus of enteric bacteria included in the NARMS program, discussed here because they are a marine-sourced risk known to be present in Massachusetts Bay (unlike some of the other enteric bacteria tested by NARMS). There are many other species of bacteria which may display antibiotic resistance and infect other parts of the body besides the gastrointestinal tract.

The CDC estimates that at least 2 million people in the U.S. experience ARB infections every year, that over 20,000 people die as a direct result of these infections, and that many more die from co-morbidities and associated complications.[167] There are numerous possible routes of exposure, include sea-bathing in waters where these bacteria are present or recreation in areas where beach sands harbor ARB.[83] One possible route of introduction for antibiotics or ABR into the marine environment is through contaminated wastewater effluent released from municipal wastewater treatment plants (WWTPs) that discharge into coastal waters.

Anthropogenic antibiotics – In the Environment. After antibiotics are consumed by patients the antibiotics and their metabolites are introduced into wastewater systems across the country. Wastewater treatment plants (WWTPs) receive wastewater laced with antibiotics and other pharmaceuticals, but treatment processes are generally designed to remove nutrients and kill or reduce microbial pathogens, not to inactivate a wide variety of chemical compounds; multiple studies have documented the presence of antibiotics[169-173] or ARB[174-179] in municipal wastewater effluent. The detection of ARB

in municipal WWTPs has led to the comparison of WWTPs as *de facto* reservoirs of such pathogens.[174; 175; 180; 181] A suggested mechanism for how WWTPs serve as reservoirs is that biological treatment processes facilitate the spread of resistance by continuously mixing bacteria with antibiotics at sub-inhibitory concentrations,[175] and that this process itself favors either the transfer of resistance genes via horizontal gene transfer or the survival of bacteria with resistance genes.[174] A generalized version of this process, and the findings by Su et al. (2014)[174] from sampling for antibiotic resistant *E. coli* at two Chinese WWTPs along different phases of treatment (from influent to effluent) are diagrammed in Figure 3-14 below. Su et al. (2014) found that wastewater treatment and disinfection decreased the total number of culturable bacteria from influence to effluent, but led to an increase in the percentage of antibiotic-resistant *E. coli* in the final effluent.[174]

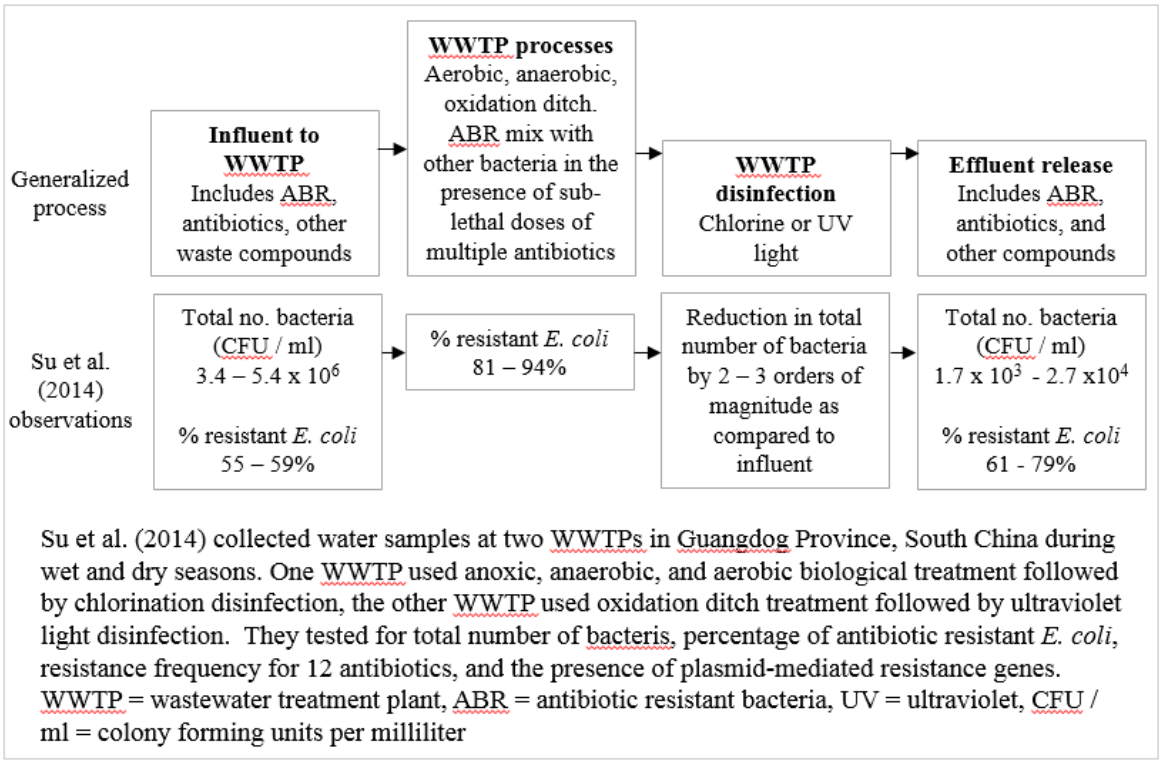

Figure 3-14. How Wastewater Treatment Plants May Act As a Source of Antibiotic Resistant Bacteria. Data source: Su et al. (2014)[174], figure by author

The persistence and ultimate fate of these ARB when they are released into the environment through effluent discharge is uncertain. However, sea-bathers downstream of wastewater releases may be at risk of high levels of exposure to ARB because of the potential for direct contact across multiple body surfaces and unintentional water ingestion associated with aquatic recreation. Given the variable virulence of bacterial species (including emerging threats such as Methicillin-resistant *Staphylococcus aureus* (MRSA)) traditional water quality monitoring methods based on numerical surveillance of *Enterococcus* may be insufficient to adequately protect vulnerable populations when they recreate in coastal waters.

Animal Husbandry as a Potential Source of Anthropogenic Antibiotics. In addition to antibiotic resistant bacteria released from WWTPs, the use of antibiotics in animal feed can contribute to the increase of ARB in the environment. The human population in Massachusetts is concentrated in the eastern half of the state, but such development has not driven out all agricultural activities in the same area. While there are animal husbandry operations in counties which overlap with Massachusetts Bay watersheds, the total livestock population is much smaller than the human population, for example, in 2007 there were an estimated 1,000 hogs in all of Middlesex County.[182] Compared to other U.S. states Massachusetts has a small number of livestock, in the year 2011 there were approximately 40,000 head of cattle in all of Massachusetts, but 6,300,000 in Kansas.[183] Livestock use of antibiotics in Massachusetts may impact water quality, but this source as an overall *pressure* for anthropogenic antibiotic inputs likely pales in comparison to those contributed to coastal ocean areas from direct human consumption.

It is difficult to separate out the potential impacts on human health risk of ARB in the marine environment from the presence of anthropogenic antibiotics unintentionally released into the environment. Freely mobile antibiotics can exert their selective pressure on bacteria until the compound is degraded by physical or chemical means. Any influence of seasonal variations in antibiotic use, and possible resulting fluctuations in populations of antibiotic–resistant bacteria in the environment, has yet to be determined. Table 3-16, below, lists known influences on the release and persistence of anthropogenic

antibiotics into the coastal environment. Figure 3-15 depicts the same information in graphical form.

Table 3-16. Influences on anthropogenic antibiotic releases and presence of antibiotic resistant bacteria in coastal waters

Influences on anthropogenic antibiotics in coastal waters	**Evidence strength**	**Reference**
Amount of antibiotics used/prescribed within coastal watershed	Strong	174; 176; 180; 181; 184; 185
Environmental conditions favorable to antibiotic stability/persistence, which may include the following: Temperature, salinity, sunlight (irradiance), pH	Limited	186
High rainfall events resulting in discharge of raw sewage	Limited	71

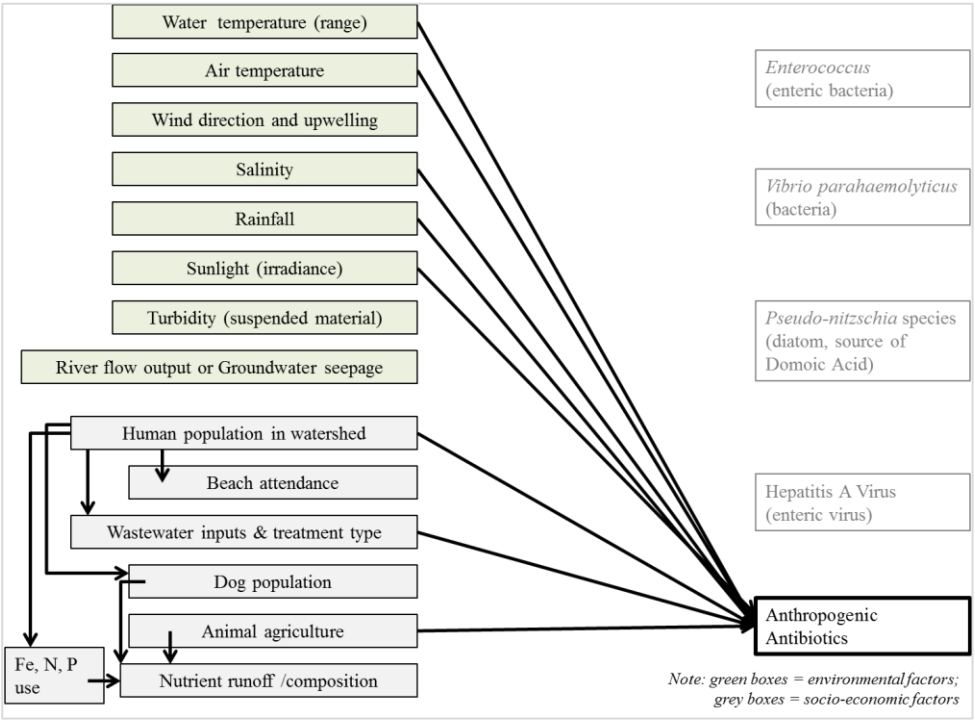

Figure 3-15. Graphical representation of known influences for anthropogenic antibiotics, presence and persistence, in coastal waters. Lines between influences (left side of figure) indicate interactions between influences.

Anthropogenic antibiotics – Estimated Usage in Massachusetts Bay Area.

Antibiotic sales are divided into two categories, human medical use and veterinary use. The amount of antibiotics used to treat food-producing animals are publicly reported by the Food and Drug Administration (FDA).[187] *Drug Use Review: Systemic Antibacterial Drug Products,* is the most recent FDA report on antibiotics sold for use in humans, this brief report covers the years 2010-2011 and was published in 2012.[184] *Drug Use Review* assembled data on sales of select systemic antibacterial drug products; in 2010 approximately 3.28 million kilograms (kg) were sold, and in 2011 the amount sold was 3.29 million kg.[184] In both years the most common active ingredient of all selected systemic antibacterial drug products sold was amoxicillin. Although the FDA does not report sales at the state level, we have estimated the amount of antibiotics sold in Massachusetts and specifically in our coastal watershed study area, the method is described below.

In 2010 U.S. healthcare providers prescribed 258 million courses of antibiotics,[185] with 3.28 million kg antibiotics sold in 2010[184], averaging out to 0.013 kg antibiotics per prescription. Antibiotic prescription rates vary by region, in the Northeast, the rate of antibiotic prescriptions is 830 prescriptions per 1,000 persons with the most commonly prescribed antibiotics coming from the penicillins category (23%), followed by macrolides (22%).[185] Based on its proportion of the U.S. population Massachusetts would be expected to account for approximately 65,600 kg of antibiotics sold in the U.S. in 2010. To narrow the estimate of antibiotic use to the six coastal watersheds that serve as our study area we multiplied the average prescription rate in the Northeast by the

population in the six coastal watersheds, resulting in an estimated 2,423,600 antibiotic prescriptions per year in the 6 coastal watersheds around Massachusetts Bay. Multiplying the average weight of antibiotics per prescription (0.013 kg) by the estimated number of prescriptions in the coastal watersheds study area (2,423,600) yields an estimate of 30,800 kg of antibiotics consumed in the study area in 2010. These numbers and calculations are summarized below in Table 3-17.

Table 3-17. Information to Estimate Massachusetts Antibiotic Usage, 2010

Informational Element	Symbol	Number	Source
U.S. total population	α	308,745,538	U.S. Census, 2010[56]
Massachusetts (MA) population	β	6,547,629	U.S. Census, 2010[56]
MA percent of U.S. population (β / α) x 100 = γ	γ	2	calculated
Estimated population in six MA Bay coastal watersheds study area (see Table 3-6)	δ	2,920,000	calculated
Percent of MA population residing in six coastal watersheds (δ / β) x 100 = ε	ε	45	calculated
U.S. total human antibiotic usage in year 2010 (kg)	ζ	3,280,000	FDA, cited in Pham 2012[184]
U.S. total number of antibiotics prescriptions in 2010	λ	258,000,000	IMS Health Xponent database, cited in Hicks et al. 2013[185]
U.S. Northeast antibiotic prescription rate (θ = 0.83)	θ	830 per 1,000 people (0.83)	Hicks et al. 2013[185]
Estimated number of total antibiotic prescriptions in MA (β x θ = μ)	μ	5,434,532	calculated
Estimated number of antibiotic prescriptions in study area, (μ x ε = ψ)	ψ	2,423,600	calculated

Table 3-17. Information to Estimate Massachusetts Antibiotic Usage, 2010

Informational Element	Symbol	Number	Source
Estimated total antibiotic usage (kg) in study area, $(\zeta / \lambda) \times \psi = 30{,}8000$ kg		**30,800 kg**	calculated

We estimate that 30,800 kg of antibiotics were consumed, excreted (after being metabolized to varying degrees), and released into wastewater treatment systems in the study area in 2010. The wastewater treatment systems in the six coastal watersheds range in size and complexity from septic systems serving individual homes to the Deer Island Wastewater Treatment Plant serving 43 Boston-area communities and processing 350 million gallons of wastewater per day.[188] No matter what size treatment plant, the goal of municipal wastewater treatment is primarily to reduce nutrient outputs into surface water, and secondarily to minimize public health threats that might spread through untreated wastewater.[180] However, as discussed above, WWTPs are not designed to eliminate the vast variety of anthropogenic pollutants that enter wastewater stream, including anthropogenic antibiotics.

Section Summary. The presence of anthropogenic antibiotics in marine environments may encourage the development of antibiotic-resistance, or favor the survival and distribution of any ARB released in wastewater effluent. The magnitude of the release of anthropogenic antibiotics and antibiotic-resistant bacteria into Massachusetts Bay through public and private wastewater treatment systems is unknown. We have presented a first order estimate of annual antibiotic use in the six coastal watersheds that serve as our study area. Direct sampling of wastewater effluent and

environmental samples would provide valuable information to refine this estimate. Due to the limited presence of livestock in the watersheds around Massachusetts Bay, and in Massachusetts overall, we expect that human use and release through wastewater is the most significant contributor to the presence of anthropogenic antibiotics in Massachusetts Bay. The next section introduced the topic of environmental modeling and describes how the information presented in this chapter on anthropogenic antibiotics and other marine-sourced risks could contribute to environmental modeling of these risks in Massachusetts Bay.

Environmental Modeling.

Modeling reduces complex systems to simplified versions of reality.[189] Environmental models often attempt to draw boundaries and describe the orderly dynamics of a system that may in reality have no boundaries and be full of chaotic interactions. In other words "our models fall far short of representing the real world fully."[189] Models may be imperfect but they still prove useful for describing relationships and forecasting changes of interest (with some expected degree of over- or under-prediction). In this work we are truly interested in the number of illnesses caused by various marine-sourced agents, but we know such epidemiological data to be limited and largely insufficient for building a predictive model for future risks in Massachusetts. Therefore, we have taken one step backwards in the chain of events and are interested in modeling the abundance of marine-sourced risks themselves, as they exist in the environment, before humans are exposed to the risk. The section below provides further

rationale for using this approach (model development methodology is described in Chapter 4).

Environmental Modeling Limitations. Research on the relationship between environmental variables and population levels of pathogenic organisms has been happening for decades. Traditionally this involved extensive sampling from field sites. In some cases there are clear relationships among a few variables and the outcome of interest. For example, temperature and rainfall were able to accurately hindcast the presence/absence of human enteroviruses in Charlotte Harbor, Florida with 97.3% accuracy.[71] However, when moving towards a scenario where *multiple* marine-sourced risks are of interest, as is the case with this research, it is likely that the relationships between variables and outcomes, and among the variables themselves, will be more complex. It is clear that all of the risk categories identified in this chapter (enteric bacteria, enteric viruses, indigenous marine bacteria, marine toxins, and anthropogenic compounds) exist in Massachusetts, and are extremely likely to be present in Massachusetts Bay at varying times and places. However there is unlikely to be data for all of the known or suspected influences and appropriate temporal or spatial scales.

The availability of spatially-relevant environmental data is often a limiting factor when investigating the relationship between environmental factors and human health risks. For example, precipitation and the associated amount of run-off into recreational water bodies can be highly localized and dependent upon local topography and existing infrastructure, but counties or cities containing multiple recreational bathing areas might be served by a single rainfall monitoring station.[190] Investigators working at the

intersection of water quality, public health, and environmental factors have suggested that "future work should investigate whether beach closures due to microbial contamination are more likely at beaches in close proximity to CSOs, downstream of heavily urbanized areas, or nearby agricultural land."[190] In other words, water quality research should consider upstream and on-land influences, a guiding principle of this work.

Examples of Environmental Modeling in the Gulf of Maine. At least one predictive environmental model exists for a marine-sourced microbiological risk to humans in Massachusetts Bay. This example comes from a causative organism of Paralytic Shellfish Poisoning -- the dinoflagellate *Alexandrium fundyense.*[4] *A. fundyense* has a long history in the maritime provinces of Canada and New England region of the United States, but a sophisticated understanding of the organism's lifecycle and its connection to environmental drivers is relatively recent.[191-193] Models to predict *A. fundyense* blooms in the Gulf of Maine have been generated that include factors such as existing field measurements of *A. fundyense* cysts in ocean sediments, circulation models of the waters in the Gulf of Maine, freshwater runoff and associated nutrient concentrations, and water temperature.[4193] These kinds of models that consider the 'whole ecosystem' (where long-term seasonal factors interact with short term environmental conditions) require the ability to incorporate multiple types of data into one model.

Through the extensive literature review that formed the bulk of the chapter we identified known or suspected environmental and socio-economically related influences on the abundance or presence of five specific marine-sourced risks -- which are

representative of larger categories of risk. Table 3-18, below, presents a simplified version of this information in a single matrix. Figure 3-16, below, displays the same information in graphical form, including arrows to represent conceptual interactions among the different influences. For example, among the socio-economic influences listed on the left side of Figure 3-16 both 'Beach attendance' and 'Dog population' are likely influenced by 'Human population in watershed', and these three influences may then influence *Enterococcus* levels (represented as a box in the upper right side). Environmental modeling with the capacity to integrate both ocean and land-based influences on marine-sourced risk would be useful to generating forecasts of environmental conditions relevant to public health.

Table 3-18. Environmental and socioeconomic influences on specific marine-sourced risks

Known or suspected environmental and socioeconomic influences	Environmental influences							Socio-economic influences					
	Water temperature (range)	Sunlight or	Wind direction & Upwelling	Salinity	Rainfall / Freshwater input	Turbidity of water	Other seasonal factors	Human population in watershed	Beach attendance	Dog population	Wastewater inputs & treatment type	Animal agriculture	Nutrient runoff & composition
Enterococcus bacteria	X	X		X	X	X		X	X	X	X	X	X
Vibrio parahaemolyticus	X	X			X		?						X
Pseudo-nitzschia species diatoms	X	X	X	X		X	?					X	X
Hepatitis A Virus	X	X		?		?		X	X		X		
Anthropogenic antibiotics	?	X		?	X	?		X			X	X	

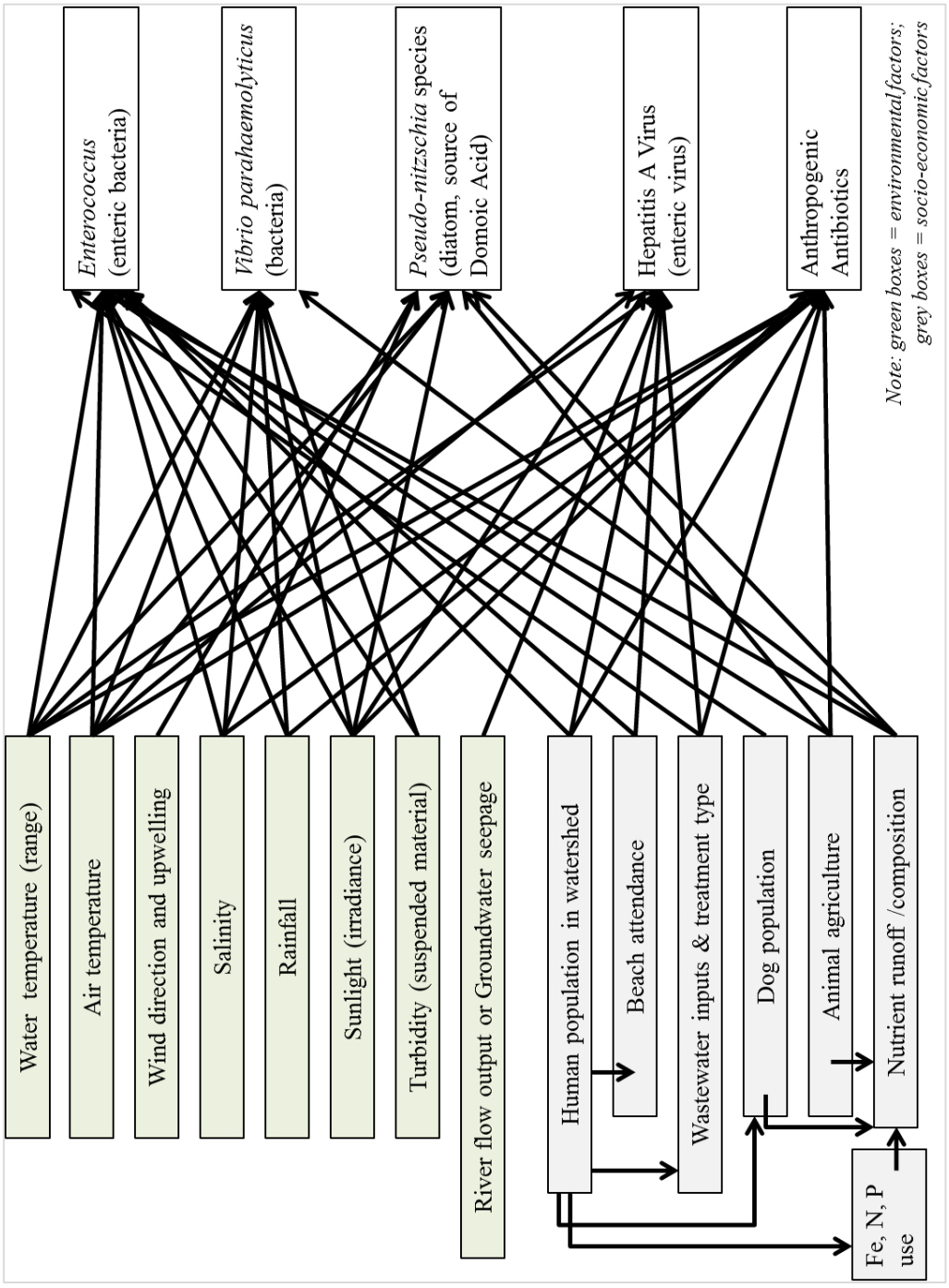

Figure 3-16. Conceptual model of influencing factors on multiple marine-sourced risks that coexist in Massachusetts Bay.

Summary Conclusion.

The current medical reporting system is widely acknowledged to be insufficient for the purpose of capturing the true incidence of many environmentally-driven illnesses.[5; 90] A greater understanding of the risk potential from marine-sources will not come from reliance on traditional epidemiological data. Rather it must come from a greater understanding of the conditions influencing the presence of these risks in their natural environment,[1] and then understanding how these factors intersect with human behavior to influence public health.

The purpose of this chapter was to 1) describe major human demographic characteristics for the six coastal watersheds around Massachusetts Bay, and 2) describe five marine-sourced risks known to exist in Massachusetts Bay, with each example representing a different category of risk. We used spatial analysis software to analyze data from the U.S. Census Bureau and MassGIS. The results of this analysis included original estimates of coastal watershed populations, changes in coastal watershed populations from the year 2000 to 2010, and selected demographic characteristics such as median age, percent of residents under age 5 or over age 65, and average median income. Population growth in the coastal watersheds is similar to the trend for Massachusetts as a whole. However, the Cape Cod watershed had distinctive features compared to the other coastal watersheds including net population loss, the largest percent of population age 65 or over (25%), the lowest percent of population age 5 or under, and the lowest average of

census tract median incomes. Together this suggests that the Cape Cod population would be the most vulnerable to any marine-source risk it encounters.

The marine-sourced risk examples discussed in this chapter represent risk categories including indigenous pathogenic bacteria, introduced enteric bacteria, introduced enteric viruses, natural marine toxins, and anthropogenic pollutants. Each marine-sourced risk example included a list of reported environmental or socioeconomic influences on the presence or persistence of that particular risk in coastal marine waters. The section on anthropogenic antibiotics also includes an original estimate of human antibiotic use in the study area. From this review it is clear that individual variables (e.g., temperature, salinity, and nutrient levels) influence multiple types of marine-sourced risks, this exercise was designed to identify such 'high value' data that could be used in attempts to model multiple risks concurrently.

An increasing human population in Massachusetts Bay coastal watersheds indicates that more people will have close access to ocean-based recreational activities, including sea-bathing. These ocean-based recreational activities, along with the consumption of seafood harvested from nearshore waters, may bring people into contact with a variety of marine-sourced risks. Given that 1) an estimated 5 million cases of gastrointestinal illness due to beach exposure occur in the U.S. every year,[6] 2) the widespread belief among the public health community that marine-sourced illnesses as a whole are currently not well quantified,[5; 6; 94] 3) seafood has the highest rate of foodborne illness in the U.S.,[3] and 4) large numbers of residents and tourists consume seafood from

or directly interact with the nearshore coastal waters of Massachusetts Bay and Cape Cod Bay, it is likely that there are unreported marine-sourced illnesses among residents and visitors to the Massachusetts Bay area. Thus, this is an area of public health research worthy of greater attention. There is a need for better epidemiological data in conjunction with a timely understanding of changing marine-sourced risk potentials to support public health forecasting. Chapter 3 discusses the current data landscape as it relates to environmental-human health research, modeling, and forecasting.

Literature Cited.

1. Doyle, T. J., Glynn, M. K., Groseclose, S. L. 2002. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. Am. J. Epidemiol. 155: 866-874.

2. Centers for Disease Control and Prevention. CDC Estimates of Foodborne Illness in the United States: CDC 2011 Estimates: Findings. Centers for Disease Control and Prevention. Atlanta, GA, USA. http://www.cdc.gov/foodborneburden/2011-foodborne-estimates.html (Accessed April 22, 2015).

3. Centers for Disease Control and Prevention. 2013. Surveillance for foodborne disease outbreaks--United States, 2009-2010. MMWR. 62: 41-47.

4. Bienfang, P. K., DeFelice, S. V., Laws, E. A., Brand, L. E., Bidigare, R. R., Christensen, S., Trapido-Rosenthal, H., Hemscheidt, T. K., McGillicuddy Jr, D. J., Anderson, D. M. 2011. Prominent Human Health Impacts from Several Marine Microbes: History, Ecology, and Public Health Implications. Int. J. Microbiol. 2011.

5. Shuval, H. 2003. Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment. J. Water & Health. 1: 53-64.

6. Ralston, E. P., Kite-Powell, H., Beet, A. 2011. An estimate of the cost of acute health effects from food- and water-borne marine pathogens and toxins in the USA. J. Water & Health. 9: 680-694.

7. Centers for Disease Control and Prevention. 2015. CDC Estimates of Foodborne Illness in the United States: Foodborne Illness Surveillance, Response, and Data Systems. Centers for Disease Control and Prevention. Atlanta, G.A. http://www.cdc.gov/foodborneburden/surveillance-systems.html (Accessed December 13, 2015).

8. United States Department of Health and Human Services, Centers for Disease Control and Prevention. 2011. National Notifiable Diseases Surveillance System: History. http://www.cdc.gov/osels/ph_surveillance/nndss/nndsshis.htm (Accessed 5/17/2011, 2011).

9. Centers for Disease Control and Prevention. 2013. Foodborne Outbreak Tracking and Reporting: Overview. Centers for Disease Control and Prevention. Atlanta, GA. http://www.cdc.gov/foodsafety/fdoss/overview/index.html (Accessed June 22, 2015).

10. Centers for Disease Control and Prevention. 2013. National Enteric Disease Surveillance: COVIS Annual Summary, 2011. Centers for Disease Control and Prevention. Atlanta, GA. 1-10.

11. Centers for Disease Control and Prevention. 2015. Waterborne Disease & Outbreak Surveillance & Reporting. Centers for DIsease Control and Prevention. Atlanta, G.A. http://www.cdc.gov/healthywater/surveillance/ (Accessed December 13, 2015).

12. Hlavsa, M. C., Roberts, V. A., Kahler, A. M., Hilborn, E. D., Wade, T. J., Backer, L. C., Yoder, J. S. 2014. Recreational water–associated disease outbreaks—United States, 2009–2010. MMWR. 63: 6-10.

13. Hlavsa, M. C., Roberts, V. A., Anderson, A. R., Hill, V. R., Kahler, A. M., Orr, M., Garrison, L. E., Hicks, L. A., Newton, A., Hilborn, E. D., Wade, T. J., Beach, M. J., Yoder, J. S. 2011. Surveillance for Waterborne Disease Outbreaks and Other Health Events Associated with Recreational Water--United States, 2007-2008. MMWR. 60: 1-37.

14. Hlavsa, M. C., Roberts, V. A., Kahler, A. M., Hilborn, E. D., Mecher, T. R., Beach, M. J., Wade, T. J., Yoder, J. S. 2015. Outbreaks of illness associated with recreational water—United States, 2011–2012. MMWR. 64: 668-672.

15. Slovic, P. 1987. Perception of risk. Science. 236: 280-285.

16. Collier, S., Stockman, L., Hicks, L., Garrison, L., Zhou, F., Beach, M. 2012. Direct healthcare costs of selected diseases primarily or partially transmitted by water. Epidemiol. Infect. 140: 2003-2013.

17. Commonwealth of Massachusetts, Executive Office of Energy and Environmental Affairs. 2006. Massachusetts Outdoors 2006: Statewide Comprehensive Outdoor Recreation Plan. Commonwealth of Massachusetts. Boston, M.A. 1-166.

18. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Massachusetts Marine Fisheries: 2014 Annual Report. Commonwealth of Massachusetts. Boston, M.A. 1-123.

19. Dufour, A. P., Evans, O., Behymer, T. D., Cantu, R. 2006. Water ingestion during swimming activities in a pool: a pilot study. J. Water Health. 4: 425-430.

20. Evans, O. M., Wymer, L. J., Behymer, T. D., Dufour, A. P. 2006. An observational study determination of the volume of water ingested during recreational swimming activities. National Beaches Conference, Niagara Falls, NY. Vol. 12Niagara Falls, NY.

21. Dorevitch, S., Panthi, S., Huang, Y., Li, H., Michalek, A. M., Pratap, P., Wroblewski, M., Liu, L., Scheff, P. A., Li, A. 2011. Water ingestion during water recreation. Water Res. 45: 2020-2028.

22. Heaney, C. D., Sams, E., Wing, S., Marshall, S., Brenner, K., Dufour, A. P., Wade, T. J. 2009. Contact with beach sand among beachgoers and risk of illness. Am. J. Epidemiol. 170: 164-172.

23. Yamahara, K. M., Sassoubre, L. M., Goodwin, K. D., Boehm, A. B. 2012. Occurrence and persistence of bacterial pathogens and indicator organisms in beach sand along the California coast. Appl. Environ. Microbiol. 78: 1733-1745.

24. Yamahara, K. M., Walters, S. P., Boehm, A. B. 2009. Growth of enterococci in unaltered, unseeded beach sands subjected to tidal wetting. Appl. Environ. Microbiol. 75: 1517-1524.

25. Halliday, E., Gast, R. J. 2011. Bacteria in Beach Sands: An Emerging Challenge in Protecting Coastal Water Quality and Bather Health. Environ. Sci. Technol. 45: 370-379.

26. Commonwealth of Massachusetts, Department of Public Health. 2015. Marine and Freshwater Beach Testing in Massachusetts, Annual Report: 2014 Season. Commonwealth of Massachusetts. Boston, M.A. 1-151.

27. Noble, R. T., Moore, D. F., Leecaster, M. K., McGee, C. D., Weisberg, S. B. 2003. Comparison of total coliform, fecal coliform, and *enterococcus* bacterial indicator response for ocean recreational water quality testing. Water Res. 37: 1637-1643.

28. National Research Council. 2004. Indicators for Waterborne Pathogens. The National Academies Press. Washington, DC. 1-332.

29. Ballester, N., Fontaine, J., Margolin, A. 2005. Occurrence and correlations between coliphages and anthropogenic viruses in the Massachusetts Bay using enrichment and ICC-nPCR. J Water Health. 3: 59-68.

30. Kress, M. M. 2015. Red's Best Seafood Booth at Boston Public Market: Signs of Shellfish For Sale by Month (unpublished photographs). unpublished.

31. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Shellfish Management. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/shellfisheries-management.html#csr (Accessed October 19, 2015).

32. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Public Health Protection: Shellfish Sanitation. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/public-health-protection.html (Accessed October 19, 2015).

33. Fraser, D. 2012. Local oysters blamed for illnesses. Cape Cod Times, Local Media Group, Inc. Hyannis, MA. http://www.capecodtimes.com/article/20121110/NEWS/211100335 (Accessed December 15, 2015).

34. Serreze, M. 2014. Cape Cod oyster 'Vibrio' poisoning case to be heard in Hampshire Superior Court. MassLive LLC. Springfield, MA. http://www.masslive.com/news/index.ssf/2014/12/toxic_cape_cod_oyster_case_to.html (Accessed January 1, 2015).

35. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Vibrio Control Plan. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/vibrio.html (Accessed October 19, 2015).

36. Earle, P. M., Crisley, F. D. 1975. Isolation and characterization of *Vibrio parahaemolyticus* from Cape Cod soft-shell clams (*Mya arenaria*). Appl. Microbiol. 29: 635-640.

37. Bartley, C. H., Slanetz, L. 1971. Occurrence of *Vibrio parahaemolyticus* in Estuarine Waters and Oysters of New Hampshire. Appl. Environ. Microbiol. 21: 965-966.

38. Commonwealth of Massachusetts, Department of Public Health. 2013. Communicable and Other Infectious Diseases Reportable in Massachusetts by Healthcare Providers (PDF File). : 1-2.

39. Commonwealth of Massachusetts, Department of Public Health. 2013. Communicable and Other Infectious Diseases Reportable in Massachusetts by Clinical Laboratories (PDF File). : 1.

40. Commonwealth of Massachusetts, Division of Marine Fisheries. 2012. New Cooling Requirements for Oysters Harvested this Summer from Eastern Cape. DMF News. Massachusetts Division of Marine Fisheries: 1-2.

41. Hunt, C. D., Borkman, D. G., Libby, P. S., Lacouture, R., Turner, J. T., Mickelson, M. J. 2010. Phytoplankton patterns in Massachusetts Bay—1992–2007. Estuaries and Coasts. 33: 448-470.

42. Commonwealth of Massachusetts. 2009. MassGIS Data: Datalayers: Land Use (2005). Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/lus2005.html (Accessed October 30, 2015).

43. Centers for Disease Control and Prevention. 2015. People at High Risk of Developing Flu–Related Complications. Centers for Disease Control and Prevention. Atlanta, Ga. http://www.cdc.gov/flu/about/disease/high_risk.htm (Accessed June 22, 2015).

44. U.S. Census Bureau. What is the Census? U.S. Department of Commerce. Washington, D.C. http://www.census.gov/2010census/about/ (Accessed July 2, 2015).

45. U.S. Census Bureau. 2014. American Community Survey: About. U.S. Census Bureau. Washington, DC. http://www.census.gov/acs/www/about_the_survey/american_community_survey/ (Accessed May 24, 2015).

46. U.S. Census Bureau, Economic Planning and Coordination Division. 2014. Economic Census: About. U.S. Census Bureau. Washington, D.C., USA. http://www.census.gov/econ/census/about/ (Accessed December 1, 2015).

47. U.S. Census Bureau. 2012. Geographic Terms and Concepts - Census Tract. U.S. Department of Commerce. Washington, D.C. https://www.census.gov/geo/reference/gtc/gtc_ct.html (Accessed July 2, 2015).

48. U.S. Census Bureau. 2012. Geographic Terms and Concepts - Block. U.S. Department of Commerce. Washington, D.C. https://www.census.gov/geo/reference/gtc/gtc_block.html (Accessed July 2, 2015).

49. U.S. Census Bureau. 2012. Geographic Terms and Concepts - Block Groups. U.S. Department of Commerce. Washington, D.C. http://www.census.gov/geo/reference/gtc/gtc_bg.html (Accessed July 2, 2015).

50. U.S. Census Bureau. 2015. Maps & Data: TIGER Products. U.S. Department of Commerce. Washington, D.C. http://www.census.gov/geo/maps-data/data/tiger.html (Accessed July 2, 2015).

51. Cutter, S. L., Boruff, B. J., Shirley, W. L. 2003. Social Vulnerability to Environmental Hazards. Social Sci. Quart. 84: 242-261.

52. ESRI. 2012. ArcGIS for Desktop. ArcMap 10.1.

53. White, E. M. 2006. Forests on the Edge: A Case Study of South-Central and Southwest Maine Watersheds. U.S. Department of Agriculture, Pacific Northwest Research Station. Corvallis, OR, USA. 1-21.

54. Commonwealth of Massachusetts. 2009. Office of Geographic and Environmental Information (MassGIS) . Commonwealth of Massachusetts. Massachusetts, USA. http://www.mass.gov/mgis/massgis.htm (Accessed December 10, 2009).

55. Center for Policy Analysis, University of Massachusetts Dartmouth. 2000. Help! Wanted: Cape Cod's Seasonal Workforce. Economic Research Series No. 26. University of Massachusetts Dartmouth. 1-90.

56. U.S. Census Bureau. 2014. 2010 Census Interactive Population Search: Massachusetts. U.S. Department of Commerce. Washington, D.C. http://www.census.gov/2010census/popmap/ipmtext.php?fl=24 .

57. Massachusetts Water Resources Authority. 2012. Water Column Monitoring in Massachusetts Bay: 1992 - 2006.

58. Libby, P. S., Fitzpatrick, M. R., Buhl, R. L., Lescarbeau, G. R., Leo, W. S., Borkman, D. G., Turner, J. T. 2014. Quality assurance project plan (QAPP) for water column monitoring 2014-2016: Tasks 4-7 and 10. Report 2014-01. Massachusetts Water Resources Authority. Boston, M.A. 1-67.

59. Costa, A., Larson, E., Stamieszkin, K. 2014. Quality Assurance Project Plan (QAPP) for Water Quality Monitoring in Cape Cod Bay 2014-2016. Report 2014-07. Massachusetts Water Resources Authority. Boston, M.A. 1-94.

60. Libby, P. S., Borkman, D. G., Geyer, W. R., Turner, J. T., Costa, A. S. 2014. 2013 Water Column Monitoring Results. Report 2014-17. Massachusetts Water Resources Authority. Boston, MA. 1-43.

61. Massachusetts Water Resources Authority. 2015. Boston Harbor and Massachusetts Bay: Water Quality Data
. Massachusetts Water Resources Authority. Boston, MA. http://www.mwra.state.ma.us/harbor/html/wq_data.htm (Accessed May 17, 2015).

62. Commonwealth of Massachusetts. 2011. MassGIS Data: Datalayers: MassDEP Ground Water Discharge Permits. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/gwp.html (Accessed October 30, 2015).

63. Commonwealth of Massachusetts. 2015. MassGIS Data: Datalayers:Designated Shellfish Growing Areas. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/dsga.html (Accessed November 7, 2015).

64. Commonwealth of Massachusetts. 2013. MassGIS Data: Datalayers: Marine Beaches. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/marinebeaches.html (Accessed October 30, 2015).

65. United States Environmental Protection Agency (U.S.EPA). 2012. Combined Sewer Overflows. USEPA. Washington, D.C. http://cfpub.epa.gov/npdes/home.cfm (Accessed 03/24, 2014).

66. Rattigan, D. 2013. Sewer overflow triggers closure of Gloucester beach Boston Globe Media Partners, LLC. Boston, M.A. http://www.boston.com/yourtown/news/gloucester/2013/06/sewer_overflow_triggers_closure_of_gloucester_beach.html (Accessed March 24, 2014).

67. United States Environmental Protection Agency (U.S.EPA), Region 1. 2013. Reducing Combined Sewer Overflows to Charles River. USEPA. http://cfpub.epa.gov/npdes/home.cfm?program_id=5 (Accessed 03/24, 2014).

68. Revere Journal Staff. 2012. Sewerage Dispute. Revere JournalThe Independent Newspaper Group. 385 Broadway, Suite 105 in the Citizens Bank Building, Revere, MA 02151.

69. Griffin, D. W., Donaldson, K. A., Paul, J. H., Rose, J. B. 2003. Pathogenic Human Viruses in Coastal Waters. Clin. Microbiol. Rev. 16: 129-143.

70. Sham, C. H., Brawley, J. W., Moritz, M. A. 1995. Quantifying septic nitrogen loadings to receiving waters: Waquoit Bay, Massachusetts. International Journal of Geographical Information Systems. 9: 463-473.

71. Lipp, E. K., Kurz, R., Vincent, R., Rodriguez-Palacios, C., Farrah, S. R., Rose, J. B. 2001. The Effects of Seasonal Variability and Weather on Microbial Fecal Pollution and Enteric Pathogens in a Subtropical Estuary. Estuaries. 24: 266-276.

72. Muddy River Restoration Project Maintenance and Management Oversight Committee. 2015. Muddy River Restoration Project: Flood Control Improvement. Muddy River Restoration Project Maintenance and Management Oversight Committee. Boston, M.A. http://www.muddyrivermmoc.org/flood-control/ (Accessed 11/29, 2015).

73. Lleò, M. M., Bonato, B., Benedetti, D., Canepari, P. 2005. Survival of enterococcal species in aquatic environments. FEMS Microbiol. Ecol. 54: 189-196.

74. Fisher, K., Phillips, C. 2009. The ecology, epidemiology and virulence of *Enterococcus*. Microbiology. 155: 1749-1757.

75. Prüss, A. 1998. Review of epidemiological studies on health effects from exposure to recreational water. Int. J. Epidemiol. 27: 1-9.

76. World Health Organization. 1999. Health-based Monitoring of Recreational Waters: The Feasibility of a New Approach (The 'Annapolis Protocol'). Outcome of an Expert Consultation, Annapolis, USA Co-sponsored by USEPA. WHO/SDE/WSH/99.1. World Health Organization. Geneva, Switzerland. 1-50.

77. Mote, B. L., Turner, J. W., Lipp, E. K. 2012. Persistence and growth of the fecal indicator bacteria enterococci in detritus and natural estuarine plankton communities. Appl. Environ. Microbiol. 78: 2569-2577.

78. Byappanahalli, M. N., Nevers, M. B., Korajkic, A., Staley, Z. R., Harwood, V. J. 2012. Enterococci in the environment. Microbiol. Mol. Biol. Rev. 76: 685-706.

79. Anderson, K. L., Whitlock, J. E., Harwood, V. J. 2005. Persistence and Differential Survival of Fecal Indicator Bacteria in Subtropical Waters and Sediments. Appl. Environ. Microbiol. 71: 3041-3048.

80. Alkan, U., Elliott, D. J., Evison, L. M. 1995. Survival of enteric bacteria in relation to simulated solar radiation and other environmental factors in marine waters. Water Res. 29: 2071-2080.

81. Shibata, T., Solo-Gabriele, H. M., Sinigalliano, C. D., Gidley, M. L., Plano, L. R. W., Fleisher, J. M., Wang, J. D., Elmir, S. M., He, G., Wright, M. E. 2010. Evaluation of conventional and alternative monitoring methods for a recreational marine beach with nonpoint source of fecal contamination. Environ. Sci. Technol. 44: 8175-8181.

82. Bonilla, T. D., Nowosielski, K., Cuvelier, M., Hartz, A., Green, M., Esiobu, N., McCorquodale, D. S., Fleisher, J. M., Rogerson, A. 2007. Prevalence and distribution of fecal indicator organisms in South Florida beach sand and preliminary assessment of health effects associated with beach sand exposure. Mar. Pollut. Bull. 54: 1472-1482.

83. Goodwin, K. D., McNay, M., Cao, Y., Ebentier, D., Madison, M., Griffith, J. F. 2012. A multi-beach study of *Staphylococcus aureus*, MRSA, and enterococci in seawater and beach sand. Water Res. 46: 4195-4207.

84. Shah, A. H., Abdelzaher, A. M., Phillips, M., Hernandez, R., Solo-Gabriele, H. M., Kish, J., Scorzetti, G., Fell, J. W., Diaz, M. R., Scott, T. M. 2011. Indicator microbes correlate with pathogenic bacteria, yeasts and helminthes in sand at a subtropical recreational beach site. J. Appl. Microbiol. 110: 1571-1583.

85. Bonilla, T. D., Nowosielski, K., Esiobu, N., McCorquodale, D. S., Rogerson, A. 2006. Species assemblages of *Enterococcus* indicate potential sources of fecal bacteria at a south Florida recreational beach. Mar. Pollut. Bull. 52: 807-810.

86. Tyagi, P., Edwards, D., Coyne, M. 2009. Distinguishing between human and animal sources of fecal pollution in waters: a review. Int. J. Water. 5: 15-34.

87. Joseph, S. W., Colwell, R. R., Kaper, J. B. 1982. *Vibrio parahaemolyticus* and related halophilic Vibrios. Crit. Rev. Microbiol. 10: 77-124.

88. Kaysner, C. A., DePaola, J., Angelo. 2004. Chapter 9: *Vibrio* (web version) . *In* Bacteriological Analytical Manual. Hammack T., Davidson, M., Feng, P. *et al*, Eds.U.S. Food and Drug Administration. Washington, D.C.

89. Thompson, C. A., Vanderzant, C. 1976. Relationship of *Vibrio parahaemolyticus* in Oysters, Waters and Sediment, and Bacteriological and Environmental Indices. J. Food Sci. 41: 117-122.

90. Iwamoto, M., Ayers, T., Mahon, B. E., Swerdlow, D. L. 2010. Epidemiology of seafood-associated infections in the United States. Clin. Microbiol. Rev. 23: 399.

91. DePaola, A., Jones, J. L., Woods, J., Burkhardt, W.,3rd, Calci, K. R., Krantz, J. A., Bowers, J. C., Kasturi, K., Byars, R. H., Jacobs, E., Williams-Hill, D., Nabe, K. 2010. Bacterial and viral pathogens in live oysters: 2007 United States market survey. Appl. Environ. Microbiol. 76: 2754-2768.

92. Centers for Disease Control and Prevention (CDC). 2005. *Vibrio* Illnesses After Hurricane Katrina --- Multiple States, August--September 2005. MMWR. 54: 1-4.

93. Centers for Disease Control and Prevention. 2014. Food Safety Progress Report for 2013. Centers for Disease Control and Prevention. Atlanta, GA. 1.

94. Centers for Disease Control and Prevention. 2013. *Vibrio* Illness (Vibriosis):*Vibrio parahaemolyticus*
. Centers for Disease Control and Prevention. Atlanta, G.A. http://www.cdc.gov/vibrio/vibriop.html (Accessed April 2, 2015).

95. U.S. Census Bureau. 2013. Income, Poverty, and Health Insurance Coverage: 2012 - Tables & Figures. Number and Percentage of People Without Health Insurance Coverage by State Using 2- and 3-Year Averages: 2009-2010 and 2011-2012 [Spreadsheet File]. U.S. Census Bureau. Washington, D.C. 1.

96. Kelly, M. T., Stroh, E. M. 1988. Temporal relationship of *Vibrio parahaemolyticus* in patients and the environment. J. Clin. Microbiol. 26: 1754-1756.

97. 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 Res. Int. 43: 1780-1790.

98. Krantz, G. E., Colwell, R. R., Lovelace, E. 1969. *Vibrio parahaemolyticus* from the blue crab *Cailinectes sapidus* in Chesapeake Bay. Science. 164: 1286-1287.

99. Colwell, R., Kaper, J., Joseph, S. 1977. *Vibrio cholerae*, *Vibrio parahaemolyticus*, and Other Vibrios: Occurrence and Distribution in Chesapeake Bay. Science. 198: 394-396.

100. Kaneko, T., Colwell, R. R. 1978. Annual Cycle of *Vibrio Parahaemolyticus* in Chesapeake Bay. Microb. Ecol. 4: 135-155.

101. Kaneko, T., Colwell, R. R. 1973. Ecology of *Vibrio parahaemolyticus* in Chesapeake Bay. J. Bacteriol. 113: 24-32.

102. Johnson, C. N., Bowers, J. C., Griffitt, K. J., Molina, V., Clostio, R. W., Pei, S., Laws, E., Paranjpye, R. N., Strom, M. S., Chen, A., Hasan, N. A., Huq, A., Noriea, N. F.,3rd, Grimes, D. J., Colwell, R. R. 2012. Ecology of *Vibrio parahaemolyticus* and *Vibrio vulnificus* in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washington (United States). Appl. Environ. Microbiol. 78: 7249-7257.

103. Parveen, S., Hettiarachchi, K. A., Bowers, J. C., Jones, J. L., Tamplin, M. L., McKay, R., Beatty, W., Brohawn, K., DaSilva, L. V., DePaola, A. 2008. Seasonal distribution of total and pathogenic *Vibrio parahaemolyticus* in Chesapeake Bay oysters and waters. Int. J. Food Microbiol. 128: 354-361.

104. Gode-Potratz, C. J., Chodur, D. M., McCarter, L. L. 2010. Calcium and Iron Regulate Swarming and Type III Secretion in *Vibrio parahaemolyticus* . J. Bacteriol. 192: 6025-6038.

105. Daniels, N. A., MacKinnon, L., Bishop, R., Altekruse, S., Ray, B., Hammond, R. M., Thompson, S., Wilson, S., Bean, N. H., Griffin, P. M. 2000. *Vibrio parahaemolyticus* infections in the United States, 1973–1998. J. Infect. Dis. 181: 1661.

106. Su, Y., Liu, C. 2007. *Vibrio parahaemolyticus*: A concern of seafood safety. Food Microbiol. 24: 549-558.

107. Cabrera-Garcia, M. E., Vazquez-Salinas, C., Quinones-Ramirez, E. I. 2004. Serologic and molecular characterization of *Vibrio parahaemolyticus* strains isolated from seawater and fish products of the Gulf of Mexico. Appl. Environ. Microbiol. 70: 6401-6406.

108. Hlady, W. G., Klontz, K. C. 1996. The epidemiology of Vibrio infections in Florida, 1981–1993. J. Infect. Dis. 173: 1176-1183.

109. Commonwealth of Massachusetts, Department of Public Health. 2012. Vibrio Control and Management in Eastern Cape Cod Bay: Information for Professional Oyster Harvesters. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eohhs/docs/dph/environmental/foodsafety/seafood/vibrio-control-brochure.pdf (Accessed December 27, 2012).

110. Anderson, C. R., Sapiano, M. R. P., Prasad, M., Long, W., Tango, P. J., Brown, C. W., Murtugudde, R. 2010. Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the Chesapeake Bay. J. Mar. Syst. 83: 127-140.

111. Silver, M. W., Bargu, S., Coale, S. L., Benitez-Nelson, C. R., Garcia, A. C., Roberts, K. J., Sekula-Wood, E., Bruland, K. W., Coale, K. H. 2010. Toxic diatoms and domoic acid in natural and iron enriched waters of the oceanic Pacific. Proc. Natl. Acad. Sci. U. S. A. 107: 20762-20767.

112. Pulido, O. M. 2008. Domoic acid toxicologic pathology: a review. Marine drugs. 6: 180-219.

113. Teitelbaum, J. S., Zatorre, R. J., Carpenter, S., Gendron, D., Evans, A. C., Gjedde, A., Cashman, N. R. 1990. Neurologic sequelae of domoic acid intoxication due to the ingestion of contaminated mussels. N. Engl. J. Med. 322: 1781-1787.

114. Lelong, A., Hégaret, H., Soudant, P., Bates, S. 2012. *Pseudo-nitzschia* species, domoic acid and amnesic shellfish poisoning: revisiting previous paradigms. 51: 168-216.

115. Fernandes, L. F., Hubbard, K. A., Richlen, M. L., Smith, J., Bates, S. S., Ehrman, J., Léger, C., Mafra, L. L., Kulis, D., Quilliam, M. 2014. Diversity and toxicity of the diatom *Pseudo-nitzschia* Peragallo in the Gulf of Maine, Northwestern Atlantic Ocean. Deep Sea Research Part II: Topical Studies in Oceanography. 103: 139-162.

116. Angus, T. H. 2015. Examining the Toxicity, Exposure, and Regulatory Approach to Potential Human Health Risks of the Algal Toxin Domoic Acid. Doctor of Philosophy thesis, University of Massachusetts Boston, Boston, MA.

117. Food and Agriculture Organization of the United Nations (FAO). 2004. Chapter 4 Amnesic Shellfish Poisoning. *In* Marine Biotoxins: FAO Food and Nutrition Paper 80. Anonymous : 97-133. Food and Agriculture Organization of the United Nations (FAO). Rome, Italy.

118. Ajani, P., Murray, S., Hallegraeff, G., Lundholm, N., Gillings, M., Brett, S., Armand, L. 2013. The diatom genus *Pseudo-nitzschia* (Bacillariophyceae) in New South Wales, Australia: morphotaxonomy, molecular phylogeny, toxicity and distribution. J. Phycol. 49: 765-785.

119. Aktan, Y. 2005. Toxic and harmful algal species in the Izmit Bay, Marmara Sea. Harmful algae news. 28: 6.

120. Amzil, Z., Fresnel, J., Le Gal, D., Billard, C. 2001. Domoic acid accumulation in French shellfish in relation to toxic species of *Pseudo-nitzschia* multiseries and *P. pseudodelicatissima*. Toxicon. 39: 1245-1251.

121. Anderson, D. M., Reguera, B., Pitcher, G. C., Enevoldsen, H. O. 2010. The IOC International Harmful Bloom Program: History and Science Impacts. Oceanography. 23: 72-85.

122. Backer, L., McGillicuddy, D. 2006. Harmful algal blooms. Oceanography. 19: 94.

123. Costa, P. R., Rosa, R., Duarte-Silva, A., Brotas, V., Sampayo, M. A. M. 2005. Accumulation, transformation and tissue distribution of domoic acid, the amnesic shellfish poisoning toxin, in the common cuttlefish,*Sepia officinalis*. Aquatic toxicology. 74: 82-91.

124. Maucher, J. M., Ramsdell, J. S. 2005. Domoic acid transfer to milk: evaluation of a potential route of neonatal exposure. Environ. Health Perspect. 113: 461.

125. Maucher, J. M., Ramsdell, J. S. 2007. Maternal–fetal transfer of domoic acid in rats at two gestational time points. Environ. Health Perspect. 115: 1743.

126. Maucher Fuquay, J., Muha, N., Wang, Z., Ramsdell, J. S. 2012. Toxicokinetics of domoic acid in the fetal rat. Toxicology. 294: 36-41.

127. Trainer, V. L., Suddleson, M. 2005. Monitoring Approaches for Early Warning of Domoic Acid Events in Washington State. Oceanography. 18: 228-237.

128. Center for Environment, Fisheries, and Aquaculture Science. 2007. Marine Microbial Communities in UK Waters from Phylogenic Studies to Remote Studies.

129. Kelly, M., Fraser, S. 1999. Toxic algal monitoring in Scotland 1998. FRS Report. 8: 99.

130. Downes-Tettmar, N., Rowland, S., Miller, P., Llewellyn, C. 2013. Seasonal variation in *Pseudo-nitzschia* spp. and domoic acid production in the Western English Channel. Continental Shelf Research. 53: 40-49.

131. Lane, J. Q., Raimondi, P. T., Kudela, R. M. 2009. Development of a logistic regression model for the prediction of toxigenic *Pseudo-nitzschia* blooms in Monterey Bay, California. Mar. Ecol. Prog. Ser. 383: 37-51.

132. Massachusetts Water Resources Authority. . 2015. pseudonitz_1992-2014.xlsx [MS Excel file]. M. Kress.

133. Northeastern Regional Association of Coastal and Ocean Observing Systems. 2014. NERACOOS: Data & Tools. The Gulf of Maine Research Institute. Portland, ME. http://neracoos.org/datatools (Accessed May 25, 2015).

134. Bates, S. S., Garrison, D. L., Horner, R. A. 1998. Bloom dynamics and physiology of domoic-acid-producing *Pseudo-nitzschia* species. *In* Physiological Ecology of Harmful Algal Blooms. Anderson D. M., Cembella, A. D., Hallegraeff, G. M., Eds.: 267-292. Springer Verlag. Berlin, Germany.

135. Lillick, L. C. 1937. Seasonal studies of the phytoplankton off Woods Hole, Massachusetts. Biol. Bull. 73: 488-503.

136. Villareal, T. A., Roelke, D. L., Fryxell, G. A. 1994. Occurrence of the toxic diatom *Nitzschia pungens f. multiseries* in Massachusetts Bay, Massachusetts, USA. Mar. Environ. Res. 37: 417-423.

137. Sharapov, U. M., Teshale, E. H., Centers for Disease Control and Prevention. 2013. Chapter 3: Infectious Diseases Related To Travel: Hepatitis A . *In* CDC Health Information for International Travel 2014 (the Yellow Book). Brunette G. W., Centers for Disease Control and Prevention, Eds.: 1-688. Centers for Disease Control and Prevention. Atlanta, GA, USA.

138. Jacobsen, K. H., Wiersma, S. T. 2010. Hepatitis A virus seroprevalence by age and world region, 1990 and 2005. Vaccine. 28: 6653-6657.

139. Pond, K. 2005. Water Recreation and Disease: Plausibility of Associated Infections: Acute Effects, Sequelae, and Mortality. World Health Organization. London, UK.

140. Kotwal, G., Cannon, J. L.,. 2014. Environmental persistence and transfer of enteric viruses. Curr. Opin. Virology. 4: 37-43.

141. Provost, K., Dancho, B. A., Ozbay, G., Anderson, R. S., Richards, G. P., Kingsley, D. H. 2011. Hemocytes are sites of enteric virus persistence within oysters. Appl. Environ. Microbiol. 77: 8360-8369.

142. Shieh, Y., Khudyakov, Y., Xia, G., Ganova-Raeva, L., Khambaty, F., Woods, J., Veazey, J., Motes, M., Glatzer, M., Bialek, S. 2007. Molecular confirmation of oysters as the vector for hepatitis A in a 2005 multistate outbreak. J. Food Protection. 70: 145-150.

143. FitzSimons, D., Hendrickx, G., Vorsters, A., Van Damme, P. 2010. Hepatitis A and E: Update on Prevention and Epidemiology. Vaccine. 28: 583-588.

144. Yanez, L. A., Lucero, N. S., Barril, P. A., Diaz, M. d. P., Tenaglia, M. M., Spinsanti, L. I., Nates, S. V., Isa, M. B., Re, V. E. 2014. Evidence of Hepatitis A virus circulation in central Argentina: Seroprevalence and environmental surveillance. Journal of Clinical Virology. 59: 38-43.

145. Nelson, N. P., Murphy, T. V. 2013. Hepatitis A: The Changing Epidemiology of Hepatitis A. Clin. Liver Dis. 2: 227-230.

146. Taylor, M. B., Cox, N., Vrey, M. A., Grabow, W. O. K. 2001. The occurrence of hepatitis A and astroviruses in selected river and dam waters in South Africa. Water Res. 35: 2653-2660.

147. Grady, G. F., Chalmers, T. G., Boston Inter-Hospital Liver Group. 1965. Viral Hepatitis in a Group of Boston Hospitals: A Prospective Controlled Epidemiologic Study. N. Engl. J. Med. 272: 662-666.

148. Koff, R. S., Grady, G. F., Chalmers, T. C., Mosley, J. W., Swartz, B. L., Boston Inter-Hospital Liver Group. 1967. Viral Hepatitis in a Group of Boston Hospitals: III. Importance of Exposure to Shellfish in a Nonepidemic Period. N. Engl. J. Med. 276: 703-710.

149. Koff, R. S., Isselbacher, K. J. 1968. Changing concepts in the epidemiology of viral hepatitis. N. Engl. J. Med. 278: 1371-1380.

150. Callahan, K. M., Taylor, D. J., Sobsey, M. D. 1995. Comparative survival of hepatitis A virus, poliovirus and indicator viruses in geographically diverse seawaters. Water Science and Technology. 31: 189-193.

151. Borchardt, M. A., Bertz, P. D., Spencer, S. K., Battigelli, D. A. 2003. Incidence of enteric viruses in groundwater from household wells in Wisconsin. Appl. Environ. Microbiol. 69: 1172-1180.

152. Gerba, C. P. 2007. Chapter 5: Virus Occurrence and Survival in the Environmental Waters. *In* Human Viruses in Water. Bosch A., Ed.: 91-108. Elsevier. Philadelphia, P.A.

153. John, D. E., Rose, J. B. 2005. A review of factors affecting microbial survival in ground water. Environ Sci Technol. 39: 7345-7356.

154. On-line Medical Dictionary. Isoelectric Point. The Joint Center for Structural Genomics. http://www.jcsg.org/help/robohelp/Definitions/Isoelectric_Point.htm (Accessed March 31, 2014).

155. Crance, J., Gantzer, C., Schwartzbrod, L., Deloince, R. 1998. Effect of temperature on the survival of hepatitis A virus and its capsidal antigen in synthetic seawater. Environ. Toxicol. Water Qual. 13: 89-92.

156. Centers for Disease Control and Prevention (CDC). 2014. Recommended immunization schedule for persons aged 0 through 18 years – United States, 2014. : 1-4.

157. Massachusetts Department of Public Health. 2015. Massachusetts School Immunization Requirements for School Year 2015-2016. : 1.

158. Fiore, A., Bell, B., Barker, L., Darling, N., Amon, J., Centers for Disease Control and Prevention. 2005. Hepatitis A Vaccination Coverage Among Children Aged 24-35 Months--United States, 2003. MMWR. 54: 141-144.

159. Hill, H. A., Elam-Evans, L. D., Yankey, D., Singleton, J. A., Kolasa, M. 2014. National, state, and selected local area vaccination coverage among children aged 19–35 months-United States, 2013. MMWR. 63: 741-748.

160. Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. 2011. Massachusetts--2010 Profile. CS212259-A. Centers for Disease Control and Prevention. 1600 Clifton Rd. Atlanta, GA 30333, USA. 1-2.

161. Massachusetts Department of Public Health. 2013. Enteric Disease in Massachusetts: 1999-2013. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eohhs/docs/dph/cdc/foodsafety-enterics-state-totals.pdf (Accessed March 17, 2015).

162. Dorell, C. G., Yankey, D., Byrd, K. K., Murphy, T. V. 2012. Hepatitis A Vaccination Coverage Among Adolescents in the United States. Pediatrics. 129: 213-221.

163. Williams, W. W., Lu, P. J., O'Halloran, A., Bridges, C. B., Kim, D. K., Pilishvili, T., Hales, C. M., Markowitz, L. E., Centers for Disease Control and Prevention (CDC). 2015. Vaccination coverage among adults, excluding influenza vaccination - United States, 2013. MMWR. 64: 95-102.

164. Centers for Disease Control and Prevention. 2015. Antibiotic/Antimicrobial Resistance: About Antimicrobial Resistance. U.S. Department of Health and Human Services. Atlanta, G.A. http://www.cdc.gov/drugresistance/about.html (Accessed October 16, 2015).

165. Shapiro, D. J., Hicks, L. A., Pavia, A. T., Hersh, A. L. 2014. Antibiotic prescribing for adults in ambulatory care in the USA, 2007-09. J. Antimicrob. Chemother. 69: 234-240.

166. Smith DeWaal, C., Roberts, C., Catella, C. 2012. Antibiotics Resistance in Foodborne Pathogens: Evidence of the Need for a Risk Management Strategy. : 1-18.

167. Centers for Disease Control and Prevention. 2014. Antibiotic/Antimicrobial Resistance: Antibiotic Resistance Threats in the United States, 2013. U.S. Department of Health and Human Services. Atlanta, G.A. http://www.cdc.gov/drugresistance/threat-report-2013/ (Accessed October 16, 2015).

168. Centers for Disease Control and Prevention. 2015. National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS): Human Isolates Final Report, 2013. U.S. Department of Health and Human Services, CDC. Atlanta, G.A. 1-81.

169. Karthikeyan, K., Meyer, M. T. 2006. Occurrence of antibiotics in wastewater treatment facilities in Wisconsin, USA. Sci. Total Environ. 361: 196-207.

170. Batt, A. L., Kim, S., Aga, D. S. 2007. Comparison of the occurrence of antibiotics in four full-scale wastewater treatment plants with varying designs and operations. Chemosphere. 68: 428-435.

171. Watkinson, A., Murby, E., Costanzo, S. 2007. Removal of antibiotics in conventional and advanced wastewater treatment: implications for environmental discharge and wastewater recycling. Water Res. 41: 4164-4176.

172. Brown, K. D., Kulis, J., Thomson, B., Chapman, T. H., Mawhinney, D. B. 2006. Occurrence of antibiotics in hospital, residential, and dairy effluent, municipal wastewater, and the Rio Grande in New Mexico. Sci. Total Environ. 366: 772-783.

173. Gulkowska, A., Leung, H. W., So, M. K., Taniyasu, S., Yamashita, N., Yeung, L. W., Richardson, B. J., Lei, A., Giesy, J. P., Lam, P. K. 2008. Removal of antibiotics from wastewater by sewage treatment facilities in Hong Kong and Shenzhen, China. Water Res. 42: 395-403.

174. Su, H., Ying, G., He, L., Liu, Y., Zhang, R., Tao, R. 2014. Antibiotic resistance, plasmid-mediated quinolone resistance (PMQR) genes and *ampC* gene in two typical municipal wastewater treatment plants. Environ. Sci. Processes Impacts. 16: 324-332.

175. Rizzo, L., Manaia, C., Merlin, C., Schwartz, T., Dagot, C., Ploy, M. C., Michael, I., Fatta-Kassinos, D. 2013. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: A review. Sci. Total Environ. 447: 345-360.

176. Varela, A. R., Ferro, G., Vredenburg, J., Yanik, M., Vieira, L., Rizzo, L., Lameiras, C., Manaia, C. M. 2013. Vancomycin resistant enterococci: From the hospital effluent to the urban wastewater treatment plant. Sci. Total Environ. 450: 155-161.

177. Fuentefria, D. B., Ferreira, A. E., Corção, G. 2011. Antibiotic-resistant *Pseudomonas aeruginosa* from hospital wastewater and superficial water: Are they genetically related? J. Environ. Manage. 92: 250-255.

178. Akiyama, T., Savin, M. C. 2010. Populations of antibiotic-resistant coliform bacteria change rapidly in a wastewater effluent dominated stream. Sci. Total Environ. 408: 6192-6201.

179. Faria, C., Vaz-Moreira, I., Serapicos, E., Nunes, O. C., Manaia, C. M. 2009. Antibiotic resistance in coagulase negative staphylococci isolated from wastewater and drinking water. Sci. Total Environ. 407: 3876-3882.

180. LaPara, T. M., Burch, T. R., McNamara, P. J., Tan, D. T., Yan, M., Eichmiller, J. J. 2011. Tertiary-treated municipal wastewater is a significant point source of antibiotic resistance genes into Duluth-Superior Harbor. Environ. Sci. Technol. 45: 9543-9549.

181. Munir, M., Wong, K., Xagoraraki, I. 2011. Release of antibiotic resistant bacteria and genes in the effluent and biosolids of five wastewater utilities in Michigan. Water Res. 45: 681-693.

182. Food & Water Watch. 2014. Factory Farm Map. Food & Water Watch. Washington, D.C., USA. http://www.factoryfarmmap.org/#animal:all;location:MA;year:2007 (Accessed 30 August, 2014).

183. U.S. Census Bureau, U.S. Department of Commerce. 2014. The 2012 Statistical Abstract. U.S. Department of Commerce. Washington, D.C., USA. http://www.census.gov/compendia/statab/cats/agriculture.html (Accessed August 31, 2014).

184. Pham, T., U.S. Food and Drug Administration. 2012. Drug Use Review: Systemic Antibacterial Drug Products. OSE RCM # 2012-544. U.S. Food and Drug Administration. Washington, D.C., USA. 1-9.

185. Hicks, L. A., Taylor Jr, T. H., Hunkler, R. J. 2013. U.S. outpatient antibiotic prescribing, 2010. N. Engl. J. Med. 368: 1461-1462.

186. Kümmerer, K. 2009. Antibiotics in the aquatic environment – A review – Part I. Chemosphere. 75: 417-434.

187. U.S. Food and Drug Administration. 2011. Summary Report On Antimicrobials Sold or Distributed for Use in Food-Producing Animals. U.S. Food and Drug Administration. Washington, D.C., USA. 1-4.

188. Massachusetts Water Resources Authority. 2009. The Deer Island Sewage Treatment Plant. Massachusetts Water Resources Authority. Boston, M.A. http://www.mwra.com/03sewer/html/sewditp.htm (Accessed October 18, 2015).

189. Meadows, D. H. 2008. Thinking in Systems: A Primer. Chelsea Green Publishing. White River Junction, V.T.

190. Bush, K. F., Fossani, C. L., Li, S., Mukherjee, B., Gronlund, C. J., O'Neill, M. S. 2014. Extreme Precipitation and Beach Closures in the Great Lakes Region: Evaluating Risk among the Elderly. Int. J. Environ. Res. Public Health. 11: 2014-2032.

191. Anderson, D. M., Keafer, B. A., McGillicuddy, D. J., Mickelson, M. J., Keay, K. E., Libby, P. S., Manning, J. P., Mayo, C. A., Whittaker, D. K., Hickey, J. M. 2005. Initial observations of the 2005 *Alexandrium fundyense* bloom in southern New England: General patterns and mechanisms. Deep Sea Research Part II: Topical Studies in Oceanography. 52: 2856-2876.

192. Anderson, D. M., Stock, C. A., Keafer, B. A., Nelson, A. B., Thompson, B., McGillicuddy, D. J., Keller, M., Matrai, P. A., Martin, J. 2005. *Alexandrium fundyense* cyst dynamics in the Gulf of Maine. Deep Sea Research Part II: Topical Studies in Oceanography. 52: 2522-2542.

193. McGillicuddy, D. J., Anderson, D. M., Lynch, D. R., Townsend, D. W. 2005. Mechanisms regulating large-scale seasonal fluctuations in *Alexandrium fundyense* populations in the Gulf of Maine: results from a physical–biological model. Deep Sea Research Part II: Topical Studies in Oceanography. 52: 2698-2714.

CHAPTER 4

INTERDISCIPLINARY DATA SCIENCE

Abstract. This chapter discusses how inter-disciplinary questions in environmental health and infectious disease research may be addressed through the use of data beyond traditional medical and epidemiological sources. The first section of this chapter discusses technologically-driven changes in data availability and data-sourcing as well as the emerging discipline of 'data science,' both important topics for synthesis-type research that seeks to gain extra utility from existing data. This section also discusses potential risks from large datasets-of-opportunity (rather than those designed to answer a specific question) and provides an illustrative example. The second section of this chapter presents a generalized workflow for interdisciplinary environmental health research. This proposed generalized workflow is supported three examples of research that successfully combined traditional epidemiological data with non-traditional remote sensing data to address environmental health questions in different parts of the world. The third and final section of this chapter poses four environmental health questions relevant to two marine-sourced risks in Massachusetts Bay (*Pseudo-nitzschia* genus

diatoms and *Enterococcus* genus bacteria). The chapter closes with a description of data collected from authoritative public sources to support the development of predictive models for those two marine-sourced human health risks.

Introduction.

The world is awash in data. Raw data from *in situ*, *in vitro* or *in silico* experiments, field data, aggregated data, anonymized data, crowdsourced data, synthetic data and big data are all options that researchers today can utilize. Multiple agencies, institutions, and individual researchers collect data, be it environmental, medical, social, or other, to answer a specific question within their focus area. A constellation of data sources may exist for any single topic- the challenge lies in combining and interpreting these data in order to chart the best research course. In some cases these data are made available to the public, either in raw, aggregated, and/or interpreted form for others to use. Combining data from multiple disparate sources in order to performance new research is the essence of a synthesis, understood as "the combining of often diverse conceptions into a coherent whole."[1] For interdisciplinary research, such as environmental health research, a synthesis approach is almost required by definition.

This dissertation uses an interdisciplinary synthesis approach to explore the topic of ocean and human health. This chapter on data science is divided into three sections. Section one of this chapter discussed the limitations that come with using data originally collected for other purposes. The expanding opportunities to using this type of data require careful consideration. Section one also provides examples of data available to

environmental health researchers from established and emerging sources. Although not exhaustive, this list is representative of the variety of data available to interdisciplinary environmental researchers and may represent a useful starting point for those interested in cross-disciplinary work.

The second section of this chapter opens with a 3-phase diagram of a generalized workflow process for researchers interested in pursuing similar interdisciplinary environmental health research. Briefly, the three phases is this diagram are 1) explore concepts and generate hypothesis, 2) develop outputs, and 3) evaluate outputs. This approach aligns well with the traditional scientific method, but is not necessarily bound by the confines of null hypothesis testing. We suggest that researchers who may not consider themselves 'interdisciplinary data scientists' could use such a framework to guide collaborative efforts with specialists in other fields. To this end we provide examples of research on three distinct environmental health topics that followed a similar generalized workflow process to engage in environmental health research. Those examples are predicting Rift Valley Fever risk area in the Horn of Africa, modeling cholera outbreaks in Bangladesh, and investigating the causative agent of Kawasaki Disease in Japan.

Section three of this chapter describes environmental and socio-economic data sets related to interdisciplinary research on marine-sourced risks in Massachusetts Bay. In this section we pose four questions related to understanding the presence of *Pseudo-nitzschia* genus diatoms and *Enterococcus* bacteria in Massachusetts Bay and the

possibility of predicting their presence. We then describe a suite of assembled data sets relevant to the Massachusetts Bay area which can be used to investigate those, and other, questions. Taken as a whole, this chapter on interdisciplinary data science provides background on the expanding possibilities for environmental health research. This chapter lays the groundwork for Chapter 4 of this dissertation, where we describe the development and testing of predictive models for two marine-sourced human health risks in Massachusetts Bay using a variety of public data sources.

The Emergence of Big Data and Data Science.

The rapid increase in the number of electronic records, along with the changing nature of available digital content, has ushered in an era of 'big data.' In this work we use the term big data to refer to any single dataset containing over 10 million records. Big data can refer to millions of computerized health records, aggregated news articles about the same topic from multiple sources over time, or financial records for large international companies. In addition, big data refers not just to datasets that can contain billions of records, but also requirements for handling data in ways that go beyond the capabilities of traditional statistical software packages.[2] For example, at present a single file in the latest version of the Microsoft Excel® software program can contain slightly over 1 million records.[3] New computer programs have been developed in response to the computational demands of big data analysis (e.g., Apache™ Hadoop®[4]) and we expect that these tools will continue to develop in response to rapidly changing technology and

user needs. The generation of big data and other new sources of data (including digitization of historical paper records) does more than simply provide more data points, this availability can spur new questions about the world and the development of research sub-disciplines.

What Makes Big Data Different. Although big data is a term with multiple popular definitions, IBM defines big data as having four elements: volume, variety, velocity, and veracity.[5] Volume refers to the scale of data, which could come from internal or external sources, on a global scale some estimate that 2.5 quintillion bytes of data are created each day.[5] This volume of data is generated from the next 'v' on the list, variety. Variety of big data refers to the different forms and sources, such as transaction data, social media, sensors, and mobile devices – including new product classes such as wearable wireless health monitors aimed at the general public.[5]This expanded volume and variety of data may also be generated and transmitted throughout an organization at a faster velocity (the third 'v') than previously seen because of computing and connectivity advances. For example, the New York Stock Exchange captures 1 terabyte of trade information during each trading session, a data stream of interest to both regulators and market analysts. The fourth 'v' refers to veracity, or uncertainty, of data.[5] With any data generation there is the potential for errors to enter a data stream. For datasets that grow rapidly and are continuously analyzed there is the potential for undetected data distortion to become magnified and for errors to propagate through dependent systems (an example of this is discussed in the section below on Google Flu Trends). IBM asserts that veracity

of data is a significant issue and that poor data quality costs the U.S. economy around $3 trillion per year while 1 in 3 business leaders don't trust the information they use to make decisions.[5]

Data veracity is a concern in every field including the life sciences where an important component of university-level coursework is proper data collection and storage. For example, in undergraduate biological laboratory courses students are often graded on the quality and clarity of record keeping in formal laboratory notebooks. Similarly, at pharmaceutical companies laboratory notebooks are considered legal documents that must be stored in locked safes when not in use. In all cases quality record keeping is the foundation of quality data. As datasets get larger there is still a need for quality control and quality assurance, and this is one element of the job of 'data scientist', a specialty title of someone who engages in 'data science.'

Data Science. The terms data science and data scientist have come into popular use in the last decade and while specific definitions differ there are broadly agreed upon common elements.[5-8] IBM defines the role of a data scientist as 'somebody who is inquisitive, who can stare at data and spot trends' and someone 'who does not simply collect and report on data, but also looks at it from many angles.[8] A 2012 article from the Harvard Business review described a data scientist as 'a hybrid of data hacker, analyst, communicator, and trusted advisor' and that what a data scientist does is 'make discoveries while swimming in data,' most notably they are people who 'bring structure to large quantities of formless data and make analysis possible.'[6] For IBM, the

educational background of a data scientist is expected to be similar to that of a traditional business or data analyst with a "solid foundation in computer science and applications, modeling, statistics, analytics and math."[8] Regardless of the specific technical background or work environment, the profession of data scientist is co-emerging with the expansion of big data. The cross-disciplinary approach and combination of multiple disparate data sources used in this dissertation could be seen as an example of data science applied to a specific question. Knowing how to find, access, and organize data across multiple disciplines is one element of synthesis research, different data sources relevant to environmental health are described in the next section.

Examples of Major Environmental Health Data Sources.

The results from purpose-designed experiments generating direct observations still constitute the highest tier of scientific data, complementing such experimental data are environmental monitoring data which may reveal changes over long time periods. Magnifying the value of the cumulative efforts of individual scientists, laboratories, and institutions are numerous national and international repositories for specific types of data or scientific publications. These data repositories may be managed or funded by non-profit organizations, academic institutions, government agencies, or some combination thereof. Major examples of such databases are described below. This list should not be considered all-inclusive, sources continually evolve based on demand and funding. These examples were chosen because of their stability, public accessibility, or relevance

to environmental health research. In addition, some of these databases have accumulated enough records to qualify as big data databases.

Multi-topic Databases. These multi-topic databases may be useful to researchers in a wide variety of fields, from biology to computer science to history.

1) U.S. Census Bureau: This website provides historical demographic, economic, health, housing, and other official statistical data for the United States. Products may combine categorical data with spatial detail at the level of census block groups (ranging from 600 to 3,000 people).[9; 10] The level of detail published by the U.S. Census allows for nuanced spatial analysis over time.

2) Google Trends: This website from Google displays stories that are 'trending' based on user-entered search terms in the free Google search engine. Topics can be filtered by categories such as 'Business' or 'Sci/Tech' and by country. Within the US, Google Trends displays a map of interest by region (state level) for an individual story.[11]

3) Amazon Web Services List of Public Data Sets: Amazon maintains a list of public data sets (including big data datasets) that customers can use.[12]

4) Dryad Digital Repository: Dryad is a non-profit long-term repository for data used in international scientific and medical literature, including data in the form of text, spreadsheets, video, photographs, and software code.[13] Datasets deposited in Dryad are free to use and citable in new publications.[14] Each Dryad data package received a unique

Digital Object Identifier that can be used when citing or locating data. Datasets are free to use but there is a small charge for depositing data packages with Dryad.[13]

Health Databases. These databases are relevant to health and medicine researchers, topics range from basic biology to clinical specialties. The websites and databases that serve specific molecular biology topics are too numerous to list here.

1) HealthData.Gov: This website is run by the U.S. Department of Health & Human Services (HHS) and aims to make data from the HHS agencies (including the CDC, FDA, and NIH) easily available and accessible to the public. This evolving website aims to make all the data it serves up to be machine-readable, downloadable, and accessible via application programming interfaces.[15]

2) PubMed: PubMed is a database of over 24 million citations from biomedical literature managed by the U.S. National Institutes of Health (NIH, a government entity).[16]

3) GenBank®: An annotated genetic sequence database of all publicly available DNA sequences maintained by the NIH since 1982. GenBank releases a public update every two months and, as part of the International Nucleotide Sequence Database Collaboration, exchanges data with the DNA DataBank of Japan and the European Molecular Biology Laboratory. Each nucleotide sequence uploaded to GenBank receives a unique Accession Number, as of mid-2015 GenBank has archived over 100 million sequence records representing over 100 billion nucleotide bases.[17]

4) Foodborne Outbreak Surveillance System (FOSS) Online Database: This database is run by the U.S. Centers for Disease Control and Prevention (CDC). FOSS receives reports from state, local, and territorial public health agencies about recorded foodborne illnesses.[18]

5) United Network for Organ Sharing (UNOS): The UNOS is a private non-profit organization that manages the U.S. organ transplant system under contract with the Federal government.[19] The UNOS database is a resource used across transplant disciplines because it contains outcomes and treatments used in transplant recipients. As described by one physician, "specialists in one field (e.g., cardiac transplants, a relatively new field) can pull information from UNOS on long-term consequences of immunosuppressive medicines that been used in one transplant type (e.g., kidney) to aid in the care of transplant patients in another type (e.g., cardiac)."[20]

Ecology and Environment Databases. These databases examples spanning multiple environment types and may include both biotic and abiotic environmental data.

1) Integrated Ocean Observing System (IOOS®): A regional-national partnership for sharing ocean, coastal, and Great Lakes data on topics including wave heights, sea level, wind, temperature, salinity, and dissolved oxygen levels. IOOS is a member of the Global Ocean Observing System (GOOS) coordinated through the United Nations.[21]

2) Ecological Society of America Data Registry: This registry describes data sets on ecology and environmental topics from articles published in the journals of the Ecological Society of America.[22]

3) TRY Plant Trait Database: The Max Planck Institute for Biogeochemistry in Germany hosts this international database developed by scientists of morphological, anatomical, biochemical, physiological, or phenological features of plants, with many geo-referenced records.[23]

Environmental Health Databases. These two databases contain information relevant to environmental topics with a close relationship to human health.

1) ENHanCed Infectious Diseases (EID2) database: This database, funded by the European Union and hosted at the University of Liverpool, contains data on pathogenic organisms and the country in which they may occur, lists of carrier organisms, genetic sequences, and publication links.[24]

2) Center for Coastal Monitoring and Assessment National Status & Trends Database (NS&T): Run by the U.S. National Oceanic and Atmospheric Administration (NOAA), the NS&T is comprised of three nationwide programs, Benthic Surveillance (discontinued in 1993), Mussel Watch and Bioeffects that are designed to describe the current status of, and detect changes in, the environmental quality of U.S. estuarine and coastal waters through environmental monitoring, assessment and related research. Starting in 1986, the Mussel Watch program is the longest running continuous contaminant monitoring program in U.S. coastal and Great Lakes waters.[25]

Remote Sensing Data Sources. These two examples are the major public sources of satellite remote sensing data and model output products derived from that data. Other remote-sensing data may available from private entities.

1) U.S. National Aeronautics and Space Administration (NASA): NASA provides satellite remote sensing data from multiple spacecraft and instruments sources with varying temporal scales, spatial scales, and image resolution. Topic areas include global precipitation, thermal anomalies, ocean color, land cover and vegetation, and snow and sea ice cover.[26]2) European Space Agency: The European Space Agency provides public data related to radar imagery, radar altimetry, optical/multi-spectral radiometry, atmospheric data, and gravimetric data from multiple missions.[27]

The example databases listed above are public repositories of data that could be relevant to OHH researchers depending on the question of interest. However, the data in each repository requires specialized knowledge to interpret. For example, foodborne illness outbreaks are a different type of data than land cover type images, but when combined they might provide new insights. The need for individuals or specialized data science teams that can combine and utilize diverse data types may grow as society poses research questions related to multiple disciplines. In addition, researchers should be aware of the potential to glean data from non-traditional sources, examples of which are provided in the next section.

Non-traditional Data Sources: Social Media and Crowdsourcing.

Social media postings can range from news photos to personal thought updates and are generated and shared publicly by numerous individuals around the world. Social media platforms use a variety of software applications on technological platforms ranging

from desktop computers to mobile devices. Companies that own and operate social media platforms have the ability to aggregate and analyze user postings, with the potential to generate what is essentially crowdsourced big data from voluntary content created by users. One dictionary defines 'crowdsourcing' as the "practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers."[28]With the spread of the Internet, and Internet-connected smartphones, the ability for spatially distant groups to communicate, give feedback, and share information in near-real time is enormous. In addition to Internet connectivity, mobile communication devices now often include the ability to share place-based, geo-referenced, information (including latitude and longitude) along with an observation record (e.g., photo or social media posting) directly from the device itself. In some cases mobile phones are able to act as sensors without conscious action by their owner, a functionality already utilized by some companies to generate location-specific crowdsourced observations.

Crowdsourcing is not strictly associated with mobile devices or social media but also result from aggregated information collected over time. Crowdsourcing can refer to the practice of allowing multiple users to contribute to a single task, such as in the online game FoldIt that "attempts to predict the structure of a protein by taking advantage of humans' puzzle-solving intuitions and having people play competitively to fold the best proteins."[29] In the case of FoldIt, crowdsourcing does not generate big data, but rather

provides an organizing mechanism to harness the unique contributions of many volunteers. Another crowdsourced product is the Encyclopedia of Life which began with the idea to provide a webpage for every species on earth, and seeks to bring together information from trusted resources such as museum, professional societies, and expert scientists into a massive database.[30]Crowdsourced data take multiple forms, with different levels of accessibility and reliability, examples of potential public sources for crowdsourced data are listed below, social media postings and Google search terms can become crowdsourced data if aggregated properly.

1) Twitter: Twitter is described on its homepage as "probably the largest publicly accessible alternative trove of social-media data."[31] Services exist to take advantage of the large amounts of user-generated real-time postings, many of which use hashtags (identifying words or phrases) to refer to specific events or topics.

2) Instagram: A social photo-sharing service that allows users to associate location data and hashtags to posted photos, visible to either private groups or the public. Photos are searchable based on multiple attributes[32] and third-party services are available to aggregate information about Instagram posts.

3) Facebook: Facebook is social network site with more than 1 billion active users that allows people to share photos, text, and other information across their social network.[33] The list of Facebook users and their basic characteristics alone is considered big data.[34]4) Google Correlate: A service from Google which helps users "finds search patterns which correspond with real-world trends."[35]While not strictly a social media source, Google

Correlate results are derived from anonymized searches made by users of the free Google search engine tool and could be considered as the results of crowdsourcing.

The sources that generate these social media data can be stationary or mobile. A user of Facebook might have an account tied to a specific city, but be able to post to their account from a mobile device anywhere in the world with an Internet connection. Due to the potentially enormous numbers of individually-generated records from social media postings or geo-referenced Internet search queries, there is interest in using these sources to monitor near real-time events, including predicting or tracking disease outbreaks or other influences on human health. The example of using location-referenced Internet search queries to predict the level of influenza activity across the United States, an important public health issue, is described in the next section.

The Example of Google Flu Trends. An example of the promise, and perils, of crowdsourced data can be found in the story of the Google Flu Trends project.[31; 36-38] Google Flu Trends (GFT) was developed in 2008 in conjunction with the Centers for Disease Control and Prevention (CDC), and involved data mining records from Google search engine queries using influenza-related search terms, in conjunction with the CDC's historical data, to develop a model that could estimate cases of influenza faster than traditional epidemiological surveillance methods.[31] A separate paper on the predictive model development was also published by Ginsberg et al. (2009).[39] Notably, GFT was released after human infection with a novel influenza A virus became a

nationally notifiable condition in 2007[40], but before the H1N1 influenza pandemic of 2009-2010.[41]

Model development involved finding the best matches among approximately 50 million Google search terms used between 2003 and 2008[39] to the CDC's 1152 historical data points[36] of influenza-like illness (ILI) related physician visits.[39] The assumption underlying this project was that Google search terms are proportional to the incidence of ILI physician visits. As Lazer et al. (2014) point out, with so many search term records available "the odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high."[36] Indeed, the model authors noted that their top 100 queries included topics like 'high school basketball', which tend to coincide with the U.S. influenza season.[39] In contrast, CDC data is derived from influenza surveillance methods based on voluntary weekly reporting from state level surveillance programs or health-care providers.[40]

Currently, the **U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet)** consists of more than 2,900 outpatient healthcare providers across the U.S.[40] As with other disease surveillance, there is a degree of underreporting because any patient with ILI must first visit a physician. Hence, the CDC cautions that while ILI surveillance answers the questions of where, when, what influenza viruses are circulating and if influenza activity is increasing or decreasing, it cannot be used to ascertain how many people have become ill with influenza during the influenza season."[40] Although passive surveillance systems do not report all disease cases, if surveillance methods and system

performance remain relatively unchanged over time it should be possible to compare the *relative* severity of different influenza seasons.[42]

In 2009 the GFT algorithm was updated after the first version badly underestimated the number of influenza-like illness (ILI) cases at the start of the H1N1 (swine flu) pandemic.[31] However, the 2009 model then ran essentially unchanged until updates were announced in October 2013.[36] During the 2010-2013 flu seasons, GFT was often overestimating flu prevalence (for 100 out of 108 weeks) with non-random errors because temporal autocorrelation meant one week's errors influenced the follow week's errors.[36] At one point during the 2012-2013 flu season GFT estimates of flu prevalence across the USA were more than double the CDC estimates of ILI,[36] a significant overprediction.[43] Despite limitations of traditional ILI surveillance, the CDC is considered the authoritative source for national ILI estimates and produces them using a historically consistent methodology.[40] It is not entirely clear what was driving GFT's persistent overestimation of ILI compared to CDC values because Google has not publicly released documentation on the specific 45 search terms it used in its GFT model training, but one possibility is that heavy media coverage of the flu during the 2012-2013 season influenced user searches and thus skewed the GFT prediction results.[36] The model developers noted this very possibility, in their 2009 paper Ginsberg et al. wrote:

> "*The search queries in our model are not, of course, exclusively submitted by users who are experiencing influenza-like symptoms, and the correlations we observe are only meaningful across large populations. Despite strong historical*

correlations, our system remains susceptible to false alerts caused by a sudden increase in ILI-related queries."[39]

In late 2014 Google announced that GFT would stop relying solely on search terms to make flu predictions, but would instead combine search terms with publicly available data from the CDC to make predictions for the 2014-2015 season.[37; 43] This announcement came after a high profile news article critical of GFT was published in the scientific journal *Nature* in February 2013.[31]

What lessons can be learned from the example of Google Flu Trends? It is a program that showed initial promise, received favorable public attention, then produced large errors and failed to perform accurately for over two years – producing estimates quite different from those produced by the CDC. As a result GFT then received negative attention (from academic and media sources), and is currently being updated by the creators with the promise (made in October 2014) of a technical paper to be published.[43] Perhaps one lesson from GFT is that crowdsourced data collected from unwitting participants should be viewed with caution, sheer volume does not assure veracity. Another caution is that episodic events (e.g., the appearance of a new disease) have the possibility to severely skew any model predictions based solely on human psychology instead of microbiological reality, and any such models should be monitored for potential skew. It remains to be seen if the refined GFT will perform as desired over multiple influenza seasons. Lazer et al. (2014) specify two problematic elements in the history of GFT that should be considered when working with large crowdsourced or social-media-

derived datasets, 1) big data hubris, and 2) algorithm dynamics. These terms are defined and discussed in the next section.

Crowdsourced Data, Big Data Hubris and Algorithm Dynamics. 'Big data hubris' is described as the 'assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis' because 'quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data.'[36] As illustrated by the example of Google Flu Trends, quantity of search terms does not replace the existing system of physician-diagnosed reporting data through established public health channels. The caution has even greater relevance when large data sets are generated through opaque processes that a researcher may not be able to account for within their analysis. For data from private sources, such as online search engines, it is also possible that algorithm dynamics may play a role in influencing any records generated. 'Algorithm dynamics' refers to the changes made by software engineers to improve the commercial service being used (for GFT the commercial service is the free Google search engine, but other commercial services include Facebook and Twitter) and changes in behavior of consumers using the service.[36]The act of suggesting additional search terms, or using an 'auto-complete' function in a text field, can influence user behavior. Lazer et al. (2014) note that 'search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider'.[36] Such actions by a service provider are known as 'blue team'

dynamics. ‘Blue team’ and ‘red team’ dynamics refer to algorithm dynamics influenced by specific groups of people.

Blue team actions come from inside a company, and include modifications to the algorithm producing the data, whereas red team actions result from research subjects (or an outside entity) attempting to manipulate the data-generating process to skew results in a particular direction.[36] In the case of the social media service Twitter, the company (the blue team) provides a list of topics that it considers to be the most timely ‘Trends,’[44] An example of how this list appears to viewers of the Twitter website search page is shown below in Figure 4-1.

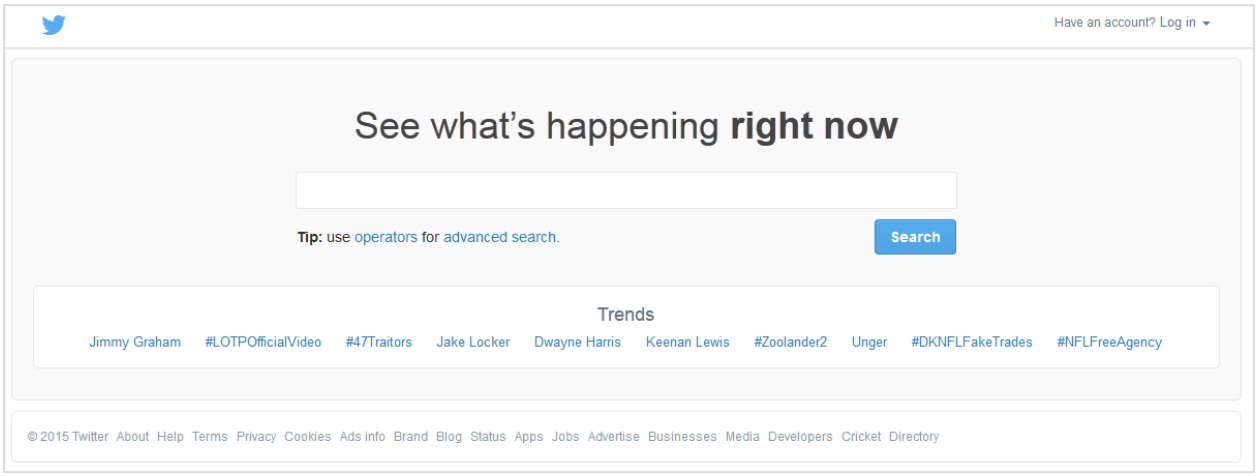

Figure 4-1. Screenshot of Twitter Search homepage, March 10, 2015 5:24 pm, displaying Trends as identified by Twitter.

The trending items listed in blue text on the screenshot shown in Figure 1 are: Jimmy Graham, #LOTPOfficialVideo, #47Traitors, Jake Locker, Dwayne Harris, Keenan Lewis, #Zoolander2, Unger, #DKNFLFakeTrades, and #NFLFreeAgency. It is not immediately

clear if these topics or hashtags could be thematically linked; but they are presented to website visitors because of the underlying Twitter algorithm that identified them for display (i.e., a blue team action).

The act of presenting these topics on the Twitter homepage means that they have the potential to be seen by more people and receive more exposure, creating a positive feedback loop. Lazer et al. (2014) note that 'campaigns and companies, aware that news media are monitoring Twitter, have used numerous tactics to make sure their candidate or product is trending.'[36] Such tactics would be considered red team dynamics because they are the work of an external group. Red team dynamics can be used to take advantage of this potential feedback loop between grassroots social media and traditional news media (e.g. television stations) to increase media exposure for a certain topic. This is one caution that researchers using crowdsourced data should be aware of, the possibility of intentional skew by persons taking advantage of algorithm dynamics and data generating processes to amplify, or echo, signals that would not rise to prominence if the system were truly operating independently, organically, and transparently.

While crowdsourced data from social media might be an accurate reflection of posted content within a single social media site, such data should not be confused with a being an accurate representative sample of anything beyond a narrowly defined universe of subjects. For example, Twitter reports that it has 288 million monthly active users, that 77% of Twitter accounts are outside of the U.S., that 80% of active users are on mobile devices, and that 500 million individual posts (known as 'tweets') are sent every

day.[45] What is less clear is the distribution of Twitter users across spatial areas or demographic groups. Persons viewing English-language Twitter posts as representative of the entire United States, or English-speaking world, would be wise to use caution when interpreting such crowdsourced products as they could differ significantly from a statistically representative sample. Active participation rates for various online social media platforms may vary between demographic groups. However, other types of crowdsourced data do not require active participation or content generation, but instead only the use of a certain level of technology (i.e., smartphones) and so may draw from a larger population.

Some crowdsourced products do not depend on conscious input from users, but instead allow personal mobile phones to act like individual sensors within a much large network, this is the case with Google Traffic. Any mobile phone with the Google Maps application and GPS-location services enabled will report movement speeds back to Google, which continuously combines data from millions of users and projects it through the Google Maps application in the form of color-coded street overlays.[46] Such recorded crowdsourced data are potentially less subject to red team dynamics than social-media sourced data because the outputs (e.g., the color-coded street overlays shown by Google Traffic) are calculated using large sample sizes and the data are contributed unconsciously and anonymously by users, any signal manipulation would require mass participation or deliberate mis-reporting of travel speeds.

To conclude this section, crowdsourced and social media data, including big data, can be seen as both rich and dangerous. Data culled from social media offers the potential of highly detailed temporal and spatial topical records that could provide fascinating social insights at low cost, but also of meaningless correlation on a grand scale. The difference between reported data (subject to human discretion) and recorded data (generated neutrally by sensors) will continue to be an important distinction. Another area of changing technology is on-animal sensors, which could allow for expanded ambient environmental condition reporting[47] as a potential compliment to human-crowdsourced data. Although this dissertation research will not rely on social media or crowdsourced data, it remains of interest for the future work given the expanding possibilities of potential environmental sensors. In contrast to the young world of social media, a more mature but still non-traditional data source being used for health research is remotely-sensed data from satellites – successful examples of this type of work are described in the next section.

Interdisciplinary Workflow and Supporting Examples.

While the use of crowdsourced or social media-sourced data for health monitoring and prediction is still maturing, other types of data that were not specifically designed for health-related research are regularly being used to investigate environmental health questions. This section will provide examples of cross-discipline data sharing being applied to predict environmental health problems, and in some cases to provide solutions in places where on-the-ground research or monitoring capacity may be limited and a

longer lead-time is needed to preposition supplies and people in anticipation of medical needs.

In 2007 the World Meteorological Organization (WMO) organized a conference titled *The International Conference on Secure and Sustainable Living: Social and Economic Benefits of Weather, Climate and Water Services.*[48] The conference called for "multi-disciplinary understanding between providers and user of weather, climate and water services as they are essential for improve decision-making and delivery of social and economic benefits."[48] Increasing the use of large-scale remote sensing data in the fields of environmental and public health requires that researchers be aware of available relevant data and able to apply it in the context of a health problem (possibly as a proxy when on-the-ground measurements are not available). A general workflow is shown in Figure 4-2, below.

Generalized Interdisciplinary Workflow. The workflow presented in Figure 4-2 divides the synthesis research work into three phases: 1) conceptual/exploratory and hypothesis generation; 2) development; and 3) evaluation. Phase 1 is similar to any initial stage of research, involving a review of existing literature on the disease, identifying any known environmental associations or suggesting new associations, and gathering potentially useful existing data sets. Phase 2 is the development phase where researchers might employ multiple techniques such as spatial-temporal mapping, statistical modeling, or model term development while utilizing datasets collected across multiple disciplines. In traditional life sciences research this would correspond to the

phase where a researcher conducts an experiment and records the results. However, in this synthesis process the type of work has shifted from direct manipulation of an environmental condition to work exploring data through modeling, mapping, or other associative methods. Phase 3 is the evaluation of the product(s) from Phase 2. Phase 3 may involve comparing the accuracy of a hindcast estimate from a model or method against recorded data, using a product from Phase 2 to guide field-based sampling to generate new data, or the further development of a theory based on revealed associations between multiple datasets.

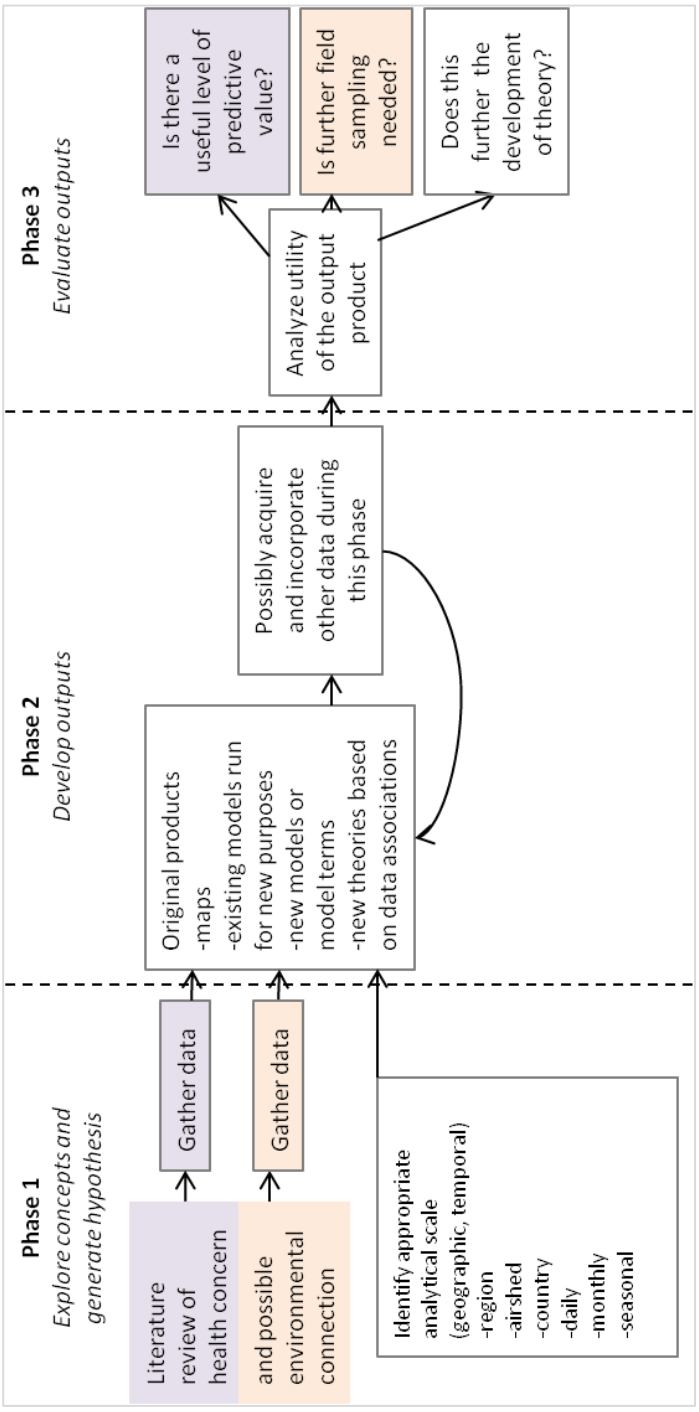

Figure 4-2. Workflow process for synthesis research. This workflow shows a generalized path for combining traditional health data and non-traditional sources such as remote sensing. Orange boxes represent environmental data, purple boxes represent epidemiological data.

Interdisciplinary Environmental Health Examples. Three examples of synthesis research are the prediction of Rift Valley fever outbreak in the Horn of Africa (see Anyamba et al. 2009), investigations into cholera outbreaks in south Asia (see Pascual et al. 2000, Koelle et al. 2005, and Huq et al. 2005), and the link between regional wind patterns and Kawasaki disease in Japan (see Rodó et al. 2014). Examples are described below.

Rift Valley Fever predictions in the Horn of Africa based on climate anomalies. Anyamba et al. (2009) investigated the historical relationship between El Niño/Southern Oscillation (ENSO) related climate anomalies and Rift Valley fever (RVF) outbreaks in the Horn of Africa to develop a model that allowed them to make a prospective spatial estimate of the 2006-2007 RVF outbreak. They combined historical satellite measurements of environmental parameters for sea surface temperature (SST), outgoing longwave radiation (OLR) as a proxy for rainfall, and vegetation measurements of photosynthetic activity transformed into a normalized difference vegetation index (NDVI), with epidemiological information on previous RVF outbreaks in the regional to develop early warning risk maps.[49] The virus that causes RVF in animals and humans is spread by mosquitoes, so NDVI was used as a proxy for persistent above-average rainfall and associated vegetation growth that provide mosquito habitat in the Horn of Africa.[49]

In late November 2006 when environmental conditions predicted an upcoming elevated risk of RVF outbreaks, government stakeholders were able to begin entomological surveillance and place public health authorities on alert weeks before any

reported human cases of RVF. After RVF transmission was confirmed in mid-December 2006 additional surveillance activities and disease mitigation activities were implemented; mitigation activities included restrictions on animal movements, distribution of mosquito nets, and information campaigns, along with domestic animal vaccination and mosquito control in specific areas.[49] Post-outbreak mapping of reported human cases found that 64% of cases were within predicted risk areas, and that most of the remaining 36% of human cases were within 50km of the outer edges of predicted risk areas.[49] These results demonstrate the feasibility of combining remote-sensing data and historical epidemiology data to make near-term predictions about disease risk for a virus whose spread is closely coupled with regional environmental conditions.

The generalized workflow employed by Anyamba et al. (2014) is diagrammed below in Figure 4-2. First, the authors brought together historical data on RVF distribution and environmental parameters associated with an increase in population of the mosquito vector. Second, the authors developed maps of historical disease range and combined them with monthly updates of vegetation growth as indicative of conditions favorable to the mosquito vector to predict areas where public health interventions should be focused. Finally, the authors evaluated the location and occurrence of recorded cases against the areas predicted by their method.

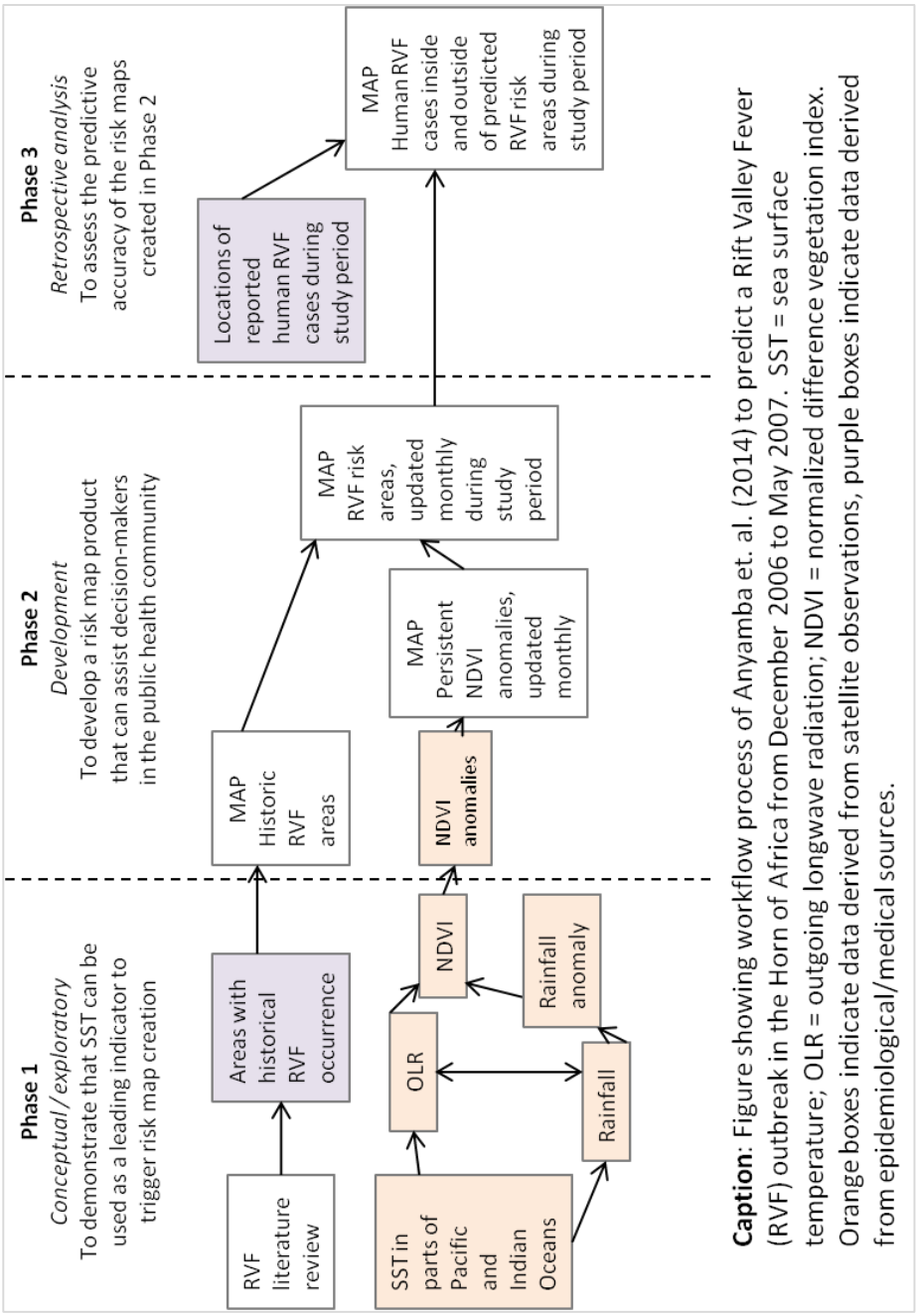

Figure 4-3. Generalized workflow employed by Aynamba et al. (2014).

Linking Cholera Dynamics in Bangladesh to Environmental Forcings. To examine the links between cholera risk and environmental forcing in Bangladesh, researchers have tied together rainfall, river discharge, flood extent, cholera cases, and temporal changes in population immunity to tease out the influence of El Niño/Southern Oscillation (ENSO) on cholera dynamics.[50] Cholera is caused by *Vibrio cholera* bacteria of which there are multiple strains, in Bangladesh the most notable strains are the El Tor and the Classical.[50] Recovery from an infection by one strain leads to acquired cross-immunity for both strains but this immunity wanes over time, this drives temporal changes in population-level immunity.[50]

Vibrio cholera bacteria are spread via fecal-oral transmission or exposure via contaminated water. At the population level susceptibility to cholera varies in a non-linear way over time; individual immunity acquired from previous exposures wanes over time, and new susceptible individuals regularly enter a population through births or aging.[50] Koelle et al. (2005) developed a model that considered immunity, disease transmission, and environmental forcings, with results that show a strong correlation between cholera transmission and climate variability. Cholera cases were found to decrease during summer monsoons, possibly due to the dilution of *V. cholera* in the environment or a change in salinity; the rise in cases after monsoons is thought to be tied to the breakdown of sanitary conditions that accompany crowding into non-flooded areas. [50] Unlike RVF, a virus transmitted by mosquitoes, cholera is a bacterial disease whose spread is heavily influences by human activity. In the case of Bangladesh researchers found that historical epidemiological data is a critical type of information that must be

considered when assessing population-level risk because acquired immunity influences the number of currently susceptible hosts.[50-52] The question of El Niño and cholera is an example of untangling the interaction between aquatic environmental conditions, seasonal factors that affect large scale human behavior, movement, and sanitation, and population level immunity that resulted from previous outbreaks of a bacterial disease using both traditional medical data and non-traditional environmental monitoring data. The generalized research process utilized by Koelle et. al. (2005) is diagrammed below in Figure 4-3. First the authors assembled data on cholera cases and the known influence of seasonality on disease transmission. Second, the authors generated equations for the model term they were interested in quantifying, solved them, then they explored environmental associations with that newly estimated model term. Lastly, the authors compared their model to the observed data and explained how their results supported their hypothesis.

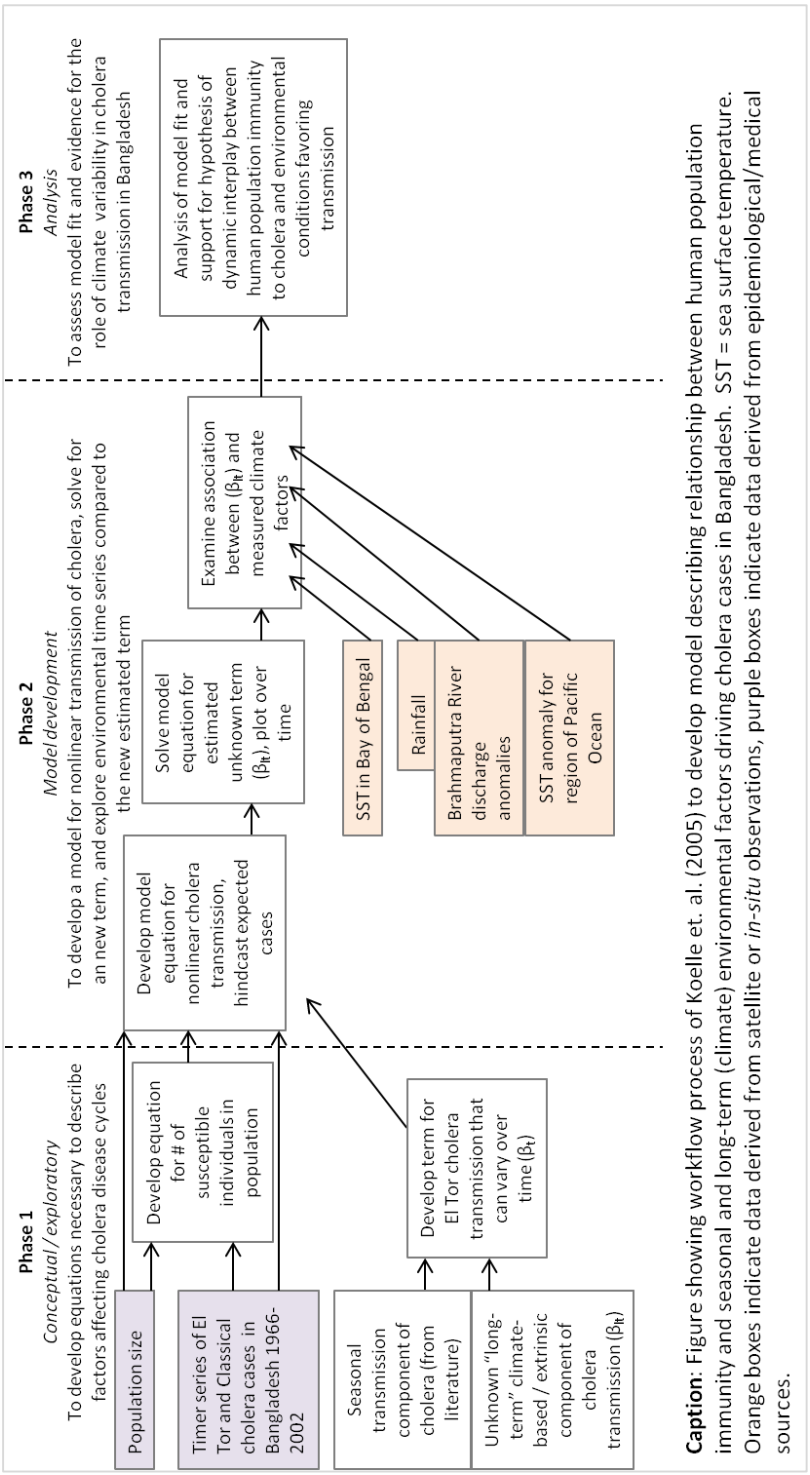

Figure 4-4. Generalized workflow employed by Koelle et al. (2005).

Predicting Kawasaki disease in Japan from regional air parcel movements. Sub-global weather patterns besides ENSO have been linked to increase risk for other diseases. In the case of Kawasaki disease (KD) in Japan, specific tropospheric wind patterns have been linked to days with high incidence of KD by Rodó et al. (2014). Kawasaki disease is an acute, coronary artery vasculitis (inflammation of blood vessels) that affects children, and despite 40 years of study no single causative agent has been identified.[53] To investigate the possibility that the trigger for KD is a form of inhaled antigen (foreign body triggering an immune response) or toxin, Rodó et al. (2014) combined KD epidemiological data from Japan, regional wind patterns over Japan, regional land cover data, and microbial profiling of tropospheric and ground aerosol samples collected at times when air was coming from region identified as the possible source of the KD trigger.[53]Characterization of the aerosols indicated that tropospheric and ground samples were significantly different, providing support for the feasibility of a windborne pathogen.[53] The researchers also found that air parcels associated with higher incidence of KD in Japan had previously moved over intensively cultivated croplands for corn, rice, and wheat in northeastern China during a time when the ground was frozen. This finding led them to speculate that the causative agent might be an aerosolized fungus or pre-formed fungal toxin associated with decaying vegetation.[53] By combining remote sensing records, field measurements, and spatially-based epidemiological data, this research has suggested new avenues of investigation to understand and predict changes in KD risk.

A generalized workflow depicting the research process used by Rodó et al. (2014) is presented below in Figure 4-4. The first phase was conceptual and exploratory and focused on gathering existing data on the topic and in the spatial area of interest. The second phase was the development of a model and generation of model outputs in the form of a map that was used in phase 3. Phase 3 included field investigation of atmospheric microbial sampling at a time and place suggested by the model developed in phase 2 and identification of atmospheric microbes.

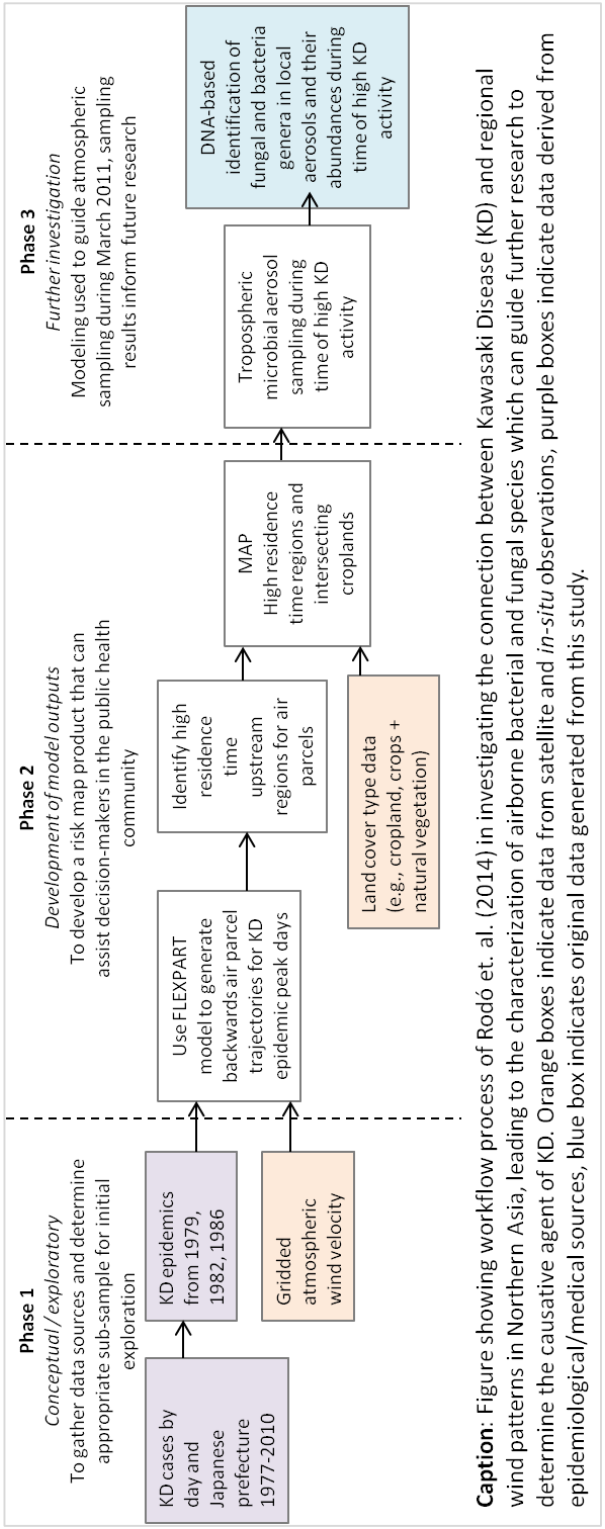

Figure 4-5. Generalize workflow employed by Rodó et al. (2014).

Section Summary. The three examples presented above all dealt with different health questions and provided different research outcomes, but the authors employed strategies with clear similarities. The research outputs for understanding Rift Valley fever in the Horn of Africa, cholera in Bangladesh, and Kawasaki disease in Japan described in this section resulted from efforts that combined traditional epidemiological data and nontraditional sources of data to provide insight into changes in disease risk probability in response to external factors. Revealing linkages between large-scale physical environmental phenomenon (such as sea-surface temperature anomalies, wind patterns, or rainfall levels) and local conditions is increasingly important in solving complex problems involving multiple states or countries. However, satellite observations are often limited to observing near the surface of an object and so cannot provide data for every environmental process of interest.[54]The examples cited above illustrate the value of applying a synthesis approach to understand how large-scale interactions between land, water, and air can determine the transport and survivability of factors ranging from raindrops to fungal spores which in turn influence local-scale human health. This inter-disciplinary mindset informs the next section, which describes data sets collected from multiple sources in support of an environmental and ocean health question.

Environmental and Socio-economic Data for Massachusetts Bay and Adjacent Coastal Watersheds.

Field studies are the foundation of scientific progress, and monitoring programs by public health authorities are crucial to environmental health science. The audience for

understanding environmental processes is wide, spanning basic science, commercial operators with business interests, and the public health and governance community. There continues to be the need for, and financial support of, carefully constructed field experiments of sufficient statistical power to reveal the symphony of biological, chemical, and physical processes that interact at multiple scales to form our world. However, after a first order understanding of relevant biological processes has been developed it is reasonable to explore methodologies that allow scientists to use the existing knowledge base to generate hypothesis and predict impacts on situations of interest. For this project, the situations of interest are the variation in the levels of *Enterococcus* bacterial populations near select bathing beaches in Massachusetts Bay and the populations of *Pseudo-nitzschia* genus diatom species in surface waters of Massachusetts Bay.

This section will describe the datasets collected to support the development of a model for linking measured changes in local indicators (identified in Chapter 2) to the measured population levels of two marine-sourced risks in Massachusetts Bay. The two marine-sourced risks are total *Pseudo-nitzschia* species as measured in Massachusetts Bay surface waters and levels of *Enterococcus* bacteria species measured in Boston Harbor at select bathing beaches. Building upon the existing body of research in the fields of *Enterococcus* and *Pseudo-nitzschia* spp. drivers of population levels in the environment we have assembled data sets of relevant parameters in order to develop a model that will attempt to estimate their influence on these populations.

The specific questions that will be asked of these data are:

1) Is it possible to hindcast levels of *Enterococcus* populations in specific areas of Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Enterococcus* growth?

2) Is it possible to hindcast levels of total *Pseudo-nitzschia* populations measured in Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Pseudo-nitzschia* growth?

3) Does there appear to be any clear relationship between *Enterococcus* levels and *Pseudo-nitzschia* levels in Massachusetts Bay?

4) Are there any field measurements for which public data do not readily exist which scientific literature suggests would likely increase the predictive ability of these models?

Such data driven exploration serves to refine theory, and identify data gaps that must be resolved to fully elucidate processes of interest, and potentially provide support for expanded field observations if necessary. This approach can be applied to other marine-sourced risks where enough evidence from experimental or field observations exists to allow for the development of predictive models based on environmental parameters.

Massachusetts Water Resources Authority Data. Massachusetts Bay, Cape Cod Bay, and Boston Harbor have long term (20+ years) coastal monitoring data for plankton, nutrients, water quality, meteorological, and hydrodynamic parameters. Much of this monitoring has been driven by the court-mandated construction of improved sewage treatment facilities overseen by the Massachusetts Water Resources Authority (MWRA).[55] The MWRA provides secondary treatment for wastewater for some two

million people in the greater Boston area; prior to September 2000 wastewater effluent was discharged into Boston Harbor with minimal treatment.[55]After September 2000, wastewater received full secondary treatment and was diverted to a discharge outfall pipe 15km offshore in Massachusetts Bay.[55] Monitoring programs collected baseline data in Boston Harbor and Massachusetts Bay started as early as 1992 and continued after the outfall went live in September 2000 in order to understand the ecological impacts of relocating the outfall site offshore.[55] The MWRA plankton sampling program identified some organisms to the species level, but others to the genus level, including *Pseudo-nitzschia* diatoms.[55] While the extent and frequency of *Pseudo-nitzschia* sampling has varied from year to year, there are some sampling station locations with 10+ year time series that will be utilized for this research project. Further details on data sources relevant to Massachusetts Bay are described in the next section.

Massachusetts Bay and Coastal Watershed Data. It is important to note that the data assembled from the MWRA and other sources listed in Table 4-1 were not collected for the purpose of developing a predictive model of *Enterococcus* bacteria or *Pseudo-nitzschia* diatom species. Rather, the MWRA monitoring program was the result of legal action, and sources such as stationary buoy sources are largely in support of weather observation and marine navigation safety. These data should be considered 'data of opportunity' and their explanatory power might be limited. The MWRA disclaimer accompanying the data states:

"*These data are from the Environmental Monitoring and Mapping System (EM&MS) and the Department of Laboratory Services LIMS system (LIMS), Oracle databases utilized by the Massachusetts Water Resources Authority (MWRA) Environmental Quality*

Department (ENQUAD). These data are collected from a variety of sources which measure the data for a variety of different uses and with different standards for accuracy and precision, and are distributed as is. MWRA cannot ensure that they are appropriate for any particular use. Neither the MWRA, the Commonwealth of Massachusetts, nor any agency to whom MWRA has given data, or from whom MWRA has obtained data, shall be held liable for any reason related to the accuracy or fitness of the provided data."[56]

The Massachusetts Water Resources Authority (MWRA) has made water quality monitoring data from Boston Harbor available for select observations starting in 1989.[56] Users can download spreadsheet files of environmental data that contain the following data elements: project ID, region, subregion [a specific beach area], Department of Environmental Protection (DEP) segment, station ID, surface or bottom [sample location within a water column], date/time, depth of measurements, temperature, salinity, specific conductance, dissolved oxygen (DO), DO Percent Saturation, pH, and turbidity. Not every observation record contains all of the data elements listed above. For example, in the spreadsheet file for Physical Data the earliest turbidity measurements occurs in 1994, but there are many records in years 1997, 1998, 2000, with no turbidity measurements. Gaps exist for other years and different observation stations as well.

The file named "bh_nutrients.xlsx" contains MWRA data on nutrient and chlorophyll measurements from 1994 to 2014.[56] Not every sample contains a record for every data element in the file. The list of data elements includes: project ID, region, subregion, DEP segment, station ID, date/time, surface or bottom, sample depth, Ammonium, Nitrate+nitrite, Total dissolved Nitrogen, particulate nitrogen, Total

Kjeldahl Nitrogen (TKN), Phosphate, Total dissolved P, Particulate P, Total Phosphorous, Particulate Carbon, Chlorophyll *a*, and Phaeophytin. However, of the records in this file, data is largely only available for the following categories: Nitrate+nitrite, Total Kjeldahl Nitrogen, Phosphate, Total phosphorus, Chlorophyll a, and Phaeophytin.[56]An examination of the spreadsheet containing these records shows that there are numerous gaps in the record. Not all samples were collected at the same time, on the same time scale, or consistently over the past 20+ years by the MWRA. In addition to the public data on the MWRA website, further data from sampling in Massachusetts Bay is available upon request. We have acquired data from station F22 and F23 for that includes most of the nutrients listed above as well as salinity, silicate, and zooplankton counts.[57] Data processing and cleaning methods used are described in Chapter 4. We acquired additional *Enterococcus* count data from marine beaches via the Massachusetts Depart of Public Health, Bureau of Environmental Health.

In addition to the MWRA, other organizations collect data about Massachusetts Bay to fulfill their own mission requirements. For example, NOAA collects and archives weather observations from various observation stations, including land-based observations at Boston Logan International Airport, and sea-based observations from buoys in Boston Harbor and Massachusetts Bay.[58] Multiple regional ocean and weather monitoring data streams are collected by the Northeastern Regional Association of Coastal and Ocean Observing Systems (NERACOOS) to support marine operations, safe navigation, and research. NERACOOS partners include multiple Universities, research institutions, and government offices at the state and federal level.[59]

The completeness of records and extent of spatial coverage for the collected datasets varies. Sources consulted for the project include federal government agencies, a non-profit organization that works with multiple government entities, state agencies, and local town offices. State agency datasets were acquired through multiple methods, including downloads from public websites and email requests to agency employees inquiring about public, but unpublished, data. Dog populations for coastal cities bordering Massachusetts Bay were assembled by the author in 2012 by emailing or calling town clerk offices or the relevant authority in charge of dog licensing. Data sources compiled are listed in Table 4-1, below. The compilation of data from such a diverse array of sources illustrates the interdisciplinary nature of environmental health work, especially ocean and human health work. Acquiring and exploring the data is part of Phase 1 (conceptual exploratory) of the synthesis research workflow process as depicted in Figure 4-2.

Table 4-1. Data Sources for Massachusetts Bay and Coastal Watersheds

Source Name	Source Type	Data Types	Sampling Frequency	Unit of spatial analysis	Source
U.S. Census, Decennial Census	Federal government	Population, age, sex, housing units, household income, other demographic data	Every 10 years, entire USA	Polygon (blocks are the smallest unit of analysis, multiple blocks in a tract). Covers entire USA	9
U.S. Census, American Community Survey	Federal government	Housing stock, wastewater treatment, other demographic data	Every year, approximately 1 in 38 U.S. households receive the survey.	Polygon, smallest unit of analysis is tract. Data can be grouped by other administrative boundary types	60

Table 4-1. Data Sources for Massachusetts Bay and Coastal Watersheds					
Source Name	**Source Type**	**Data Types**	**Sampling Frequency**	**Unit of spatial analysis**	**Source**
Dog population in coastal cities bordering Massachusetts Bay	Compiled by author via phone and email survey in 2012	Number of dogs registered to town, some estimates of unregistered dogs	One time survey	Polygon, city	Unpub-lished data
U.S. Geological Survey, River discharge data	Federal government	Average daily discharge rate at 8 stations along waterways, no data on Cape Cod	Continuous, June-Aug records 2007-2014	Point recordings, flow rates. Stations 0110- 0000, 55566, -5876, -5870, -5730, -5608, -5583, -5585, -2345.	61,62
Northeastern Regional Association of Coastal and Ocean Observing Systems	Non-profit organization partner of federal, state, local government, academia, and industry. Part of U.S. Integrated Ocean Observing System	Ocean and weather conditions from buoys in the northeast. Air temperature, water temperature at multiple depths, salinity, chlorophyll, turbidity, wind direction, current direction.	Minute by minute, but reports of daily averages are available. Annual data acquired for of 2000-2014.	Point records at buoy locations. Buoy A01 is in Massachusetts Bay.	59
NOAA Buoy Station BHBM3, Boston Harbor	Federal government, data also served via NERACOOS	Air temperature, water temperature	6-minute intervals	Point, on the shoreline of Boston.	58
NOAA National Climatic Data Center	Federal government	Precipitation, air temperature to tenth of degree, average daily wind speed	Daily average	Point, Sampling station at Boston Logan Airport	63

Table 4-1. Data Sources for Massachusetts Bay and Coastal Watersheds					
Source Name	**Source Type**	**Data Types**	**Sampling Frequency**	**Unit of spatial analysis**	**Source**
Massachusetts Department of Public Health, Beach Water Quality Testing	State government	*Enterococcus* sampling results	Weekly, monthly, or daily during summer bathing season depending on location and previous test results	Point data, representing polygon beach area	64
Massachusetts Water Resources Authority	State government	Boston Harbor bacteria counts	Varies, weekly or monthly.	Point data from defined sampling locations	56
Massachusetts Water Resources Authority	State government	Boston Harbor nutrient data: Ammonium, Nitrate + nitrite, Total Kjeldahl Nitrogen, Phosphate, Total phosphorus, Chlorophyll *a*, Phaeophytin.	Varies, weekly or monthly, Acquired data spanning 1992 -2014	Point data from defined sampling locations in Boston Harbor	56
Massachusetts Water Resources Authority	State government	Ammonium, Nitrate + nitrite, Phosphate, Total P/N, Particulate P/N/C, Chlorophyll *a*, Silicate, Salinity, zooplankton.	Varies, weekly or monthly, Acquired data spanning 1995 -2014	Point data from defined sampling locations in Massachusetts Bay	57

Table 4-1. Data Sources for Massachusetts Bay and Coastal Watersheds					
Source Name	**Source Type**	**Data Types**	**Sampling Frequency**	**Unit of spatial analysis**	**Source**
Massachusetts Water Resources Authority	State government	Beach water quality, bacteria counts and precipitation in the form of rainfall	Spring-fall. Daily during summer bathing season	Point, representative of polygon for beach bathing area	56
Massachusetts Water Resources Authority	State government	*Pseudo-nitzschia* species count data	Approximately monthly from 1992 -2014 (date range varies by station)	Point, from defined sampling locations	65
U.S. Environmental Protection Agency	Federal government	Location and information for facilities within a National Pollutant Discharge Elimination System (NPDES) permit.	Monthly	Point, facilities have associated latitude and longitude data	66
Massachusetts Department of Public Health	State government	Enteric diseases diagnosed in the Commonwealth of Massachusetts	Annually, with 1+ year lag for public release	U.S. State, Commonwealth of Massachusetts	67

Summary Conclusion.

This chapter discussed the changes in the type and amount of data available to environmental health researchers. These changes include increasing numbers of observations from multiple sources, including non-traditional sources such as satellites,

the digitization and publication of previously paper-based records, and the increasing speed of data generation from traditional sources as well as new sources such as social media or mobile phone-based applications. In addition, this chapter provided three examples of public health research benefitting from the acquisition and assimilation of data from sources not traditionally utilized by medical researchers. These three examples highlighted the use of satellite-derived remote sensing data related to regional climate conditions and how that data was combined with other data sources to develop models used to predict changes in risk probability for seasonal Rift Valley Fever in the Horn of Africa, the severity of cholera outbreaks in Bangladesh, and high-incidence days of Kawasaki disease in Japan. Diagrams describing the workflow employed in these examples of synthesis research were included, along with a generalized workflow diagram that other researchers might follow for their own project. Finally, this chapter presented a list of relevant data sets that have been assembled from authoritative sources in order to develop a basic model for hindcasting the presence of *Pseudo-nitzschia* species of diatoms in Massachusetts Bay and *Enterococcus* bacteria within the Boston Harbor embayment of Massachusetts Bay. Model specifics and results are described in Chapter 4.

Literature Cited.

1. Merriam-Webster Incorporated. 2015. Dictionary definition of synthesis. Encyclopedia Britannica. Springfield, M.A. http://www.merriam-webster.com/dictionary/synthesis (Accessed September 27, 2015).

2. Lohr S. 2013. The Origins of 'Big Data': An Etymological Detective Story. The New York Times Company. New York, NY.

3. Microsoft Corporation. 2016. Excel specifications and limits. Microsoft Corporation. Redmond, W.A. https://support.office.com/en-nz/article/Excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3 (Accessed January 13, 2016).

4. The Apache Software Foundation. 2015. Apache™ Hadoop®. The Apache Software Foundation. Forest Hills, M.D. https://hadoop.apache.org/ (Accessed January 13, 2016).

5. IBM. 2015. The Four V's of Big Data. IBM. Armonk, NY. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg (Accessed July 12, 2015).

6. Davenport T. H., D. J. Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century. Harv. Bus. Rev. 90: 70-76.

7. Press, G. 2013. A Very Short History Of Data Science. Forbes.com. http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/ (Accessed July 12, 2015).

8. IBM. 2015. What is a Data Scientist? IBM. Armonk, NY. http://www-01.ibm.com/software/data/infosphere/data-scientist/ (Accessed July 12, 2015).

9. U.S. Census Bureau, U.S. Dept of Commerce. 2015. U.S. Census Bureau Homepage. Washington, D.C. http://www.census.gov/# (Accessed March 13, 2015).

10. U.S. Census Bureau. 2012. Geographic Terms and Concepts - Block Groups. U.S. Department of Commerce. Washington, D.C. http://www.census.gov/geo/reference/gtc/gtc_bg.html (Accessed July 2, 2015).

11. Google Inc. 2015. Google Trends. Google, Inc. Mountain View, CA. https://www.google.com/trends/ (Accessed July 13, 2015).

12. Amazon Web Services, I. 2015. Amazon Web Services: Public Data Sets. Amazon Web Services. http://aws.amazon.com/datasets (Accessed July 13, 2015).

13. Dryad. 2015. Dryad: Frequently Asked Questions. Dryad. Durham, NC. http://datadryad.org/pages/faq (Accessed October 10, 2015).

14. Dryad. 2015. Dryad: The organization: Overview. Dryad. Durham, NC. http://datadryad.org/pages/organization (Accessed October 10, 2015).

15. U.S. Department of Health and Human Services. 2015. HealthData.Gov: About. U.S. Department of Health and Human Services;. Washington, D.C. http://www.healthdata.gov/content/about (Accessed July 13, 2015).

16. U.S. National Library of Medicine. 2015. PubMed. National Institutes of Health. Bethesda, MD. http://www.ncbi.nlm.nih.gov/pubmed (Accessed May 31, 2015).

17. National Center for Biotechnology Information. 2014. GenBank Overview. U.S. National Library of Medicine. Bethesda, MD. http://www.ncbi.nlm.nih.gov/genbank/ (Accessed May 31, 2015).

18. Centers for Disease Control and Prevention. 2013. The Foodborne Outbreak Online Database (FOOD Tool). Centers for Disease Control and Prevention. Atlanta, GA. http://www.cdc.gov/foodsafety/fdoss/data/food.html (Accessed May 31, 2015).

19. United Network for Organ Sharing. 2015. United Network for Organ Sharing: About Us. United Network for Organ Sharing. Richmond, VA. http://www.unos.org/about/index.php (Accessed May 31, 2015).

20. Ploutz, M. 2015. Interview of Dr. Michelle Ploutz by Marin Kress. M. Kress.

21. National Oceanic and Atmospheric Administration. 2014. Integrated Ocean Observing System: About (IOOS). National Oceanic and Atmospheric Administration. Washington, D.C. http://www.ioos.noaa.gov/about/welcome.html (Accessed August 8, 2015).

22. Ecological Society of America. 2015. Ecological Society of America Data Registry. Ecological Society of America. Washington, DC. http://data.esa.org/esa/style/skins/esa/index.jsp (Accessed May 31, 2015).

23. Boenisch, G., Kattge, J. 2014. TRY Plant Trait Database: About. Max Planck Institute for Biogeochemistry. Jena, Germany. https://www.try-db.org/TryWeb/About.php (Accessed May 31, 2015).

24. University of Liverpool. 2015. ENHanCEd Infectious Diseases (EID2) Database. University of Liverpool. Liverpool, UK. http://www.zoonosis.ac.uk/eid2 (Accessed May 31, 2015).

25. National Oceanic and Atmospheric Administration. 2012. NS&T Program Download Page. National Oceanic and Atmospheric Administration. Silver Spring, MD. http://ccma.nos.noaa.gov/about/coast/nsandt/download.aspx (Accessed Jun 24, 2015).

26. National Aeronautics and Space Administration. 2014. NASA's Earth Observing System. National Aeronautics and Space Administration. Washington, DC. http://eospso.gsfc.nasa.gov/content/nasa-earth-science-data (Accessed May 31, 2015).

27. European Space Agency. 2015. ESA Earth Online. European Space Agency. Paris, France. https://earth.esa.int/web/guest/data-access (Accessed July 13, 2015).

28. Merriam-Webster Incorporated. 2015. Crowdsource: Definition. Merriam-Webster, Incorporated. http://www.merriam-webster.com/dictionary/crowdsourcing (Accessed March 8, 2015).

29. University of Washington. The Science Behind FoldIt. University of Washington. Washington, USA. http://fold.it/portal/info/about (Accessed March 9, 2015).

30. Encyclopedia of Life Contributors. 2015. Encyclopedia of Life. Encyclopedia of Life Secretariat. Washington, DC. http://eol.org/about (Accessed May 31, 2015).

31. Butler D. 2013. When Google got flu wrong. Nature. 494: 155-156.

32. Instagram. 2015. Instagram. Instagram. Menlo Park, CA. https://instagram.com/# (Accessed March 9, 2015).

33. Facebook. 2015. Facebook: Homepage. Facebook. California, USA. https://www.facebook.com/?_rdr (Accessed March 9, 2015).

34. Marr, B. 2014. Big Data: 20 Free Big Data Sources Everyone Should Know. Social Media Today LLC. http://www.smartdatacollective.com/bernardmarr/235366/big-data-20-free-big-data-sources-everyone-should-know .

35. Google Inc. 2011. Google Correlate. Google, Inc. Mountain View, CA. http://www.google.com/trends/correlate (Accessed March 9, 2015).

36. Lazer D. M., R. Kennedy, G. King, A. Vespignani. 2014. The parable of Google Flu: traps in big data analysis. Science. 343: 1203-1205.

37. Comstock, J. 2014. Google Flu Trends will supplement crowdsourced approach with CDC data. Chester Street Publishing, Inc. Cambridge, MA. http://mobihealthnews.com/37836/google-flu-trends-will-supplement-crowdsourced-approach-with-cdc-data/ (Accessed March 8, 2015).

38. Madrigal A. C. 2014. In Defense of Google Flu Trends. The Atlantic. The Atlantic Monthly Group.

39. Ginsberg J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature. 457: 1012-1014.

40. Centers for Disease Control and Prevention. 2015. Overview of Influenza Surveillance in the United States. U.S. Department of Health and Human Services. Atlanta, G.A. http://www.cdc.gov/flu/weekly/overview.htm .

41. Centers for Disease Control and Prevention. 2015. Epidemiology and Prevention of Vaccine-Preventable Diseases: Influenza. (The Pink Book: Course Textbook , 13th Edition (2015)). U.S. Department of Health and Human Services. Atlanta, G.A. http://www.cdc.gov/vaccines/pubs/pinkbook/flu.html (Accessed January 13, 2016).

42. Proff R., K. Gershman, D. Lezotte, A. C. Nyquist. 2009. Case-based surveillance of influenza hospitalizations during 2004-2008, Colorado, USA. Emerg. Infect. Dis. 15: 892-898.

43. Stefansen, C. 2014. Google Flu Trends gets a brand new engine. Google. Mountain View, CA. http://googleresearch.blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html (Accessed March 10, 2015).

44. Twitter Inc. 2015. Twitter Search. Twitter, Inc. San Francisco, CA. https://twitter.com/search-home (Accessed March 10, 2015).

45. Twitter Inc. 2015. About Twitter, Inc. Twitter, Inc. San Francisco, CA. https://about.twitter.com/company (Accessed March 11, 2015).

46. Barth, D., Google Inc. 2009. The bright side of sitting in traffic: Crowdsourcing road congestion data. Google Inc. Mountain View, C.A. http://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html (Accessed September 19, 2015).

47. Kays R., M. C. Crofoot, W. Jetz, M. Wikelski. 2015. Terrestrial animal tracking as an eye on life and planet. Science. 348: 1222.

48. Habib S., M. Plessis-Fraissard, S. D. Ambrose. 2008. Space-based Earth observations for societal benefit. WMO Bulletin. 57: 22-28.

49. Anyamba A., J. P. Chretien, J. Small, C. J. Tucker, P. B. Formenty, J. H. Richardson, S. C. Britch, D. C. Schnabel, R. L. Erickson, K. J. Linthicum. 2009. Prediction of a Rift Valley fever outbreak. Proc. Natl. Acad. Sci. U. S. A. 106: 955-959.

50. Koelle K., X. Rodó, M. Pascual, M. Yunus, G. Mostafa. 2005. Refractory periods and climate forcing in cholera dynamics. Nature. 436: 696-700.

51. Pascual M., X. Rodó, S. P. Ellner, R. Colwell, M. J. Bouma. 2000. Cholera dynamics and El Nino-Southern Oscillation. Science. 289: 1766-1769.

52. Huq A., R. B. Sack, A. Nizam, I. M. Longini, G. B. Nair, A. Ali, J. G. Morris Jr, M. Khan, A. Siddique, M. Yunus. 2005. Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. Appl. Environ. Microbiol. 71: 4645-4654.

53. Rodó X., R. Curcoll, M. Robinson, J. Ballester, J. C. Burns, D. R. Cayan, W. I. Lipkin, B. L. Williams, M. Couto-Rodriguez, Y. Nakamura, R. Uehara, H. Tanimoto, J. A. Morgui. 2014. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proc. Natl. Acad. Sci. U. S. A. 111: 7952-7957.

54. Fekete B. M., R. D. Robarts, M. Kumagai, H. P. Nachtnebel, E. Odada, A. V. Zhulidov. 2015. Time for in situ renaissance. Science. 349: 685-686.

55. Hunt C. D., D. G. Borkman, P. S. Libby, R. Lacouture, J. T. Turner, M. J. Mickelson. 2010. Phytoplankton patterns in Massachusetts Bay—1992–2007. Estuaries and Coasts. 33: 448-470.

56. Massachusetts Water Resources Authority. 2015. Boston Harbor and Massachusetts Bay: Water Quality Data
. Massachusetts Water Resources Authority. Boston, MA.
http://www.mwra.state.ma.us/harbor/html/wq_data.htm (Accessed May 17, 2015).

57. Massachusetts Water Resources Authority. . 2015. 1995-2015_F22_F23.xlsx [MS Excel file]. M. Kress.

58. National Oceanic and Atmospheric Administration. 2015. National Data Buoy Center: Station BHBM3 - 8443970 - Boston, MA. U.S. Department of Commerce. Stennis Space Center, MS.
http://www.ndbc.noaa.gov/station_page.php?station=bhbm3 (Accessed May 25, 2015).

59. Northeastern Regional Association of Coastal and Ocean Observing Systems. 2014. NERACOOS: Data & Tools. The Gulf of Maine Research Institute. Portland, ME.
http://neracoos.org/datatools (Accessed May 25, 2015).

60. U.S. Census Bureau. 2014. American Community Survey: About. U.S. Census Bureau. Washington, DC. http://www.census.gov/acs/www/about_the_survey/american_community_survey/ (Accessed May 24, 2015).

61. U.S. Geological Survey. 2015. National Water Information System. U.S. Geological Survey. Washington, DC. http://waterdata.usgs.gov/nwis (Accessed May 24, 2015).

62. U.S. Geological Survey. 2015. National Water Information System: Mapper. U.S. Geological Survey. Washington, DC. http://maps.waterdata.usgs.gov/mapper/index.html (Accessed May 24, 2015).

63. National Oceanic and Atmospheric Information, National Centers for Environmental Information. 2015. National Centers for Environmental Information: Climate Data Online Search. U.S. Department of Commerce. Silver Spring, M.D. http://www.ncdc.noaa.gov/cdo-web/confirmation (Accessed December 31, 2014).

64. Massachusetts Department of Public Health. 2012. BeachWaterQualityData_2003-2011.xls [MS Excel file]. Anonymous .

65. Massachusetts Water Resources Authority. 2015. pseudonitz_1992-2014.xlsx [MS Excel file]. M. Kress.

66. U.S. Environmental Protection Agency. 2015. Envirofacts: Data Downloads: Custom Search - PCS. U.S. Environmental Protection Agency. Washington, D.C. http://www.epa.gov/enviro/facts/datadownloads.html (Accessed September 25, 2015).

67. Massachusetts Department of Public Health. 2013. Enteric Disease in Massachusetts: 1999-2013. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eohhs/docs/dph/cdc/foodsafety-enterics-state-totals.pdf (Accessed March 17, 2015).

CHAPTER 5

MODEL DEVELOPMENT AND TESTING

Abstract. This chapter discusses the development of a quantitative hypothesis-driven approach to explain potential predictive influences for two marine-sourced risks known to exist in Massachusetts Bay. Those two risks are the diatom *Pseudo-nitzschia delicatissima* complex which can produce the neurotoxin Domoic Acid, and *Enterococcus* bacteria which are the standard indicator bacteria used to assess recreational water quality and are associated with mammalian fecal pollution. The *P. delicatissima* complex model is based on data from two stations approximately 20 miles apart. The *Enterococcus* model is based on data from three ocean-facing beach locations along the north coast of Massachusetts Bay because of their proximity to the offshore sampling site for *P. delicatissima* complex and their location as the 'upstream' end of the general circulation pattern for the Bay. We identified potential explanatory variables through the literature review (described in Chapter 2) and then identified available data, sourced primarily from state and federal monitoring programs (described in Chapter 3). Testing of the probabilistic models derived from these data sources revealed that, for *P. delicatissima* complex the presence/absence can be poorly-to-adequately predicted, and for *Enterococcus* presence/absence at the level of >10 bacteria per 100mL seawater

cannot be predicted with any confidence. Analysis of the data did not reveal any discernable relationship between the presence of *Enterococcus* in recreational waters at the sampled locations and the presence of *P. delicatissima* complex at Station F22 when sampled in the same month. At present, the use of *Enterococcus* and other fecal indicator bacteria as an indicator of biological water quality is not informative of the presence of *P. delicatissima* complex. The same lack of relationship is expected for other fecal indicator bacteria and other species of *Pseudo-nitzschia*. These results suggest that direct sampling for marine-sourced risks in recreational- and shellfish-harvesting wasters is the most appropriate monitoring action for protecting public health at present.

Introduction.

The preceding chapters have described how to think about an environmental health topic in terms of an overarching framework that places the topic within a larger system (Chapter 1), the biology and state of knowledge of five marine-sourced risks that can exist in Massachusetts Bay (Chapter 2), and the way that the changing availability of information sources beyond traditional epidemiological data allows us to explore new questions the environment and human health (Chapter 3). This chapter synthesizes the information from the previous chapters into an information theoretic framework to determine the probability of the presence/absences for two marine-sourced risks in Massachusetts Bay. The four specific questions we are addressing with these models are:

1) Is it possible to hindcast levels of *Enterococcus* populations in specific areas of Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Enterococcus* growth?

2) Is it possible to hindcast levels of total *Pseudo-nitzschia* populations measured in Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Pseudo-nitzschia* growth?

3) Does there appear to be any clear relationship between *Enterococcus* levels and *Pseudo-nitzschia* levels in Massachusetts Bay?

4) Are there any field measurements for which public data do not readily exist which scientific literature suggests would likely increase the predictive ability of these models?

In the previous chapter we presented a figure describing the three phases of this type of interdisciplinary environmental health/data science work. Those phases are:

- Phase 1: Explore concepts and generate hypothesis
- Phase 2: Develop outputs
- Phase 3: Evaluate outputs

Our process up until this point is depicted in Figure 5-1, below, in the section titled 'Phase 1'. In Phase 1 we identified our topic of interest, carried out a literature review and data gathering process to understand the current state of those risks and how they play out in Massachusetts and the Massachusetts Bay area (our spatial area of interest). We did not pre-select a temporal scale in Phase 1, but waited until Phase 2 after we had examined the assembled data. The section of Figure 5-1 depicting Phase 2 shows the variety of products that have been produced as part of this research (graphics, maps, and candidate model sets), some of these outputs have been included in previous chapters (examples include the series of maps in Chapter 2).

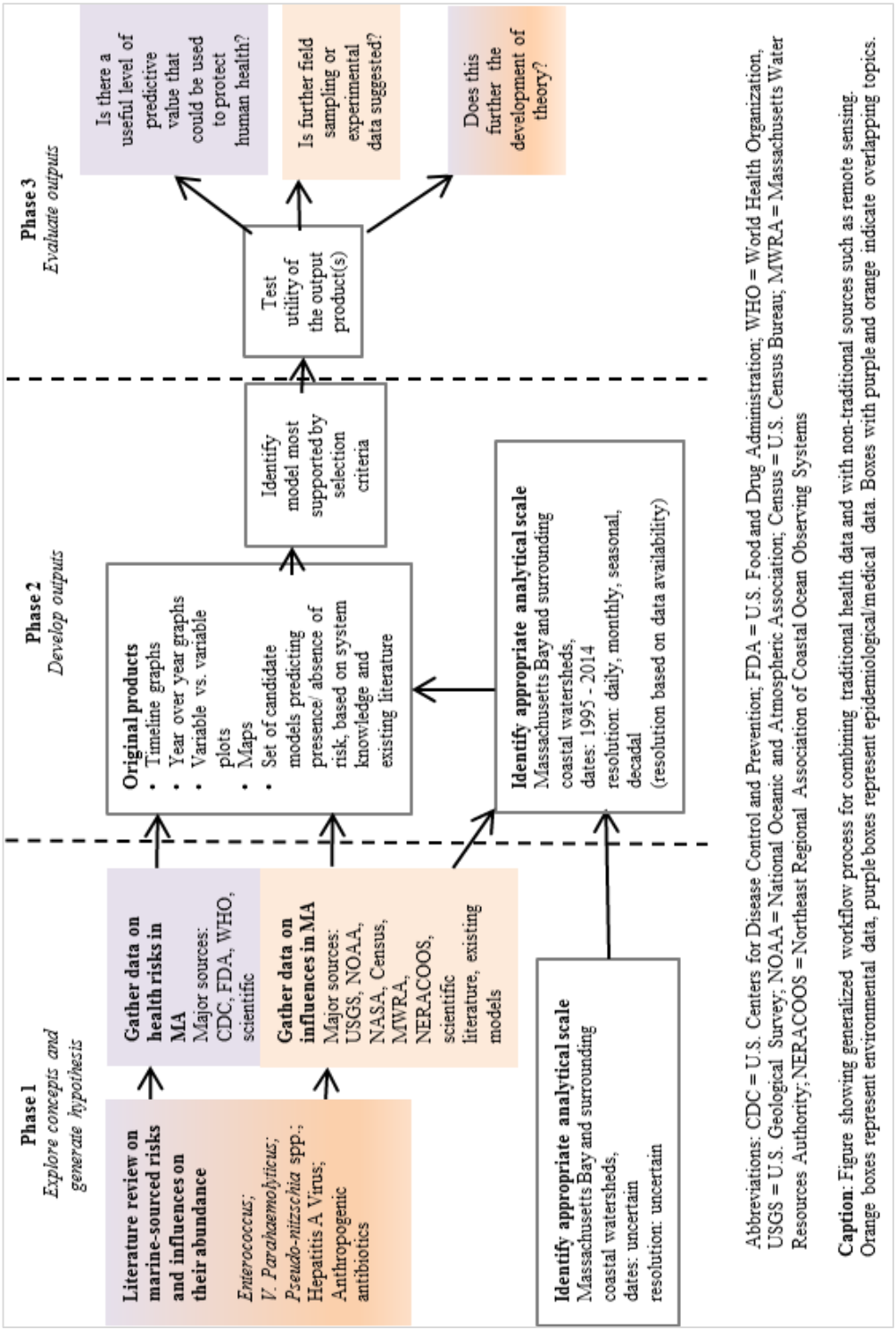

Figure 5-1. Depiction of three phases of interdisciplinary data science with status of our work to this point.

The remainder of this chapter will discuss the Phase 2 development work, then the Phase 3 evaluation work, and will end with a summary conclusion section.

Phase 2: Develop Outputs

The first task in Phase 2 was to examine the available Massachusetts Bay data and compare it to other studies that have investigated the relationship between *Pseudo-nitzschia* and environmental variables. Four examples of such work from the past decade are summarized below, notably these four examples all used different analytical approaches. This reflects both the diversity of statistical methods available to researchers and the lack of a widely accepted standard approach for describing environmental influences on *Pseudo-nitzschia* abundance. As diatoms *Pseudo-nitzschia* are taxonomically distant from more well-known causative agents of harmful algal blooms, namely the dinoflagellates *Alexandrium fundyense* and *Karenia brevis* which belong to a different phylum.[1]

Canonical correspondence analysis and *Pseudo-nitzschia* species in the Quoddy Region, Bay of Fundy, Canada. Kaczmarska et al. (2007) examined the relationship between *P. delicatissima* and environmental factors is the Quoddy Region, Bay of Fundy, Canada. The authors identified seven species of *Pseudo-nitzschia* in samples at 5 stations collected weekly for 11 weeks (from 29 July to 14 October, 2003) and then related species abundance to environmental factors.[2] The environmental

variables investigated were transparency, fluorescence, silicate, phosphate, nitrite plus nitrate, ammonia, nitrite, oxygen, sigma-t, tidal level, tidal state and total depth of water column at sampling site.[2] The seven *Pseudo-nitzschia* species were clustered into three groups based on morphology, the *P. seriata*-group, *P. delicatissima*-group and *P. americana* (a group containing a single species). The authors note that species from the *P. delicatissima*-group dominated most of their samples, with *P. delicatissima* being the most temporally persistent.[2] The authors used canonical correspondence analysis (CCA) to identify environmental factors that explained the greatest amount of variance in temporal and spatial distribution patterns at both the species and group levels. A total of 52.4% of the variance in species data was explained by the first four CCA extracted ordination axes (27.3, 18.4, 4.0, and 2.7% respectively).[2] To determine statistical significance between species abundance and an environmental gradient Kaczmarska et al. constructed a biplot of *t*-values using the Van Dobben method and found that of eleven environmental variables tested, only 'nitrite plus nitrate' was significantly (positively) correlated with *P. delicatissima* and *P. pseudodelicatissima*.[2] The same variable was significantly negatively correlated with *P. pungens*, and not significantly correlated with any other species. Overall, the authors found that different environmental variables correlated with different *Pseudo-nitzschia* species, suggesting that each may exploit distinct environmental conditions.[2]

Logistic regression model for the prediction of toxigenic *Pseudo-nitzschia* blooms in Monterey Bay, California. Lane et al. (2009) used logistic regression to

model relationships between *Pseudo-nitzschia* blooms and environmental factors in Monterey Bay, California, which the authors note was first time logistic regression had been applied to *Pseudo-nitzschia* bloom prediction.[3] The authors transformed continuous *Pseudo-nitzschia* count data into a dichotomous response variable (bloom / no bloom) using a threshold of 10,000 *Pseudo-nitzschia* cells per liter seawater.[3] Their analysis considered 31 environmental variables, but only 6 variables were identified as statistically significant across the three models developed (annual, spring-summer, and fall-winter models). The six significant variables were water temperature, the upwelling index level, the natural log of chlorophyll *a*, natural log of silicic acid, natural log of the Pajaro River freshwater discharge, and nitrate concentration. Only two variables, natural log of chlorophyll *a* and natural log of silicic acid, were significant in all three models.[3] The annual model was built using 422 total cases (of which 64 bloom cases), performance at the optimized prediction point of 0.145 resulted in 5% false negatives and 62% false positives.[3] Massachusetts Bay does not experience the same intensity of upwelling as Monterey Bay, an important distinction between the two areas that must be considered when attempting to relate findings from one region to another.

Generalized Linear Model for predicting *Pseudo-nitzschia* blooms in the Chesapeake Bay. Anderson et al. (2010) used logistic regression to predict a dichotomous response variable (bloom / no bloom) at three threshold levels for *Pseudo-nitzschia* species in the Chesapeake Bay on the eastern coast of the United States.[4] The

bloom threshold levels explored were small (≥10cells/mL), medium (100 cells/mL), and large (1,000 cells/mL) at any time of entire year (they did not build separate seasonal models like Lane et al. (2009)).[3; 4] The authors utilized surface phytoplankton abundance and water quality data to build their model, noting that the data were originally collected for monitoring purposes and thus subject to sampling biases due to the frequency of "event-response" type of data collection.[4] Environmental variables identified by Anderson et al. (2010) for small blooms were month of year, water temperature, latitude, longitude, freshwater discharge, phosphate concentration, and nitrate plus nitrite concentration.[4] Medium-bloom model variables were month of year, water temperature, salinity, freshwater discharge, phosphate, dissolved organic carbon concentration, nitrate. Large-bloom model variables were water temperature, salinity, latitude within Chesapeake Bay, silicic acid, nitrate plus nitrate, and turbidity as measured by a Secchi disk.[4] The small bloom model performed best of all three and was the main focus of their analysis. At the optimized prediction point of 0.19 the small bloom model had a Heidke Skill Score of 0.53, probability of detection score of 0.75, false alarm ratio of 0.52, and probability of false detection score of 0.09.[4]

Although the Chesapeake Bay and Massachusetts Bay are both the eastern coast of the U.S., the study sites in Anderson *et al.* (2010) have environmental characteristics that differ from Massachusetts Bay, notably a salinity range of 0.5 to >18psu,[4] and recorded sea surface temperatures in Chesapeake Bay ranging from approximately 0 to 30ºC.[5] Massachusetts Bay sea surface temperature records from NOAA Buoy 44013

from 2010 to 2014 do not indicate any temperature above 25°C.[6] Massachusetts Bay also has a different salinity profile than the Chesapeake Bay. Salinity records from Station 142 at the mouth of Boston Harbor indicate salinity ranging from 25 to 35 psu,[7] and offshore surface salinity records from Buoy A01 range from approximately 24 to 32 psu.[8] Again, these regional differences may limit the transferability of findings between regions.

BEST analysis examination of variation in *Pseudo-nitzschia* species in the Western English Channel. Downes-Tettmar et al. (2013) collected weekly samples of phytoplankton and environmental parameters at Station L4 (50°15'N, 4°13'W) in the Western English Channel from January to December 2009 and divided *Pseudo-nitzschia* into three groups based on size and morphology.[9] The three groups were the *P. delicatissima*-group, *P. pungens/multiseries*-group, and *P. seriata*-group. The authors used a technique known as BEST analysis, which examines the similarities between pairs of samples (in this case the abundance of *Pseudo-nitzschia* group species) and pairs of variables (environmental data).[9] Although this may sound similar to the ordination plots produced by methods such as canonical correspondence analysis, the authors note that their data did not meet the assumptions required for the use of that technique. Unlike the logistic approaches used by Lane et al. (2009) and Anderson et al. (2010) to develop predictive models, the BEST analysis used by Downes-Tettmar et al. (2013) does not result in an equation to predict presence or absence of a bloom. However, the BEST

analysis and Spearman's rank correlation (between *Pseudo-nitzschia* group abundances and environmental variables) used by the authors did result in a list of environmental variables correlated to group abundance, along with the results of a significance test at $p < 0.02$. For the *P. delicatissima*-group the significant environmental variables (with Spearman's rank correlations) were water temperature (0.54), hours of light (0.67), salinity (0.56), phosphate concentration (-0.69), and rainfall (-0.22).[9] The abundance of each *Pseudo-nitzschia* group was associated with different environmental variables, and no single variable was shown to be significant for all three groups.[9] This supports the argument for group-level, or preferably species-level, identification for use in model development as discussed by Downes-Tettmar et al. (2013) and others, rather than total *Pseudo-nitzschia* spp. abundance.[92]

Section Summary. As illustrated in the four examples summarized above, there are multiple ways to assess the relationship between *Pseudo-nitzschia* abundance and environmental factors.[2-4; 9] These research projects have been carried out in different regions with different environmental regimes, and Table 5-1, below, shows that they have yielded different results in terms of environmental variables identified as important for *Pseudo-nitzschia* abundance. As yet there is no single best way to examine the relationship between environmental factors, *Pseudo-nitzschia* presence or bloom size, and Domoic Acid production.

Table 5-1. Variables identified in recent modeling and correlation work

If reported, variables are identified as positively significant (++), negatively significant (--), positively associated but not significant (+), negatively associated by not significant (-).				
Paper	Kaczmarska et al. (2007)[2]	Lane et al. (2009), annual model[3]	Anderson et al. (2010), small bloom model[4]	Downes-Tettmar et al. (2013)[9]
Method	Canonical correspondence analysis	Logistic regression	Generalized linear model	BEST analysis, Spearman rank
***Pseudo-nitzschia* type measured**	*P. delicatissima*	Total *Pseudo-nitzschia*	Total *Pseudo-nitzschia*	*P. delicatissima*-group
Variables Water Temperature		--	--	++
Maximum light				+
Hours of light				++
Salinity				++
Nitrate				-
Nitrite				-
Nitrate + nitrite	++		--	
Phosphate			--	--
Ammonia				-
Silicate				-
Silicic Acid		--		
Chlorophyll *a*		++		+
Rainfall				--
Upwelling		++		
Month			--	
Latitude			--	
Longitude			++	
Freshwater discharge			--	

Given the variety of environmental conditions, some researchers have focused on untangling the relationship between individual environmental variables and abundance at

the *Pseudo-nitzschia* group or species level, using techniques such as canonical correspondence analysis or BEST analysis.[2; 9] Other researchers interested in predictive capacity have used regression methods to develop models which may be specific to bloom size, times of year, morphological groups, or individual species.[3; 4] The goal of our work is to develop a predictive model which could be used to support public health protection efforts in Massachusetts Bay. We are not attempting to delineate the mechanisms of action by which individual environmental factors influence *Pseudo-nitzschia* abundance, although we recognize that model predictions might serve to generate hypothesis which may be tested by others. With that goal in mind, we decided to use a model selection process that differs from those presented above. The model selection process (Phase 2 work), and then our model testing results (Phase 3 evaluation work), are described below.

Phase 2: Model Development Using Information Theory

We have chosen to use an information-theoretic approach for model selection that seeks to identify the 'best approximating model' from a suite of candidate models.[10] The candidate set of models is developed by choosing models based on our current understanding of the phenomenological processes that affect *Pseudo-nitzschia* or *Enterococcus* abundance. The models in that candidate set are then ranked relative to one another using information criteria (there are multiple kinds) to identify the best approximating model and calculate the difference between the models in the candidate

set.[10] We will use Akaike's Information Criterion (AIC)[11] as our information criteria. This section will give a brief overview of information theory and AIC to provide context for the remainder of the chapter. For an extensive treatment on these and other topics including information and likelihood theory, weight of evidence approaches, and the difference between these and other multivariate modeling approaches the reader is referred to the book Model Selection and Multi-Model Inference: A Practical Theoretic Approach (2nd Edition) by K.P. Burnham and D.R. Anderson.[10]

Information theory is a concept that arose in the 1940s, but model selection based on information theory has only been introduced into biology and ecology relatively recently.[10] It is philosophically different than null hypothesis testing and has no equivalent dichotomy of 'significant' or 'non-significant' variables.[10] Together, the concept of information theory and the use of AIC provides a general, yet powerful, method for selecting a model for the data of interest. This approach differs from other model development methods that sequentially add variables to a model based on significance tests because it requires the researcher to develop an *a priori* set of candidate models (model specification) based on their existing understanding of the system.[10] That is, variable inclusion is based on an understanding of biological relevance rather than statistical significance. These candidate models may include different variables and interaction terms, so variable selection is a key part of the model development, and ultimately model selection.[10]

After the candidate model set is specified all of the models in that set are ranked relative to one another using AIC to identify the 'most supported model.' When multiple models receive similar levels of support because their AIC scores are close together the process of 'model averaging' allows this uncertainty among models to be considered when obtaining parameter estimates.[12] By using model averaging (also called multi-model inference) it is possible to make inferences from several models in the candidate set.[10]

AIC scoring is a mathematical process that calculates a distance from a candidate model to an (unknown) constant representing the 'true' model with parameters that reflect reality.[10] The model with the lowest distance among all the models within a candidate set is ranked highest. AIC does not give an absolute measure of how good a particular model is, only whether one model is better than another model in the candidate set. Since AIC is provides a relative comparison among models in the set, the variables and models included in that set must be carefully selected. A fundamental part of AIC is based upon understanding the concept of the 'relative distance' (a mathematical construct based on the entire distribution of the model) between a model and full reality. The equation for this relative directed distance (also known as the Kullback-Leibler distance) between full reality (*f*) and a model (*g*) estimating that reality is[10],

$$I(f,g) - C = E_f[\log(g(x|\theta))]$$

Where: $I(f, g)$ denotes the information lost when approximating f using g

C denotes the unknown true distribution, or E_f [log (f (x))]
E_f denotes a statistical expectation
x denotes a set of data
θ denotes parameters of the candidate models, the model space
log denotes the natural logarithm

The left side of the equation (the relative distance) remain unknown, but the right side can be quantified given the model and the data. There is no parameter estimation at this point.[10] The equation for an AIC score goes beyond the equation above for relative directed distance to provide an estimate of the expected relative distance between the fitted model and the (unknown) truth that underlies the observed data.[10] The equation for the AIC score is shown below.

$$AIC = -2log\left(\mathcal{L}(\hat{\theta}|y)\right) + 2K$$

Where: $\mathcal{L}$ denotes likelihood
$log\left(\mathcal{L}(\hat{\theta}|y)\right)$ is the value of the log-likelihood at its maximum point
$\hat{\theta}$ denotes estimated parameters
y denotes an independent random sample from the distribution
K denotes the number of estimable parameters

The application of AIC to model selection is straightforward, AIC values are computed for each model, then models are ranked based on AIC scores. The model with the lowest AIC is estimated to be the closest to that unknown reality among the models in the set.[10] In addition to the AIC, the AIC differences (written as ΔAIC) are an important part of model selection consideration. ΔAIC is the difference in AIC scores between the 'most supported model' and each of the other models in that set. ΔAIC values among one candidate set are not comparable to the ΔAIC of another candidate set. As ΔAIC

increases there is diminishing support that a fitted model is the best model, given the data. For nested models (which may contain some of the same parameters) the guideline for evaluating ΔAIC is that values from 0 to 2 indicate that a model is substantially similar to the best model in the set (the model with the lowest AIC) and is considered indistinguishable from the best model, values of 4 to 7 indicate considerably less support for that model (but there is still a likelihood that is it a supported model), and ΔAIC values greater than 10 indicate essentially no support for a model over the best model in the set. We will apply this model selection method after generating a set of candidate models for the presence/absence of *P. delicatissima* complex and the presence/absence of *Enterococcus*, our response variables of interest (the next section describes the response variable data in further detail). After identifying the most supported model from each candidate set we will use it to generate probabilities of *P. delicatissima* complex or *Enterococcus* presence with values from the test data set, i.e., we will hindcast the presence of these taxa. As part of Phase 3 evaluation we will compare the predictive (hindcast) accuracy of each model against the observed presence/absence outcomes. This comparison of predicted vs. observed outcomes will have four parts (model sensitivity, model specificity, false positive rate, and false negative rate) and will be discussed in the section on Phase 3 work.

Phase 2: Response Data Description

After assembling data from multiple sources our examination of the combined data revealed typical problems such as missing observations within a larger data set and temporal discontinuity between different data sets. For example, land-based weather observations and buoy-based oceanographic observations are collected sub-daily, with daily summaries available for different points for many years. Similarly, turbidity measurements from Buoy A01 are not available before 2007, so they were applicable to the *Enterococcus* model but not the *P. delicatissima* complex model. In addition, model-derived historical estimates of oceanographic variables (including chlorophyll, nitrates, diatoms) based on satellite data are available at a geographical scale that dwarfs our study area, and a monthly temporal scale lacking coverage for many years. Due to these limitations we did not use historical satellite-derived information for our model development.

The data on *Pseudo-nitzschia* counts from the MWRA were collected approximately monthly, across all seasons, but not in every month of every year. Macronutrient and phytoplankton samples were collected concurrently with *Pseudo-nitzschia* observations, but not in the days prior. Water quality samples for most recreational beaches are collected weekly, but only in the summer (roughly late May through September). These and other conditions required us to pare down the available data so that for the *P. delicatissima* complex model we had a total of 229 observations combined from Stations F22 and F23 spanning the years 1995 to 2014. 75% of these

observations were used for the training data set, and 25% were set aside for the testing data set. For the *Enterococcus* model we had 349 observations combined from three sampling points (Devereux Beach, Marblehead; Singing Beach Station 1, Manchester-By-The-Sea; and Good Harbor Beach, Gloucester) across summer months in the years 2007 – 2014, with 25% of those observations set aside for the test data set. Further details about the response variable data sets are provided below.

***Pseudo-nitzschia* Abundance Data.** The Massachusetts Water Resources Authority has multiple monitoring stations throughout Boston Harbor and Massachusetts Bay.[7] However, this work utilized sampling results collected between 1995 and 2014 at Station F23 (latitude: 42.339, longitude: -70.942) and Station F22 (latitude: 42.4798, longitude: -70.617).[13] These two stations are shown on Figure 5-2, below. Also shown on Figure 5-2 are Buoy A01 (latitude: 42.521, longitude: -70.565) and NOAA Buoy 44013 (latitude: 42.346, longitude: - 70.651) which provide oceanographic data and Boston Logan International Airport which houses an observing station that provides weather data.[8; 14; 15]

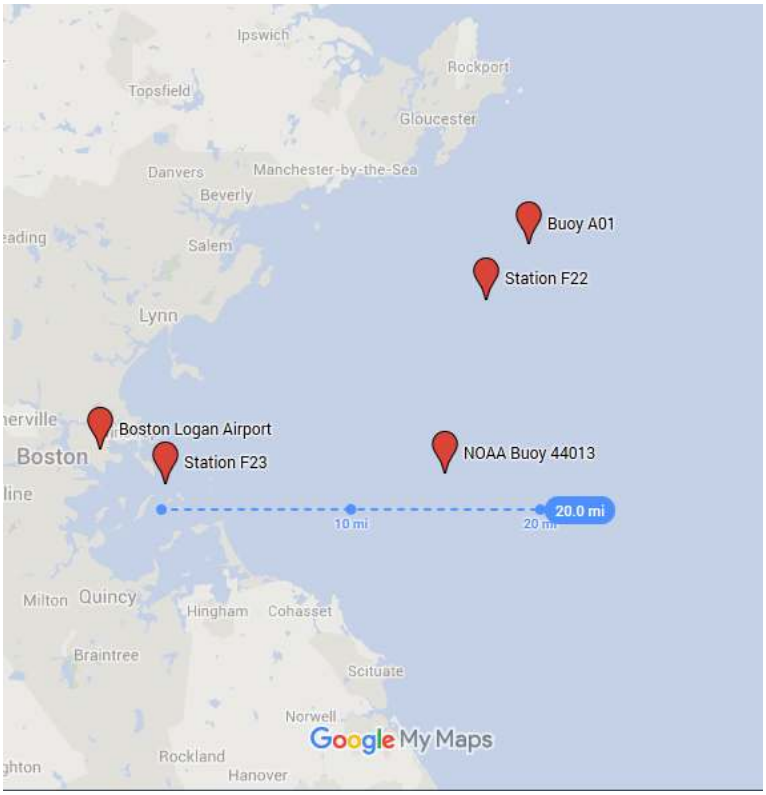

Figure 5-2. Map of Station F22 and F23, Buoy A01, Buoy 44013, and Boston Logan Airport locations.

The Massachusetts Water Resources Authority (MWRA) database lists eight *Pseudo-nitzschia* species or group category descriptions used in the monitoring program (see Table 5-2). By examining the observations of *Pseudo-nitzschia* categories we see that only two have been commonly detected in sampling efforts at Stations F22 and F23, the *P. delicatissima* complex and *P. pungens*. *Pseudo-nitzschia* genus taxonomy has changed during the period of MWRA monitoring in Massachusetts Bay, so some shifts in the abundance of different categories might be the result of changes in organism classification (referred to as 'binning') practices or from changes in project staff.[16] Also,

MWRA sampling at Station F23 began in 1992, but we selected August 1995 as our dataset start date because historical records for *Pseudo-nitzschia*, macronutrients, and other variables appeared more consistent after that date. Throughout the rest of this work any references to a start date of 1995 should be interpreted as starting in August 1995.

Table 5-2. Massachusetts Water Resources Authority categories for *Pseudo-nitzschia* species classification. Results of shallowest surface sample per day at Stations F22 and F23, August 1995 – December 2014.

	129 Samples at F23 (Outer Boston Harbor)		100 Samples at F22 (Massachusetts Bay)	
	Number of samples		Number of samples	
Category	Count = 0	Count ≠ 0	Count = 0	Count ≠ 0
Pseudo-nitzschia spp.	124	5	98	2
Pseudo-nitzschia sp. 1 (*delicatissima*?)	129	0	100	0
Pseudo-nitzschia delicatissima	129	0	100	0
Pseudo-nitzschia delicatissima complex	74	55	47	53
Pseudo-nitzschia pungens	104	25	86	14
Pseudo-nitzschia cf. *pungens*	128	1	100	0
Pseudo-nitzschia seriata	129	0	100	0
Pseudo-nitzschia cf. *americana*	129	0	99	1
Note: all sample counts are in units of cells/L				

Clearly, the *P. delicatissima* complex has been detected more often than *P. pungens* at Stations F23 and F22, it has the highest number of samples where the count was not equal to 0 cells/L (see Table 5-2). Past research has suggested that the *P. delicatissima* and the *P. delicatissima*-group prefer slightly different environmental conditions than *P. pungens*, with the result that their abundance peaks are temporally separated (see Downes-Tettmar et al. (2013)).[9] As part of Phase 2 output development work we

graphed *P. delicatissima* complex and *P. pungens* on separate timelines (see Figure 5-3, below). In Figure 5-3, the top two timelines show observations at Station F23 from August 1995 to 2014; the bottom two timelines show samples taken at Station F22 from 2000 to 2014. Note that sampling at *Pseudo-nitzschia* sampling at Station F22 did not start until the year 2000.

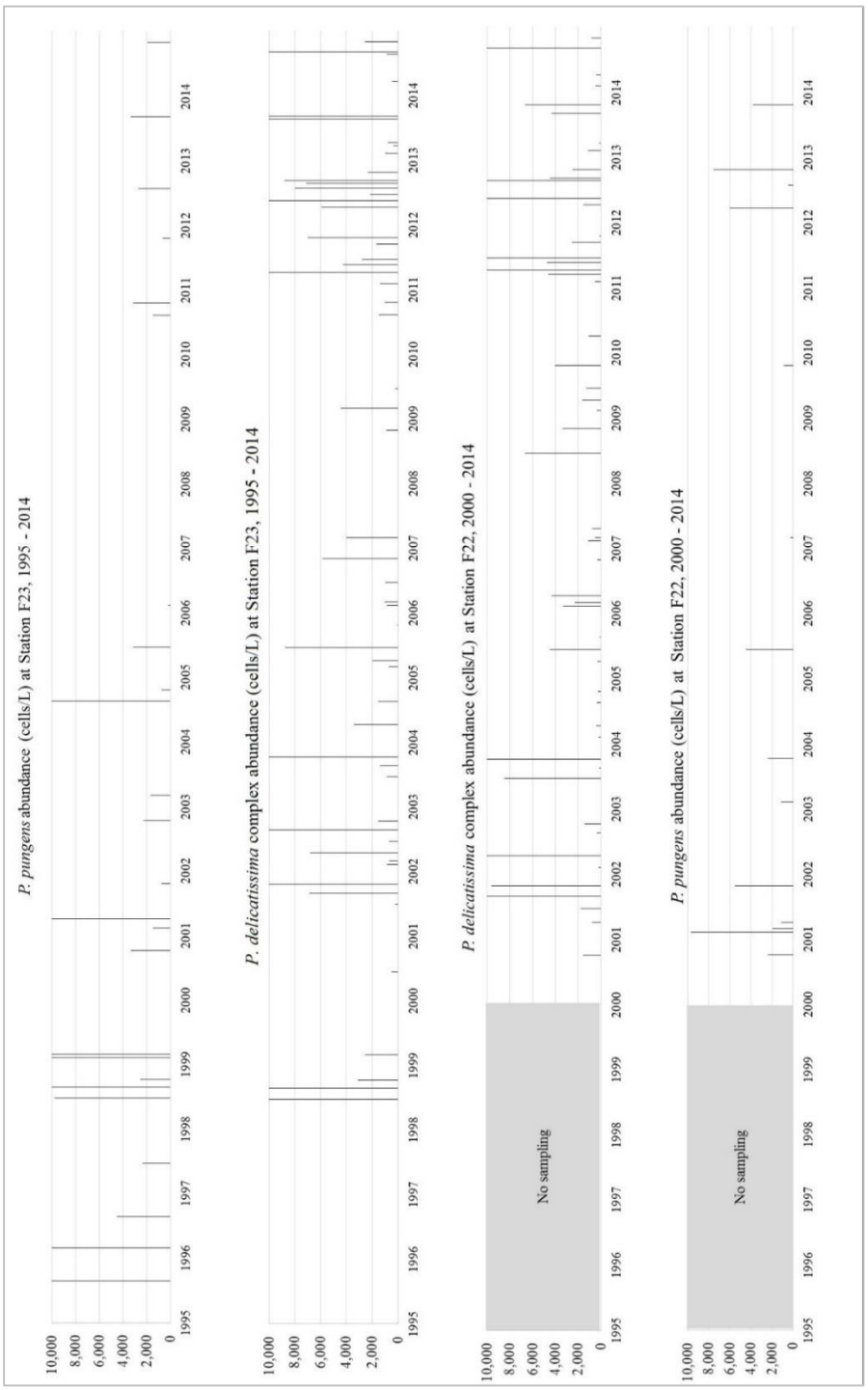

Figure 5-3. Abundance of *P. delicatissima* complex and *P. pungens* at Stations F22 and F23, 1995 - 2014. Data source: Massachusetts Water Resources Authority. Note that the y-axis is limited to 10,000 cells/L for reasons of scale.

Figure 5-4, below, shows the full y-axis scale of *P. delicatissima* abundance at Station F23 from 1992 to 2014. The massive event in 1998 dwarfs all other measurements, hence our use of a truncated scale in other figures.

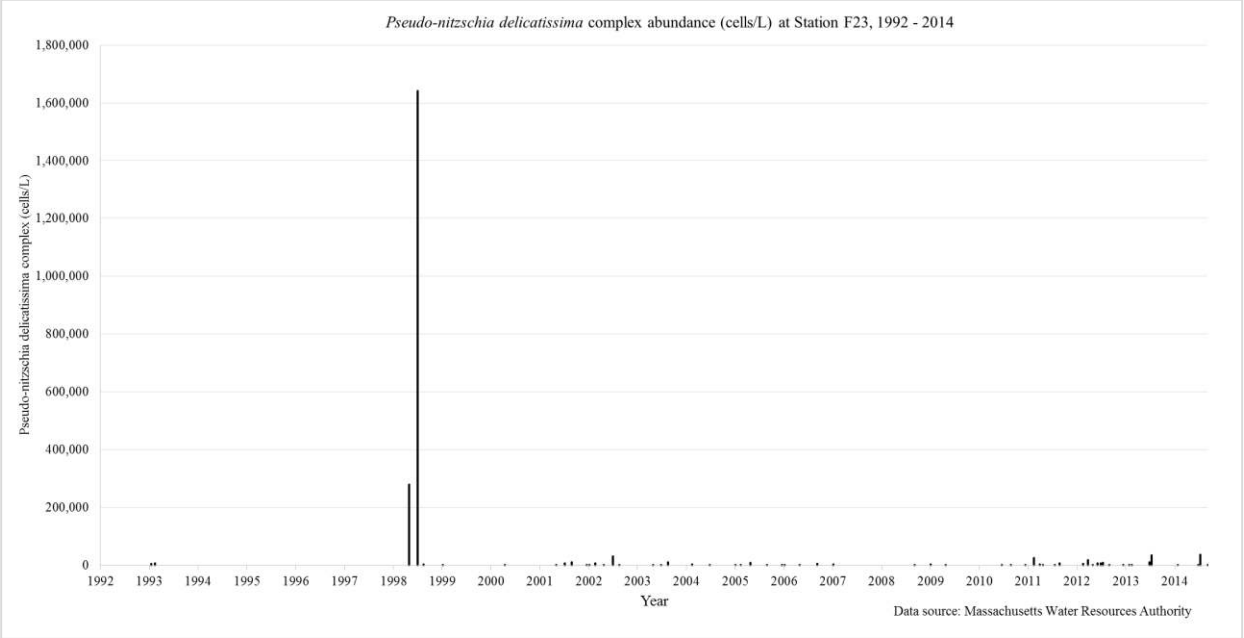

Figure 5-4. *P. delicatissima* complex at Station F23 only, 1992 - 2014.

An enlarged view of *P. delicatissima* complex abundance at Station F23 from 1995 to 2014 is shown below in Figure 5-5, with the y-axis limited to 10,000 cells/L. Based on this monitoring dataset there is no clear seasonal or annual signal in the presence of *P. delicatissima* complex. However, it does appear that Station F23 had a greater number of *P. delicatissima* complex observations with counts over 1,000 cells/L during the years 2011-2013.

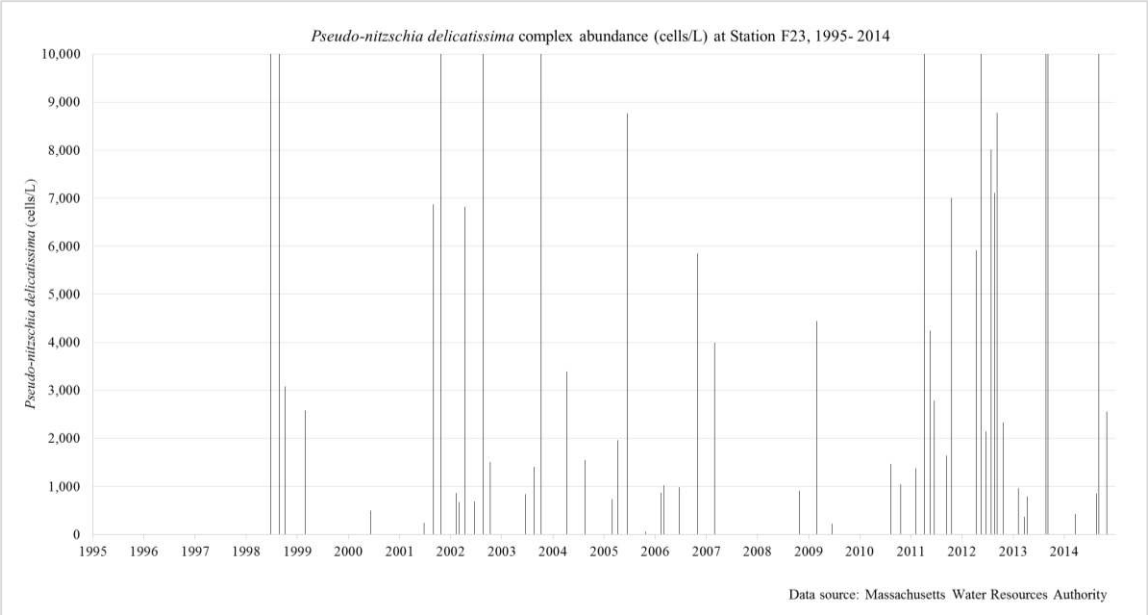

Figure 5-5. *P. delicatissima* complex abundance at Station F23, 1995 - 2014, limited scale.

Figure 5-6, below, shows *P. delicatissima* abundance at Station F22 between 2000 and 2014, with the y-axis truncated at 10,000 cells/L. As shown in the figure, *P. delicatissima* complex has frequently been found at Station F22 since sampling started in the year 2000, but there is no clear annual cycle.

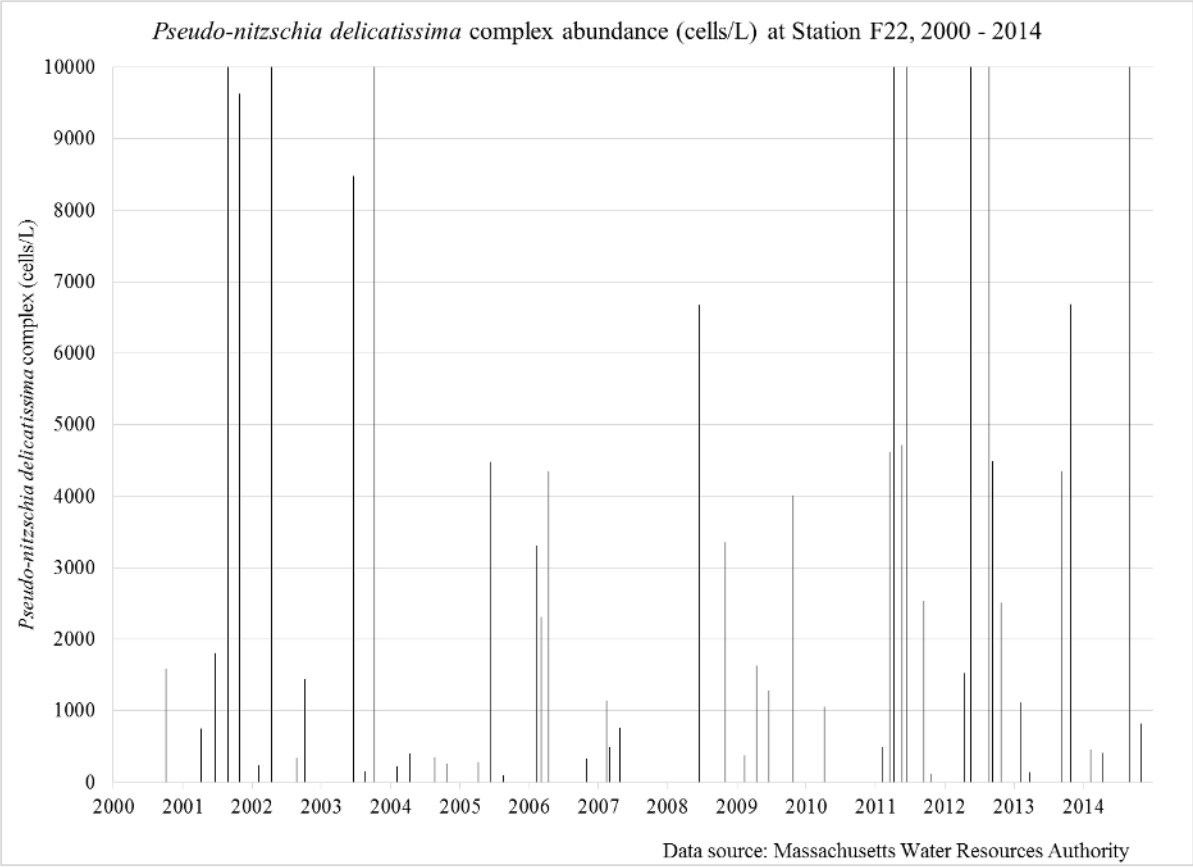

Figure 5-6. *P. delicatissima* complex abundance at Station F22, 2000-2014.

Given the abundance of non-zero-count samples for *P. delicatissima* complex as opposed to other categories of *Pseudo-nitzschia*, and the suggestion that morphology-based groups, or individual species, have different environmental niches, we focused solely on modeling the *P. delicatissima* complex.

Between 1995 and 2014 MWRA surface sampling for *Pseudo-nitzschia* at Station F22 and F23 ranged from 4 to 12 samples per station per year.[17] The observations made at Station F22 and F23 were part of a much larger monitoring program as part of the cleanup of Boston Harbor and the construction of the Deer Island Wastewater Treatment

Plant and the offshore outfall pipe disposal site (referred to as 'the outfall').[13; 18; 19] One of the goals of this cleanup effort was to reduce the amount of macronutrients being released into Boston Harbor, and there has been an approximately 80% decrease in ammonium concentrations as a result of this effort.[16] The drop in ammonium levels in outer Boston Harbor after the outfall went online is shown below in Figure 5-7. Station 142 and F23 are both located in outer Boston Harbor, including data from both stations provides finer temporal resolution of the ammonium concentration in outer Boston Harbor since nutrient sampling at Station F23 and Station 142 were usually offset by a few days.

Figure 5-7 shows that measured levels of ammonium at these stations dropped dramatically after the outfall went online in September 2000.

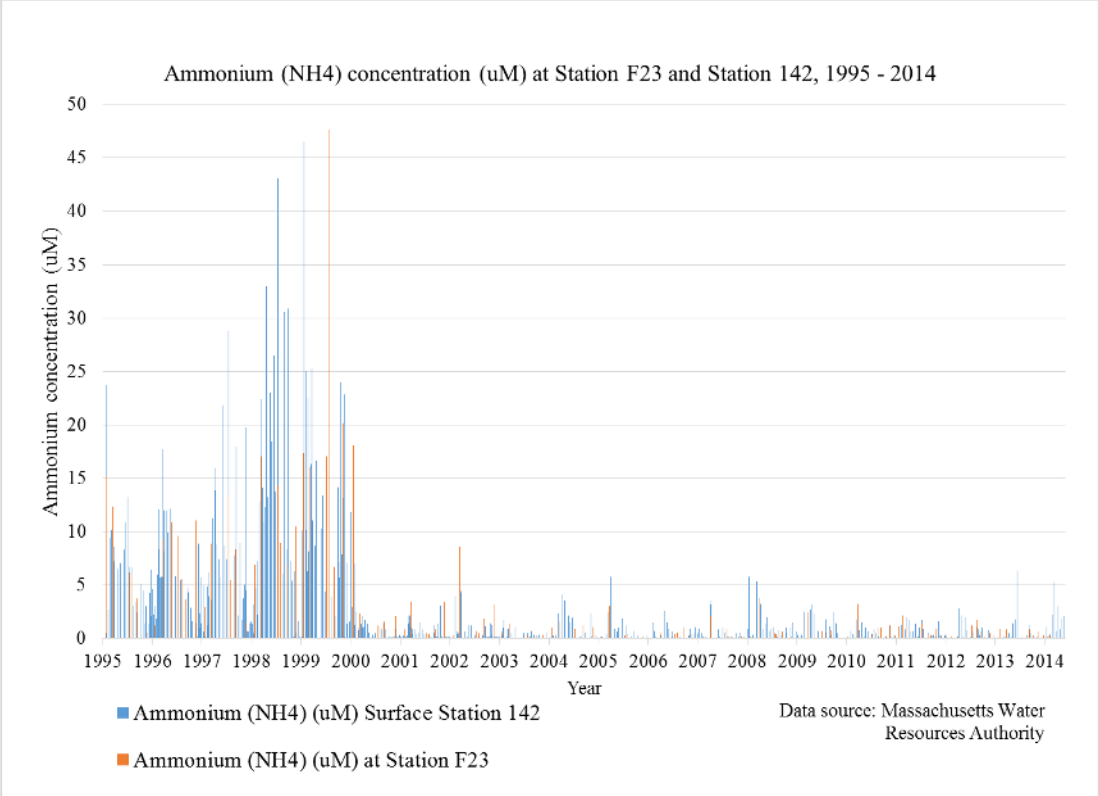

Figure 5-7. Ammonium concentrations at Station F23 and 142, 1995 - 2014.

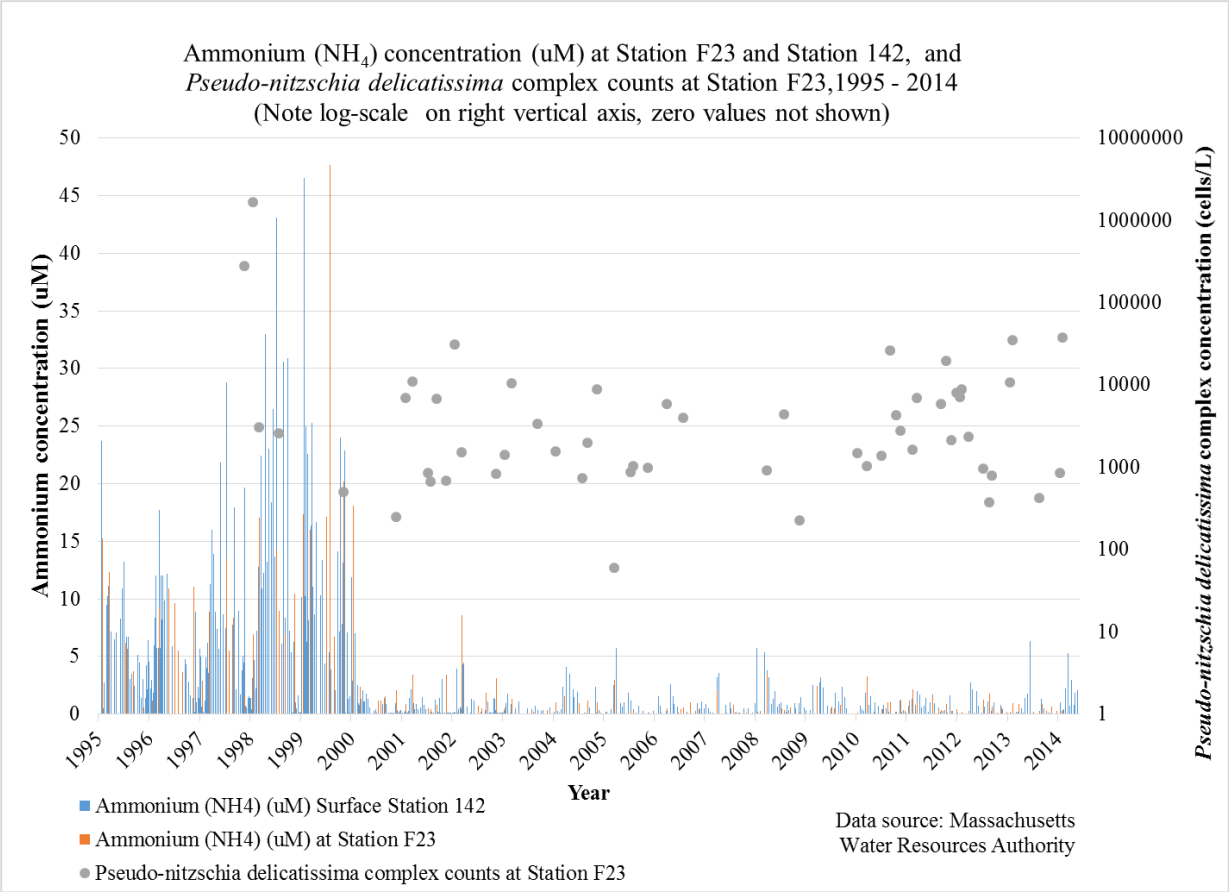

Figure 5-8. Ammonium concentrations at Stations F23 and 142, and *P. delicatissima* complex abundance at Station F23, 1995 -2014. Note log10 scale on right vertical axis.

Figure 5-8, above, shows the concentrations of ammonium at Stations F23 and 142, and *P. delicatissima* complex counts (using a log scale) at Station F23 from 1995-2014. The largest recorded *P. delicatissima* complex bloom event at Station F23 occurred in August 1998, with a concentration of over 1.6 million cells/L. There has not been a bloom of the same magnitude recorded at Station F23 since the outfall went online in September 2000, however there have been multiple blooms with concentrations over 10,000 cells/L.

In some parts of the world researchers have identified seasonal patterns in total *Pseudo-nitzschia* abundance.[3; 20] To visualize the possibility of broad seasonal trends in *P. delicatissima* bloom sizes in Massachusetts Bay we categorized abundance counts into six different size classes and graphed them by month of observation at Station F23 (see Figure 5-9) and Station F22 (see Figure 5-10).

Through the MWRA monitoring program *P. delicatissima* complex has been detected at Station F23 at the mouth of Boston Harbor in Massachusetts Bay at least one time in every month from February to October (see Figure 5-9). Blooms of 10,000 cells/L or more have been detected in April, May, August, September, and October, but not frequently. Of the 129 water samples examined from Stations 23, 74 samples had a count of 0 cells/L, and 55 samples had a count of greater than 0 cells/L. The most frequent count for all observations at Station F23 is 0 cells/L, shown as grey bars in Figure 5-9. Figure 5-9 also shows that sampling efforts are not evenly distributed across all months. For example, in the years 1995 to 2014 only 4 samples with *P. delicatissima* complex counts have been taken at Station F23 during the month of May and only 1 sample has been taken in December. February has the highest number of total samples (29), followed by August (20) and October (20). The striking finding is that despite extremely limited winter sampling (essentially only February), *P. delicatissima* complex has repeatedly been detected in every season at Station F23.

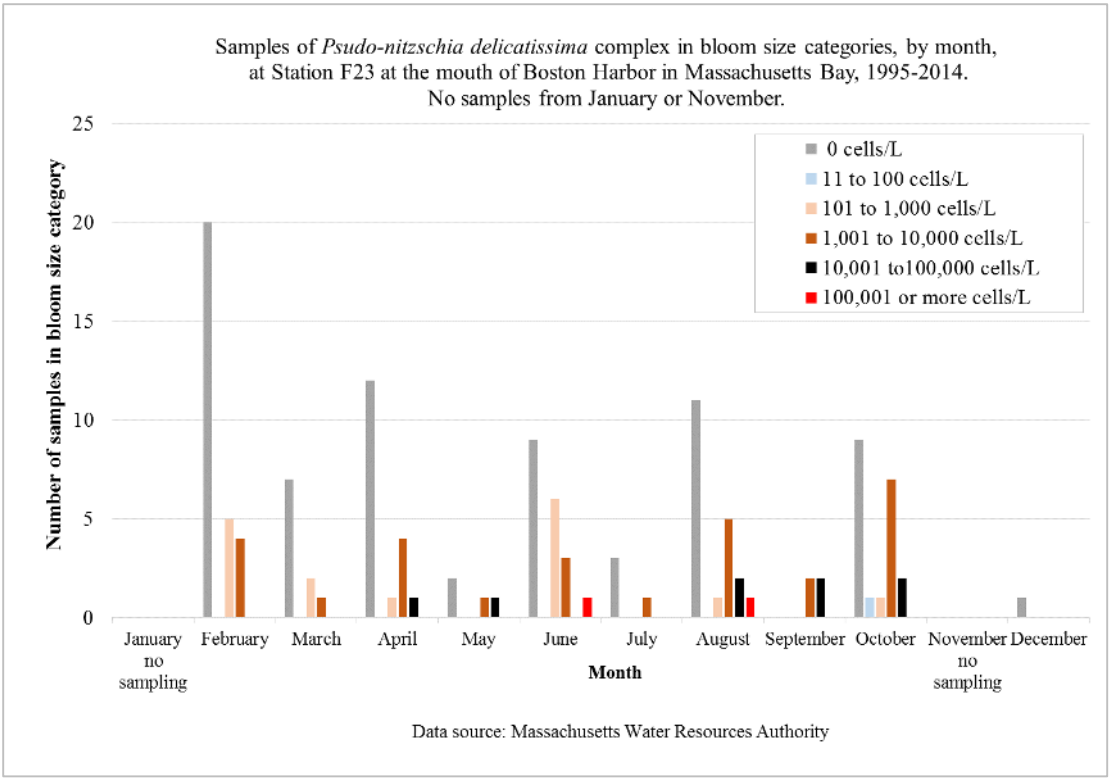

Figure 5-9. Samples for *P. delicatissima* complex at Station F23 from 1995 – 2014. A total of 129 sample were collected at Station F23, results are shown by bloom size category and month of sampling. For example: between 1995 and 2014 there were 29 samples collected in the month of February, 20 of those had counts of 0 cells/L (gray bars), 5 had counts ranging from 101 to 1,000 cells/L (pink bars), and 4 samples had counts between 1,001 to 10,000 cells/L (brown bars).

To provide further spatial and temporal nuance, the 100 samples from Station F22 are shown below in Figure 5-10, grouped by bloom size category and month of sampling. Note that no samples have been taken at Station F22 from November to January, and in this 15-year data set only 4 samples have been take in the month of July. No sample at Station F22 has recorded a bloom of 100,000 cells/L or more (see Figure 5-10). Of the Station F22 samples, 47 samples found 0 cells/L for *P. delicatissima* complex. However, at Station F22 sampling in the months of February through June and August through

October has recorded bloom sizes of 1,000 cells/L or greater for *Pseudo-nitzschia delicatissima* complex on multiple occasions. Despite temporal sampling coverage limitations at Station F22, *P. delicatissima* complex has been detected in every season. *P. delicatissima* complex sampling at Station F22 did not start until after the outfall went online, so we have limited insight into the effects of the outfall on *P. delicatissima* complex at that site. However, the dominant counter-clockwise circulation pattern in Massachusetts Bay puts Station F22 slightly 'upstream' of the outfall.[21] The repeated presence of *P. delicatissima* complex at Station F22 since the year 2000 suggests that these diatoms may also be present at other 'upstream' locations with limited influence from the outfall-driven nutrients and greater influence from regional oceanic processes.

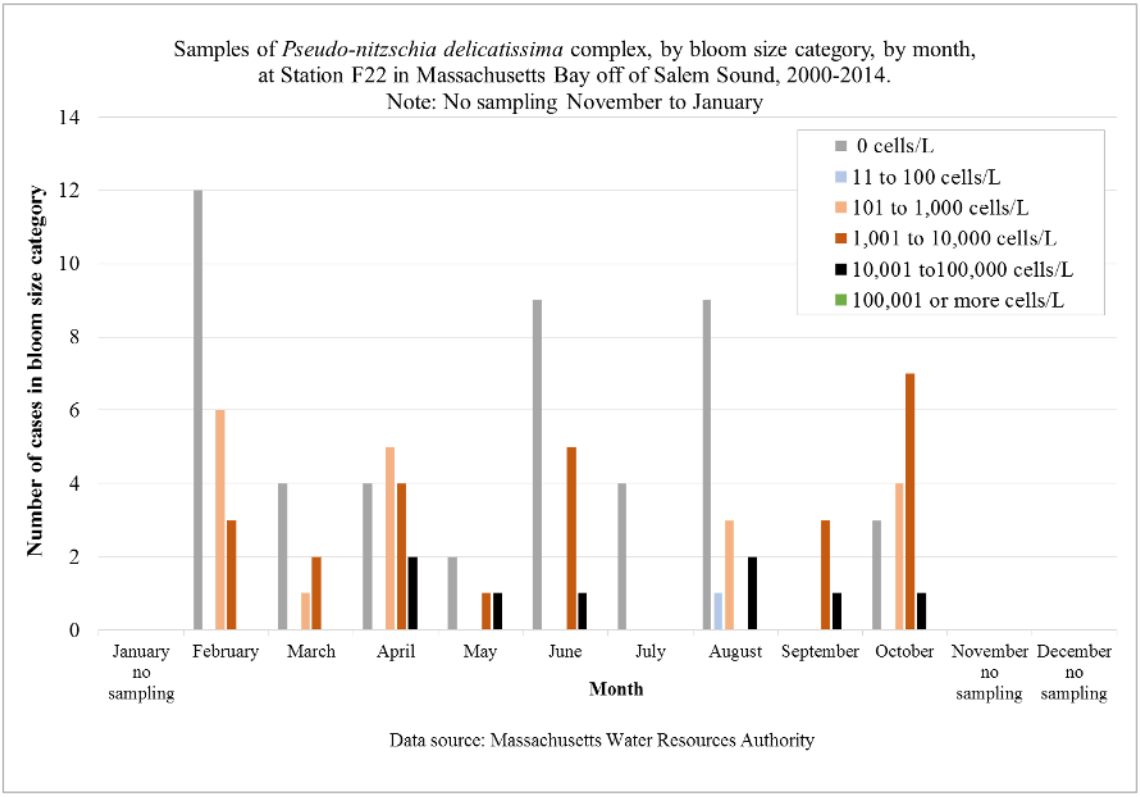

Figure 5-10. Samples for *P. delicatissima* complex at Station F22 from 1995 to 2014, by bloom size category and month of sample. Between 1995 and 2014 there were a total of 100 samples taken at Station F22.

Based on the 229 samples taken between 1995 and 2014 at Stations F22 and F23 it is clear that *P. delicatissima* complex can be present, in varying abundance, in Massachusetts Bay from February to October. There does not appear to be a clear seasonal pattern for *P. delicatissima* complex presence or abundance (bloom size category). Given its demonstrated presence in late fall (October) and late winter (February) it is possible that *P. delicatissima* complex may be present in Massachusetts Bay from November to January as well. The suggestion that *P. delicatissima* complex

may be present year-round in and of itself warrants further attention from public health authorities.

***Enterococcus* Abundance Data.** Water quality data containing the laboratory test results of *Enterococcus* abundance in recreational waters are published by the Commonwealth of Massachusetts, Department of Public Health (MA-DPH).[22] Beaches classified as 'Tier Two' are sampled weekly during the summer bathing season.[23] We selected three marine beaches in the Tier Two category along the northern end of Massachusetts Bay to use as our study sites for developing an *Enterococcus* model. These three beaches are all within 15 miles of Buoy A01 in northern Massachusetts Bay, and similarly proximal to Station F22 (one of the *P. delicatissima* complex sampling locations) as shown in Figure 5-11 below. These beaches are all at the upstream end of the counter-clockwise circulation pattern generally found in Massachusetts Bay.[21] It stands to reasons that if there is any ocean-driven influence on *Enterococcus* levels at coastal bathing beach areas such a signal would be most clear in these locations, close to where waters from the Gulf of Maine enter Massachusetts Bay.[21] Due to their proximity to Station F22, we believe that these ocean-facing marine beaches represented the best chance of detecting any potential relationship between beach water quality as measured by *Enterococcus*, and *P. delicatissima* complex abundance at Station F22. Despite their physical proximity, it should be noted that the sampling program at beaches is not coordinated with the MWRA sampling at offshore stations in Massachusetts Bay.

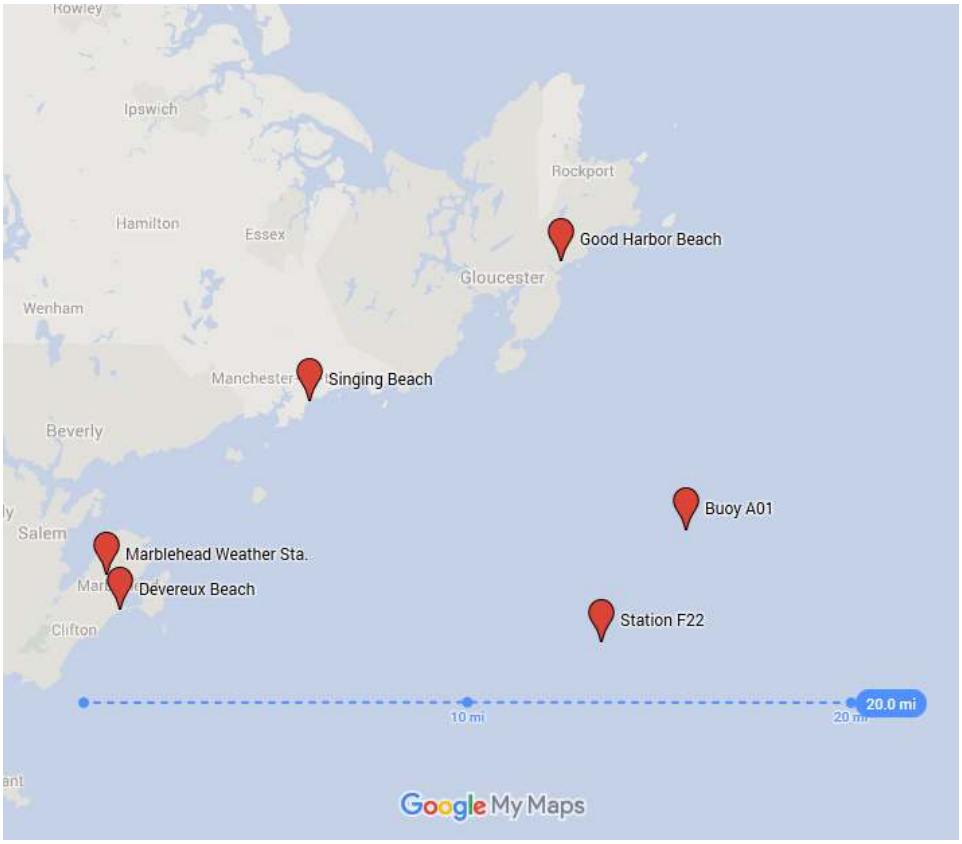

Figure 5-11. Map of *Enterococcus* sampling locations and other data collection points.

In addition to being proximal to the offshore sampling locations, all three of these beaches are within Essex County. According to the 2010 census data, the human population in the tract containing each beach is similar.[24] Devereux Beach is located in census tract 2031, population 4,557; Singing Beach is located within census tract 2181, population 5,136; and Good Harbor Beach is located within census tract 2213, population 4,532.[24] The results of summer recreational water quality sampling at these three beaches

from 2007 to 2014 are shown below in Figure 5-12, note the log scale on the vertical axis.

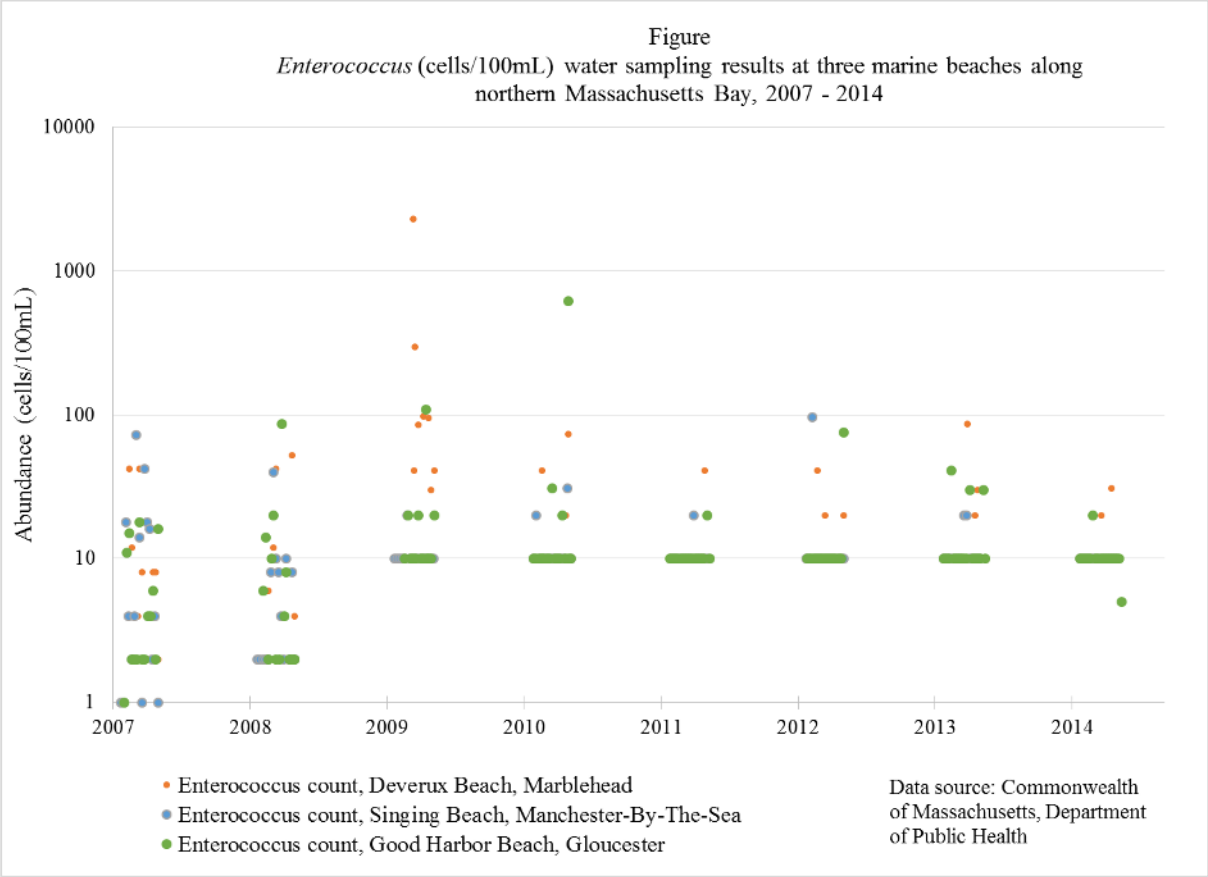

Figure 5-12. *Enterococcus* abundances at three study beaches, 2007 - 2014.

The threshold for an *Enterococcus* exceedance in recreational waters is 104 cells/100mL, and as shown in Figure 5-12 (above) it is clear that these three beaches had very few exceedances between 2007 and 2014. Although the threshold for exceedance is 104 cells/mL, we decided to create a binary response variable for *Enterococcus* with

'presence' equal to any count over 10 cells/100mL, and 'absence' as any count of 10 cells/100mL or lower.

Phase 2: Model Development and Selection

In this section we describe our model development and selection process for a suite of probabilistic models. First we describe the variables considered in model development, the set of candidate models, and model selection for the probabilistic presence/absence models of *P. delicatissima* complex in Massachusetts Bay. Then we go through the same process for the probabilistic presence/absence models of *Enterococcus* at three marine beaches in along the northern coast of Massachusetts Bay.

Probabilistic Model Development Using Logistic Regression. Our desired product is a probabilistic predictive model that could be used for public health purposes, not a mechanism-of-action explanation for *P. delicatissima* complex blooms or *Enterococcus* abundance. However, we expect that our results could be used to generate hypotheses for future research. Burnham and Anderson (2002) strong advise against "highly iterative and interactive" model development and caution that such activities should "be reserved for early exploratory phases of initial investigation."[10] To our knowledge this work is the first attempt to model *P. delicatissima* complex abundance in Massachusetts Bay. *Pseudo-nitzschia* modeling as a whole has a relatively short history, and previous studies have produced mixed results with regards to environmental

predictors and their potential influence on diatom abundance.[25] We are not aware of any predictive model for *Enterococcus* levels in northern Massachusetts Bay.

For both taxa our outcome variable of interest was a dichotomous outcome (presence / absence) and all of the candidate models were logistic regression models. A logistic regression equation is solved to provide probabilities of a dichotomous success outcome at the observed values of the predictors, and the total probability curve is 'S-shaped'.[26] We used logistic regression models instead of linear models because our data do not satisfy two important assumptions required for linear least squares modeling.[26] First, our outcome takes on only two possible values (0 or 1) and is therefore not normally distributed (a requirement for least squares modeling).[4; 26] Second, a linear model structure assumes that a specific change in the predictor variable is associated with the same change in the response probability no matter what the value of the predictor is (i.e., where that predictor value occurs along the straight line), this is unlikely to be true when the response variable is dichotomous.[26] Overall, the advantage of logistic regression with the logit function is that it models the log probability of a dichotomous outcome event as a linear combination of the predictor variables.[4]

There are three main elements to parsing a logistic regression function, the odds of a success, the logit function, and the odds ratio.[26] These elements are described briefly below.

- Odds of success: the ratio of the probability of a success, $\pi(x)$, to the probability of failure. Odds vary between 0 and ∞ as the probability varies between 0 and 1.
- Logit function: the natural log of the odds of success. The logit function varies between $-\infty$ and ∞ as the probability of success varies between 0 and 1. Logistic regression hypothesizes that the logit is linearly related to the predictors.

$$Logit(x) = \ln\ [\![\pi(x)/(1-\pi(x)\,)]\!]$$

- Odds ratio: The ratio of the odds for $x_j + 1$ to the odds for x_j. Where x represents the predictor variables. The logistic regression model suggests an additive/multiplicative relationship between a predictor and the odds, with a multiplicative change in the odds of success associated with a one unit increase in x_j, holding all else in the model fixed. The existence of a relationship between predictor variables and the dichotomous outcome is based on odds ratios through the use of the logit function.[26]

The statistical software package can calculate the probability of a success outcome for a given logistic regression equation and values of predictor variables. It is this predictive probability value that we are interested in, ultimately we will use these values to test the hindcast performance of the selected model. In addition to the probability we will have

to selection a prediction point above which the model will predict presence. The default prediction point is 0.5, but values can range from 0 to 1. In practice, alternate prediction points (based on odds ratios) can be used to optimize the predictive power of a model or be applied for selective risk management.[3] In this work any probability *above* the prediction point suggests the presence of the modeled organism, so any lowering of the prediction point below 0.5 would make a prediction of *presence* more likely. If used in a real-life public health decision-making space, where the prediction of organism presence results in costly response activities, modelers might benefit from direct consultation with response managers to find a realistically useful prediction point or suite of prediction points. For a discussion of these tradeoffs in the context of harmful *Pseudo-nitzschia* blooms see Lane et al. (2009) and Anderson et al. (2010).[3,4]

Model Selection. Our chosen approach was an information-theoretic approach rather than strict null hypothesis significance testing, this allowed for the inclusion of exploratory models and a limited degree of iterative model development. The information theoretic approach weighs competing models, the information criteria measure used provides a quantitative measure of relative support for each model.[12] For our information criteria measure we use Akaike's Information Criterion-corrected (AICc), which is the AIC with an additional corrective term for small sample sizes.[10] We used R Studio software program and the 'AICcmodavg' package for R for all AICc calculations.[27; 28] Correlation matrix graphics (Figures 5-13 and 5-14) were generated using the 'corrplot' version 0.73 package in R Studio.[29]

Model Cross-Validation. There are multiple cross-validation methods for testing models, involving some division between a 'training' dataset used to develop a model and other data reserved to form the 'test' dataset on which to test the model's predictive performance (also known as cross-validation). One common practice is using 75% of the total dataset for training and reserving 25% of the dataset for testing the model (this may also be referred to as model validation).[30] After identifying the most supported model according to AIC, we followed the cross-validation method used by Anderson et al (2010).[4] In brief, this involved leaving out one year of data, fitting the model, and then testing the model on the reserved year – this process was repeated for every year with available data.[4]

***Pseudo-nitzschia* models and results of model selection using AICc.** Here we describe the development of a suite of predictive models for the presence / absence of *P. delicatissima* complex in Massachusetts Bay and the selection of one model from that suite. Note that in this work models do not attempt to predict bloom size, although that may be of interest in future modeling work. The relationship between environmental influences, bloom size, and domoic acid product by any species within the *Pseudo-nitzschia* genus is the subject of ongoing research. Regional variation only adds to the potential complexity. *Pseudo-nitzschia* presence is a precursor to potential domoic acid (DA) production, but presence alone is not sufficient to indicate the presence of DA. Therefore, in this first modeling attempt for Massachusetts Bay, we focus strictly on

presence / absence of *P. delicatissima* complex. Additionally, we see this modeling exercise using monitoring data as an opportunity to generate hypotheses which may be further explored through purpose-designed experiments.

Pseudo-nitzschia model generation. We assembled a suite of response variables and predictor variables from a variety of sources, all of which are biologically reasonable according to the existing literature.[12] These variables are shown below in Table 5-3. The continuous variable of *P. delicatissima* complex count data (variable name 'pn.count') was transformed into a binary response variable (variable name 'binary0' with values of '0' representing absences, and values of '1' representing presence).

The total data set was split into two parts, the training dataset and the test dataset. The training dataset is the portion of the total dataset used to develop the model coefficients for predictor variables, it includes both the response and predictor variables. The test dataset is the portion where we use the model and observed predictor variable values to make a prediction about the probability of success, we can then compare the model's predictions (i.e., the hindcast) to the actual observed outcome and assess model performance. Our training dataset consisted of all possible cases, equal to 229 cases for *P. delicatissima* complex samples, 32 of these cases were dropped during model selection due to missing data for one or more predictor variables, leaving 197 cases. Of these 197 cases, 96 were observations with *P. delicatissima* complex present, the other 101 were absence cases. Each single-year cross validation test set included from 3 to 18 cases.

There was no test set for 1995 because of missing predictor variable data from that year. The variables considered during the development of candidate models are shown below in Table 5-3.

Table 5- 3. Variables use in *P. delicatissima* complex model development

Variable	**Variable description**	**Units**	**Source**
date	Date, in the form of year.month.day	na	
month	Month	na	
year	Year	na	
station	Station, either F22 (Massachusetts Bay, or F23 (entrance to Boston Harbor)	na	MWRA
pn.count	Count of *Pseudo-nitzschia delicatissima* complex only, all other *Pseudo-nitzschia* categories excluded.	cells/ L	MWRA
pn.ln	natural log of (pn.count)	na	calculated
binary0	Binary (0=false, 1= true) for presence of *P. delicatissima* complex	na	calculated
sal.station	salinity at the at the station when sampling	psu	MWRA
chl.station	Chlorophyll a at the station when sampling	µg/L	MWRA
chl.station.ln	Natural log of chl.station	na	calculated
nh4	NH4, Ammonium	µM	MWRA
no2	NO2, Nitrogen Dioxide	µM	MWRA
no3	NO3, Nitrate	µM	MWRA
no2no3	NO2 + NO3, sum of individual measures	µM	MWRA
DON	Dissolved Organic Nitrogen, DON = tdn – (no2 + no3 + nh4) [31]	µM	calculated
partP	particulate Phosphorous	µM	MWRA
po4	Phosphate	µM	MWRA
sio4	Silicate	µM	MWRA
sio4.ln	Natural log of sio4	na	Calculated
tdn	Total dissolved Nitrogen	µM	MWRA
tdp	Total dissolved Phosphorous including dissolved orthophosphate and dissolved organic phosphate	µM	MWRA
PON	particulate organic Nitrogen	µM	MWRA
watertemp.44013	Daily average water temp at NOAA Buoy 44013, Massachusetts Bay	°C	NOAA

Table 5- 3. Variables use in *P. delicatissima* complex model development			
Variable	**Variable description**	**Units**	**Source**
watertemp.5d.avg.44013	Average of water temp for 5 days at NOAA Buoy 44013, Massachusetts Bay	ºC	NOAA
watertempavg.a01	Daily average water temperature at Buoy A01, Massachusetts Bay. Latitude: 42° 31'19“ N Longitude: -70° 33'55“ W	ºC	NERACOOS
turbidavg.a01	Daily average turbidity 1meter depth at Buoy A01, Massachusetts Bay. Latitude: 42° 31'19“ N Longitude: -70° 33'55“ W	ntu	NERACOOS
zoo	Zooplankton, sum of individual copepodites, nauplii, trochophore, veliger, zoea, and unidentified organisms.	ind/ m^3	MWRA
zoo.ln	natural log of (zoo)	na	calculated
nn.p	ratio of (NO2+NO3) to phosphate (PO4), (no2no3/(po4))	na	calculated
si.no3	Ratio of silicate to nitrate (sio4/no3)	na	calculated
si.po4	Ratio of silicate to phosphate (sio4/no3)	na	calculated
prcp.bos	Precipitation at Boston Logan Airport weather station, code GHCND:USW00014739	tenths of mm	NOAA
prcp.day.before	Precipitation at Boston Logan Airport weather station one day before sampling, code GHCND:USW00014739	tenths of mm	calculated
prcp.5day.total	Total precipitation at Boston Logan Airport weather station for 4 days before and on day of sampling, code GHCND:USW00014739	tenths of mm	calculated
river.dis	Merrimack River flow rate at USGS station 01100000 "Merrimack River BL Concord River at Lowell, MA", cubic feet per second	Ft^3/ sec	USGS
river.ln	natural log of (river.dis)	na	calculated
river.2wkavg	Average flow rate for Merrimack River for 2 weeks preceding sampling (including on day of sampling)	Ft^3/ sec	USGS
river.1wkavg	Average flow rate for Merrimack River for 1 week preceding sampling (including on day of)	Ft^3/ sec	USGS
river.30davg	Average flow rate for Merrimack River for 30 days preceding sampling (including on day of)	Ft^3/ sec	USGS
river.30day.ln	natural log of (river.30davg)	na	calculated

Table 5- 3. Variables use in *P. delicatissima* complex model development			
Variable	**Variable description**	**Units**	**Source**
fluo.avg	fluorescence (averaged for the day if more than one observation at station, negative values removed and treated as missing data)	ug/L	MWRA, calculated

A graphical correlation matrix is shown below in Figure 5-13. The size and intensity of blue circles indicate positive correlations, red circles indicate negative correlations. Blank or white-fill cells indicate a correlation coefficient near to zero. The correlation matrix visualization clearly displays some notable strong positive correlations for example, month and watertemp.44013, PON and partP and chl.station. Slightly less strong positive correlations are visible between the variables po4, no3, no2no3, sio4, and tdn. Strong negative correlations are apparent for the variables watertempavg.a01 with both no3 and no2no3, along with zoo and po4. A slightly less strong negative correlation is apparent between the variables si.po4 and sal.station, as well as month and no3.

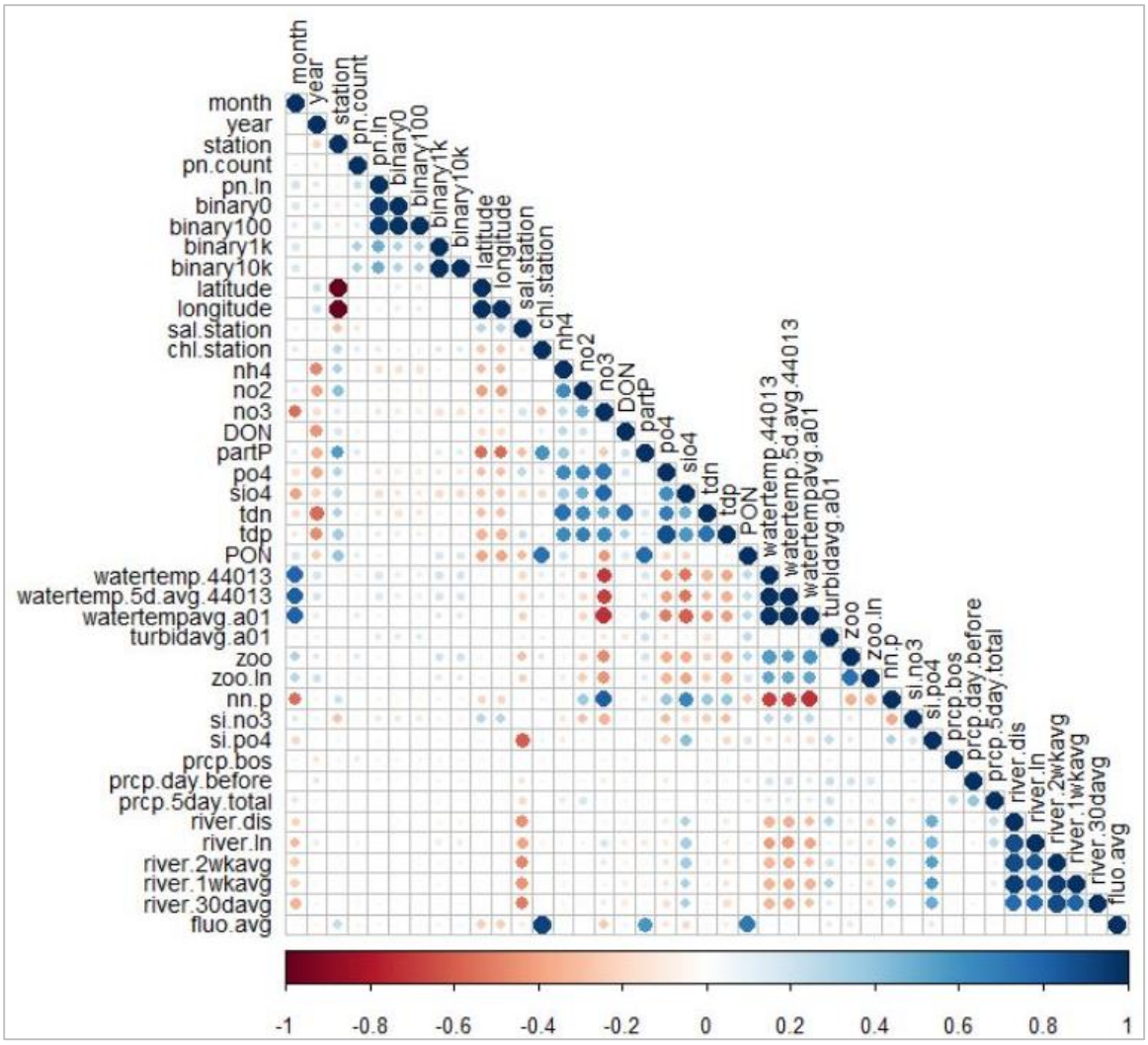

Figure 5-13. Graphical correlation matrix for variables considered in *P. delicatissima* complex model development.

We developed a suite of *a priori* models to predict presence / absence of *P. delicatissima* complex in Massachusetts Bay. These models were based on knowledge of major physical and seasonal drivers influencing the Massachusetts Bay system, basic diatom biology, and published studies where other researchers have attempted to model, or

correlate environmental variables with, total *Pseudo-nitzschia* or *Pseudo-nitzschia delicatissima* in other locations. Our set of ten candidate models is listed below in Table 5-4, the R Studio code for these models is included in Appendix A.

Table 5-4. Candidate models for *P. delicatissima* complex presence/absence

Model number	**Variables predicting *Pseudo-nitzschia delicatissima* complex**	**Number of parameters**	**Rationale for model**
1	temperature, sio4, nh4, no3, po4, zoo.ln, prcp.day.before, salinity	8	Lelong et al. (2012) describe multiple variables influencing *Pseudo-nitzschia* species.[25]
2	temperature, hours of light (no proxy), rainfall (proxy = prcp.5day.total), phosphate, salinity	4	Downes-Tettmar et al. (2013) significant correlations for *P. delicatissima.*[9]
3	temperature, po4, (no3+no2), month	4	Anderson et al. (2010) variables for **10** cells/ml bloom threshold, discarding latitude and longitude.[4]
4	temperature, po4, salinity, sio4, freshwater discharge (proxy = river.1wkavg), month	6	Anderson et al. (2010) variables for **100** cells/ml bloom threshold,[4]
5	temp, ln(silicic acid) (proxy = sio4.ln), ln(chl *a*)	3	Lane et al. (2009), **spring** season model, total *Pseudo-nitzschia*.[3]
6	ln(silicic acid) (proxy = sio4.ln), chl.station.ln, river.30day.ln, no3	4	Lane et al. (2009), **fall** season model, total *Pseudo-nitzschia*.[3]
7	no2no3, ratio of nitrate to phosphate (nn.p)	2	Kaczmarska et al. (2007), **summer/fall** variable significantly positively correlated with *P. delicatissima* and *P. pseudodelicatissima.* [2]

Table 5-4. Candidate models for *P. delicatissima* complex presence/absence			
Model number	**Variables predicting *Pseudo-nitzschia delicatissima* complex**	**Number of parameters**	**Rationale for model**
8	no2, sio4, prcp.day.before	3	Exploratory, based on results in correlation matrix and evidence that *P. delicatissima* complex can be present in all seasons, might be a short-term response.
9	tdn, si.no3, prcp.day.before	3	Exploratory, based on results in correlation matrix.
10	temp, silicate, chl.station, tdn, si.no3, prcp.day.before, latitude, longitude	6	Combination of earlier exploratory models.
11	temp, silicate, chl.station, no2no3, si.no3, prcp.day.before, DON, nh4	8	Loureiro et al. (2009) show that in laboratory cultures *P. delicatissima* preferentially acquires NH4 but in limiting conditions may use urea as alternative N source, thus levels of dissolved organic nitrogen (DON) may influence abundance.
12	zoo.ln, sio4, po4, nh4, prcp.day.before	5	The most common grazers of diatoms are large organisms such as copepods (Sarthou et al 2005, quoting Smetacek 1999)[32], so zoo.ln is a logical potential predictor variable.
13	temp, sio4, chl.station, nh4, po4, prcp.day.before	6	Exploratory model based on diatom bloom principles, precipitation might be depositing dust from the air.
14	temperature, po4, (no3+no2), month, latitude, longitude	6	Anderson et al. (2010) Chesapeake Bay, best-fit logistic GLM, variables for **10** cells/ml bloom threshold (total *Pseudo-nitzschia*), including latitude, longitude.

Table 5-4. Candidate models for *P. delicatissima* complex presence/absence			
Model number	**Variables predicting *Pseudo-nitzschia delicatissima* complex**	**Number of parameters**	**Rationale for model**
15	tdn, si.no3, prcp.day.before, latitude, longitude	3	Exploratory model based on results in correlation matrix, exploratory.
16	Latitude, longitude	2	Only latitude and longitude

P. delicatissima complex model selection. The AICc results for this candidate set are shown below (see Table 5-5). As a reminder, the AICc is the same as the AIC described above with the addition of corrective term for small sample sizes.[10]

Table 5-5. AICc scoring for *P. delicatissima* complex candidate model set

Model Number	K	AICc	ΔAICc	AICcWt	Cum.Wt
10	8	263.17	0	0.98	0.98
11	10	270.97	7.8	0.02	1
5	4	288.08	24.91	0	1
13	7	288.21	25.03	0	1
4	7	292.36	29.19	0	1
1	9	293.6	30.43	0	1
3	6	293.81	30.64	0	1
9	4	295.27	32.1	0	1
15	5	295.29	32.12	0	1
14	7	295.31	32.14	0	1
2	5	297.91	34.74	0	1
6	5	312.84	49.67	0	1
12	6	312.91	49.74	0	1
8	4	316.33	53.16	0	1
16	2	318.35	55.17	0	1
7	4	319.43	56.26	0	1

The AICc output tables have the following component columns:

- K, the number of parameters estimated,
- AICc, the estimated distance from the proposed model to the true model,
- ΔAICc, the difference in AICc score between that model and the most supported model in the set (listed first among all the models in the set),
- AICcWt indicates the total model weight.
- Cum.Wt, the cumulative weight of the models from the most supported on down,

As shown in the results above, Model 10 is the most supported model in this candidate set. The other models have ΔAICc values greater than 7, so there is no support for those models when compared to Model 10. There is no need to perform model averaging because the weight (shown in column "AICcWt") of the 'best' model is greater than 0.9.[12] Based on the AICc score for our candidate set, the most supported predictive model for the presence/absence of *P. delicatissima* complex in Massachusetts Bay is as follows:

$$LOGIT(p) = -234.823 +$$

$$0.0505(watertemp.44013) - 0.0362(tdn) - 0.0251(si.no3) - 0.0089(sio4) -$$

$$0.0032(prcp.day.before) + 0.0367(chl.station) + 5.545(latitude)$$

Note that there is no coefficient provided by longitude. The performance of this model on both the training and test datasets will be discussed in the later section on model performance as part of Phase 3 (evaluate outputs).

Enterococcus candidate models and results of model selection using AICc.

Similar to the work for the *P. delicatissima* complex models described above, the

Enterococcus candidate models were developed for a dichotomous response, presence or absence of *Enterococcus* counts above 10 cells/100mL. The current threshold for a recreational water quality exceedance is 104 cells/100mL[33], and our presence/absence point of 10 cells/100mL is an order of magnitude lower than that. Of the 264 cases in the training dataset 156 had a value of 10 cells/100mL. The training dataset was comprised of recreational water quality samples taken weekly during the summer bathing seasons in 2007 to 2014 at three ocean-facing marine beaches: Devereux Beach in the town of Marblehead, Singing Beach Point 1 in the town of Manchester-By-The-Sea, and Good Harbor Beach in the town of Gloucester. Water quality results for all these and other beaches are available online from the Commonwealth of Massachusetts.[22]

Enterococcus model generation. We selected these beaches because of their proximity to sampling station F22 (where part of the *P. delicatissima* complex data was collected) and to Buoy A01 which provided data on water temperature, turbidity, and chlorophyll *a* levels.[8] We limited our temporal range to 2007 to 2014 to allow for the consideration of offshore turbidity data from Buoy A01in the model development process. The variables considered during model development are shown below in Table 5-6.

Table 5-6. Variables considered in *Enterococcus* presence/absence model development

Variable name	**Variable description**	**Units**	**Source**
Date	Date, in the form of year.month.day	na	
Month	Month	na	
Year	Year	na	
Latitude	Latitude of sampling location, estimated to the second decimal place.	na	GoogleMap

Table 5-6. Variables considered in *Enterococcus* presence/absence model development			
Variable name	**Variable description**	**Units**	**Source**
Longitude	Longitude of sampling location, estimated to the second decimal place.	na	GoogleMap
entero.count	*Enterococcus* counts from water quality testing by local public health officials, reported to state-level Department of Public Health.	cells / 100mL	MA-DPH
entero.ln	Natural log of (entero.count)	na	Calculated
entero.exceed	Binary variable, if entero.count $\geq$ 104 coded as '1', if entero.count <104 coded as '0'.	na	Calculated
entero.over10	Binary variable, if entero.count > 10 coded as '1', if entero.count $\leq$10 coded as '0'.	na	Calculated
human.pop.tract	Human population in the census tract containing the beach sampling point	count	U.S. Census
dog.pop	Dog population in the town containing the beach sampling point	count	Collected by author
tmax.bos	Maximum daily temperature recorded at Boston Logan Airport weather station, code GHCND:USW00014739	tenths of ^{o}C	NOAA
prcp.bos.3day	Cumulative precipitation for sampling day plus 2 previous days (3 days total) recorded at Boston Logan Airport weather station, code GHCND:USW00014739	tenths of mm	Calculated
prcp.marblehead	Total precipitation on the day of sampling recorded at Marblehead weather station, code GHCND:USC00194502	tenths of mm	NOAA
tmax.marblehead	Maximum daily temperature recorded at Marblehead weather station, code GHCND:USC00194502	^{o}C	NOAA
prcp.mblhd.day. before	Total precipitation on the day before the sampling day recorded at Marblehead weather station, code GHCND:USC00194502	tenths of mm	NOAA
prcp.mblhd.2day	Cumulative precipitation for sampling day plus 1 previous days (2 days total) recorded at Marblehead weather station, code GHCND:USC00194502	tenths of mm	Calculated
prcp.mblhd.5day	Cumulative precipitation for sampling day plus 4 previous days (5 days total) recorded at Marblehead weather station, code GHCND:USC00194502	tenths of mm	Calculated

Table 5-6. Variables considered in *Enterococcus* presence/absence model development			
Variable name	**Variable description**	**Units**	**Source**
chl.a01	Daily average chlorophyll *a* levels measured at Buoy A01, Massachusetts Bay. Latitude: 42° 31'19“ N Longitude: -70° 33'55“ W	ug/L	NERACOOS
watertempavg.a01	Daily average water temperature at Buoy A01, Massachusetts Bay. Latitude: 42° 31'19“ N Longitude: -70° 33'55“ W	°C	NERACOOS
turbidavg.a01	Daily average turbidity 1meter depth at Buoy A01, Massachusetts Bay. Latitude: 42° 31'19“ N Longitude: -70° 33'55“ W	ntu	NERACOOS
prcp.bos	Precipitation at Boston Logan Airport weather station, code GHCND:USW00014739	tenths of mm	NOAA
prcp.day.before	Precipitation at Boston Logan Airport weather station one day before sampling, code GHCND:USW00014739	tenths of mm	calculated
river.dis	Merrimack River flow rate at USGS station 01100000 "Merrimack River BL Concord River at Lowell, MA", cubic feet per second	Ft^3/ sec	USGS
river.ln	natural log of (river.dis)	na	calculated
river.2wkavg	Average flow rate for Merrimack River for 2 weeks preceding sampling (including on day of sampling)	Ft^3/ sec	USGS
river.1wkavg	Average flow rate for Merrimack River for 1 week preceding sampling (including on day of sampling)	Ft^3/ sec	USGS
river.30davg	Average flow rate for Merrimack River for 30 days preceding sampling (including on day of sampling)	Ft^3/ sec	USGS
river.30day.ln	natural log of (river.30davg)	na	calculated

The graphical correlation matrix for the variables considered in the *Enterococcus* model development is shown below in Figure 5-14. The response variable of interest is ‘entero.over10’ and there are some faint positive correlations between entero.over10 and the variables representing precipiation and the variable chl.a01, as well as a faint negative correlation with the variable tmax.marblehead. Other physical environmental

correlations are visible, such as the strong postive correlation between month and water temperature, and the faint positive correlation between ‘river.dis’ and ’prcp.mblhd.5day.’

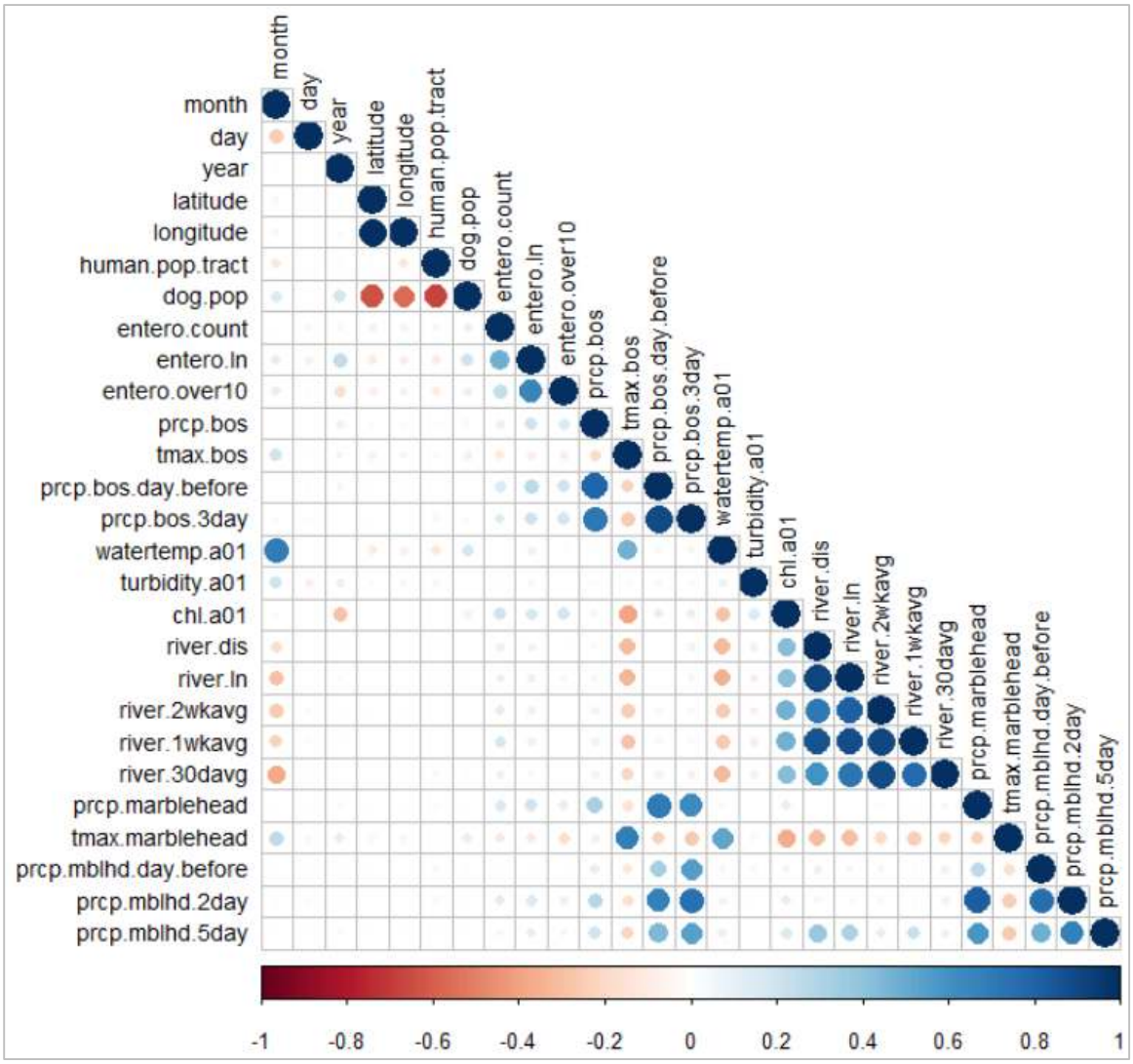

Figure 5-14. Graphical correlation matrix of variables considered during *Enterococcus* model development.

Based on the literature review results identifying potential influences on levels of *Enterococcus* in the marine environment, the available data, and insights from early

exploratory model development, we developed a set of 12 models for *Enterococcus.* These models are described in Table 5-7 (below), the R Studio code for these models is included in Appendix A.

Table 5-7. Candidate model set for *Enterococcus* presence/absence model.

Model Number	Predictor variables	Number of variables	Rationale
1	prcp.mblhd.day.before	1	Marblehead weather station is closer to the sampling sites, and might prove a better fit than the Boston-based weather data. Or they might be equal. Precipitation alone is often used as a reason to close beaches.[23]
2	human.pop.tract	1	Mammalian fecal waste can contribute *Enterococcus* to the system, humans living in the census tract might be more likely to contribute *Enterococcus* via direct shedding.[23]
3	dog.pop	1	Mammalian fecal waste (including dog waste) can contribute *Enterococcus* to the local system.[23]
4	turbidity.a01	1	*Enterococcus* might persist longer at higher levels of turbidity (which may provide growth substrate)[34]
5	watertemp.a01	1	Higher water temperatures may encourage *Enterococcus* persistence.[34; 35]
6	river.2wkavg	1	Correlation matrix showed slight positive association, and rivers may carry Enterococcus from land-based sources.[34]
7	prcp.mblhd.2day	1	Precipitation is currently used as a reason to close beaches before testing is finished.[23]

Table 5-7. Candidate model set for *Enterococcus* presence/absence model.			
Model Number	**Predictor variables**	**Number of variables**	**Rationale**
8	chl.a01 + tmax.marblehead + prcp.bos.day.before +prcp.mblhd.2day + human.pop.tract + dog.pop	6	A mixture of logical predictor variables.
9	latitude + longitude	2	Location may be a predictor.
10	chl.a01 + tmax.marblehead + prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude +river.1wkavg	9	A mixture of logical predictor variables, including latitude and longitude for location information.
11	chl.a01 + tmax.marblehead + prcp.bos.day.before + prcp.mblhd.2day + human.pop.tract + dog.pop	6	The correlation matrix shows a slightly stronger relationship with prcp.bos.day before than with prcp.mblhd.day.before, but both are positively correlated. Precipitation is highly associated with surface runoff that moves fecal matter containing *Enterococcus* from land into water. Daily maximum air temperature is used as a proxy for local sunshine (negatively correlated with *Enterococcus* levels). Nutrients and water column stratification that favor higher levels of chlorophyll might indicate favorable conditions for *Enterococcus* persistence.
12	latitude + year + watertemp.a01 +river.1wkavg	4	Water temperature, river output, and proximity to riverine output may influence *Enterococcus* levels.

Enterococcus model selection. The AICc scoring for the *Enterococcus* presence/ absence predictive model candidate set is shown below (see Table 5-8).

Table 5-8. AICc scoring results for *Enterococcus* candidate model set

Model Number	K	AICc	ΔAICc	AICcWt	Cum.Wt
10	10	295.14	0	0.98	0.98
8	7	303.45	8.31	0.02	0.99
11	6	305.69	10.55	0.01	1
12	5	317.64	22.5	0	1
9	3	319.93	24.79	0	1
2	2	320.27	25.13	0	1
7	2	320.97	25.83	0	1
3	2	321.05	25.91	0	1
4	2	321.3	26.16	0	1
1	2	323.4	28.26	0	1
5	2	323.82	28.68	0	1
6	2	324.36	29.22	0	1

Based on the AICc scores reported above, Model 10 is the most supported model. Relative to Model 10, none of the other models in the candidate set are supported, their ΔAICc values are all greater than 7. Additionally, the AICcWt for Model 10 is greater than 0.9, so multi-model averaging does not need to be considered for model parameters in this candidate set. Based on the AICc scores, the most supported predictive model for *Enterococcus* presence/ absence at the three ocean-facing north coastal watershed beaches used for our study is:

$$\begin{aligned} LOGIT(p) = \; & 594.2 + 0.4087(chl.a01) - 0.0116(tmax.marblehead) \\ & + \; 0.0040(prcp.bos.day.before) - 0.0022(prcp.mblhd.2day) \\ & - 0.1733(year) + \; 0.1766(watertemp.a01) - 68.14(latitude) \\ & + \; 38.11(longitude) - 0.00003(river.1wkavg) \end{aligned}$$

The cross-validation performance of this model will be discussed in the section on model performance as part of Phase 3 (evaluate outputs). The R Studio code used to generate, selection, and perform cross-validation tests whose results are described in this section is included in Appendix A.

Phase 3: Evaluate Outputs

As shown in Figure 5-1, Phase 3 of this research process involves evaluating the outputs developed during Phase 2. Here we discuss the performance of the models developed to predict the presence/absence of *P. delicatissima* complex and *Enterococcus.* We will also address the following three questions related to model performance:

- Is there a useful level of predictive value (based on hindcast performance) that could be used to protect human health?
- Is further field sampling or experimental data suggested?
- Does this further the development of theory?

The discussion on model performance will be followed by a summary and conclusion.

Model Performance. We evaluated model based on four performance aspects, sensitivity, specificity (also referred to as selectivity), false positive rate, and false negative rate. Model sensitivity is the chance of detecting a true positive (TP), specificity is the chance of detecting a true negative (TN).[36] The false positive ratio (referred to as

'Type I' error) is the chance of getting a false positive (FP) and the false negative ratio (also known as 'Type II' error) is the chance of getting a false negative (FN).[36] The equations for calculating these performance measures are summarized below:

- Sensitivity (chance of correctly predicting true positive) = TP/(TP+FN)
- Specificity (chance of correctly predicting true negative) = TN/(TN+FP)
- False Positive Rate (Type I error) = FP/(FP+TN)
- False Negative Rate (Type II error) = FN/(FN+TP)

In an ideal world both sensitivity and specificity would be high, and the false positive and false negative rates would be low. However, there are likely to be tradeoffs between each aspect of performance depending on the prediction point used. The default prediction point for a dichotomous response model comparison is usually 0.5.[3] However, it is possible to use alternate prediction points if there are specific aspects of model performance that are considered more important to the user. For example, if false positives are extremely costly the user might want to minimize their likelihood. Performance at default and alternative prediction points for the predictive models of presence/absence for *P. delicatissima* complex and *Enterococcus* are discussed below.

Pseudo-nitzschia delicatissima complex presence/absence prediction model. Using an information-theoretic approach we developed a model for the presence/absence of *P. delicatissima* complex diatoms in Massachusetts Bay. The motivating question for this model development was presented in chapter 3:

- Is it possible to hindcast levels of *Pseudo-nitzschia* populations measured in Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Pseudo-nitzschia* growth?

We developed a model using a dataset containing 197 cases, re-fit the model with one year left out, and then tested the model on that year. This created an ensemble of 19 different cross-validation experiments. Mean performance metrics for the areas of sensitivity, specificity, false positive rate, and false negative rate for the cross-validation experiments are summarized in Table 5-9. A graph showing the mean performance metric scores at multiple prediction points is shown in Figure 5-15. The R Studio code used to generate, selection, and perform cross-validation tests is included in Appendix A.

Table 5-9. Ensemble of Cross-Validation Performance Metrics: *P. delicatissima* presence/absence prediction

Mean model performance values: *P. delicatissima* complex presence/absence prediction		
	Default prediction point	**Alternate prediction point**
Prediction point value	**0.5**	**0.3**
Sensitivity	0.54	0.91
Specificity	0.52	0.20
False Positive Rate	0.48	0.80
False Negative Rate	0.46	0.09

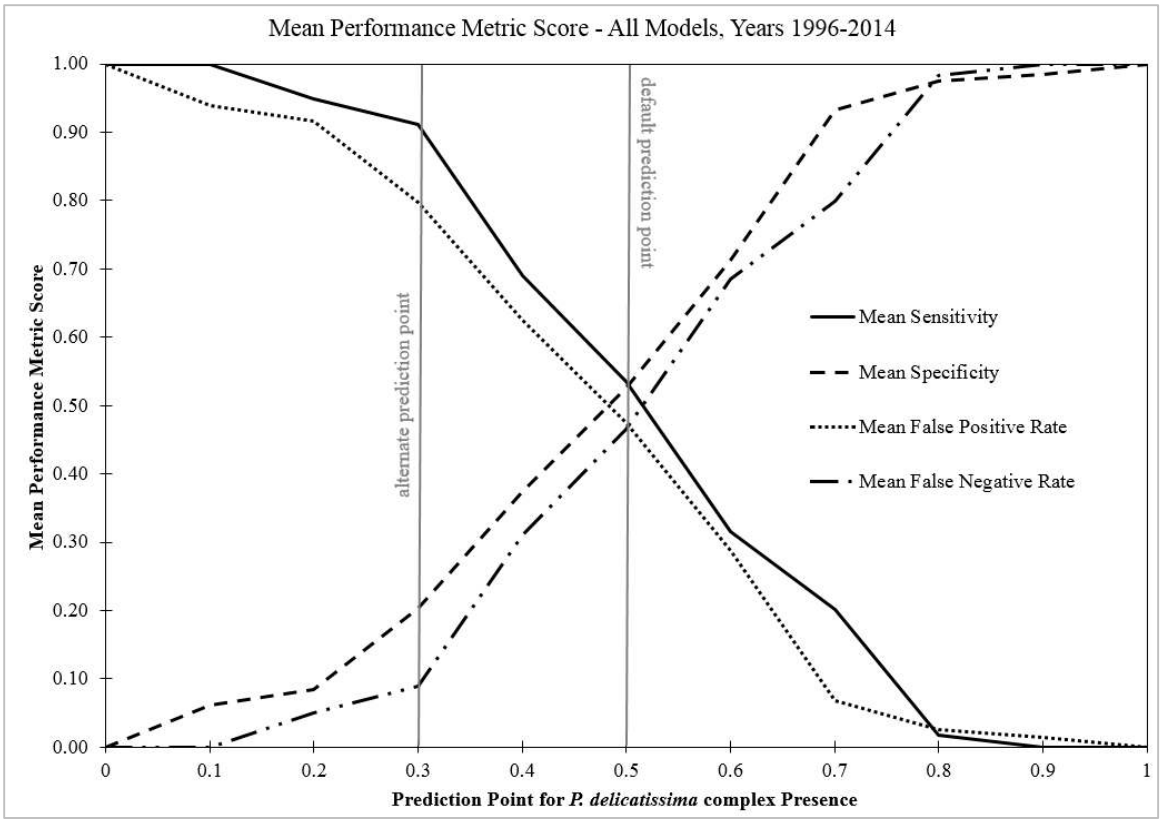

Figure 5-15. Mean performance metric score for ensemble of 19 cross-validation experiments for *P. delicatissima* model.

At the default prediction point of 0.5 the model ensemble performed reasonably well - seeming to balance sensitivity (0.54) and specificity (0.52), along with similar rates of false positive (0.48) and false negative (0.46) predictions. In general, on the test dataset at the default prediction point of 0.5 the model over-predicted the presence of *P. delicatissima* complex.

In addition to the default prediction point of 0.5 we present the results from an alternate prediction point of 0.3. We chose 0.3 because the test results indicated that at this prediction point the sensitivity was over 0.90. Such a prediction point would be

more conservative from a public health perspective as it would presumable capture almost all of the true 'presence events' of *P. delicatissima* complex. The tradeoff for this is poor performance (0.2) at detecting true negatives (cases where no *P. delicatissima* complex were observed). In order to make a very rudimentary comparison of model performance, we compared our model performance results to the only other dichotomous prediction logistic regression model that we know of for *Pseudo-nitzschia* species on the east coast of the U.S., the work of Anderson et al. (2010).[4] Although they used a different approach for model development and testing they did report sensitivity and false positive rates for their logit model.[4] Significant differences between this work and that of Anderson et al. (2010) are summarized below in Table 5-10.

Table 5- 10. Differences in *Pseudo-nitzschia* predictive modeling efforts for Chesapeake Bay and Massachusetts Bay.

Authors	Anderson et al. (2010)[4]	Current work
Dichotomous 'success' outcome used in model	Small blooms (≥10 cells/mL) of total *Pseudo-nitzschia* species diatoms in Chesapeake Bay	Observed presence of *P. delicatissima* complex diatoms at two stations in Massachusetts Bay
Total samples in dataset	6,989	229
Number of cases used for model training	6,989	197
Time span	1985 to 2007	1995 to 2014
Latitude span (approximate)	37 to 38.8 North	41.7 to 42.5 North
Surface salinity range at sampling sites	0.5 to ≥18 psu	24 to 35 psu*
Default prediction point 0.5: Model sensitivity	0.34	0.54
Default prediction point 0.5: Model false positive rate	0.03	0.48
Alternate prediction point	*0.19* *Chosen by Anderson et al.*	*0.3*
Alternate prediction point: Sensitivity	0.75	0.91
Alternate prediction point: False positive rate	0.09	0.80
*Recorded at Station 142[7] and offshore Buoy A01[8] in Massachusetts Bay.		

The notable differences between the study by Anderson et al. (2010) in Chesapeake Bay and our work in Massachusetts Bay only allows us to make a very tentative comparison between these two models. In comparison to the work of Anderson et al. (2010)[21], our model was developed using a dataset over an order of magnitude smaller (197 cases). A larger sample size might reduce the incidence of false positive predictions because more incidents of *P. delicatissima* complex presence in the dataset would likely

refine the predictive parameters. Another addition difference is that our study area has a smaller salinity range with higher salinity values reflecting the openness of Massachusetts Bay. Our study area is also further north and heavily influenced by the Gulf of Maine circulation system.[21] A model from the Chesapeake Bay region based on samples from high salinity nearshore areas would allow for more direct inter-model comparison. Simply identifying Chesapeake Bay *Pseudo-nitzschia* to the group or species level might reveal interesting regional differences in abundance and bloom event timing. In summary, our model is more over-predictive of *P. delicatissima* complex presence in Massachusetts Bay than the model developed by Anderson et al. (2010) for the presence of small blooms comprised of total *Pseudo-nitzschia* species in Chesapeake Bay. Overall our model performance can be described as adequate.

Enterococcus presence/absence prediction model. Using an information-theoretic approach we developed a model for the presence/absence of *Enterococcus* bacteria in recreational waters at three beaches along northern Massachusetts Bay. The motivating question for this model development was presented in chapter 3:

- Is it possible to hindcast levels of *Enterococcus* populations in specific areas of Massachusetts Bay with reasonable accuracy using the datasets collected for factors with known biological relevance to *Enterococcus* growth?

We developed a model using a training dataset containing 261 cases (50 presence, 211 absence), and then tested the model's hindcasting accuracy on a dataset containing 80

cases (consisting of 10 presence and 70 absence cases). Model performance in the four areas of sensitivity, specificity, false positive rate, and false negative rate for the training and test datasets is summarized below in Table 5-11. The R Studio code used to generate, selection, and perform cross-validation tests is included in Appendix A.

Table 5-11. Ensemble of Cross-Validation Performance Metrics: *Enterococcus* presence/absence prediction

Ensemble of Cross-Validation Performance Metrics: Enterococcus presence/absence prediction		
	Default prediction point	**Alternate prediction point**
Prediction point value	**0.5**	**0.1**
Sensitivity	0.07	0.69
Specificity	0.86	0.37
False Positive Rate	0.03	0.52
False Negative Rate	0.82	0.20

At the default prediction point of 0.5 the model cross-validation performed extremely poorly terms of sensitivity (0.07). The mean sensitivity was unable to correctly predict true positives. However, the model had a very high mean specificity (0.86) at the default prediction point. At an alternate prediction point of 0.1 the performance improved. The mean cross-validation ensemble sensitivity increased to 0.68, specificity declined to 0.37, and the false negative rate was low at 0.20. Overall the cross-validation results of the model for presence/absence of *Enterococcus* counts over 10 cells/100mL had poor performance. The range of mean performance metrics across

prediction points is shown below in Figure 5-16. In Figure 5-16 the default and alternate prediction points are indicated with a grey vertical dotted line. The alternate prediction point of 0.1 is more conservative from a public health perspective because it gives preference to higher sensitivity but also a higher false positive rate.

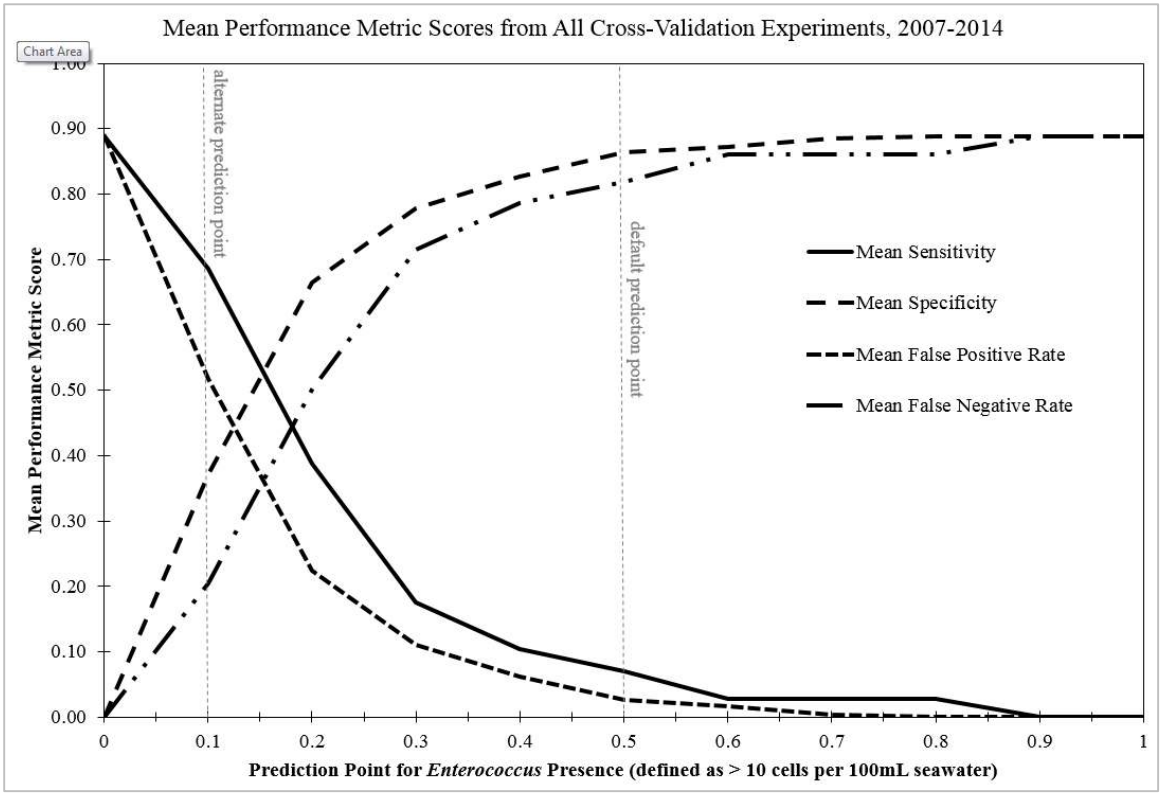

Figure 5-16. Mean performance metric score for ensemble of 8 cross-validation experiments for *Enterococcus* model.

Discussion of Predictive Models.

The previous sections have described model generation, model selection, and model cross-validation using year-by-year predictions generated with a model fitted by leaving out the predicted year. Below we discuss the overall utility of the selected

models for *P. delicatissima* complex and *Enterococcus*. We also discuss the question of correlation between the abundance of *P. delicatissima* complex and *Enterococcus* in the northern part of Massachusetts Bay. This section ends with suggestions for potential future modeling efforts.

***Pseudo-nitzschia delicatissima* complex model.** When predicting the presence/absence of *P. delicatissima* complex our model performance can be described as 'poor to adequate' with a bias towards over-prediction of *P. delicatissima* complex presence. One potentially encouraging result is that the model displayed high mean sensitivity (0.91) and a low mean false negative rate (0.09) when tested at the alternate prediction point (0.3). However the false positive rate was far higher than the false negative rate at both the default (0.5) and alternate (0.3) prediction points when tested. This bias towards over-prediction can be viewed as potentially more protective of public health, but the cost of over-prediction depends on who is using the model and to what purpose. In Massachusetts Bay the observed *Pseudo-nitzschia* abundance has varied across seasons and years.[17] Additionally, the literature reports intra-genus diversity in terms of environmental variables that influence *Pseudo-nitzschia* abundance in different regions,[2-4; 9] and there gaps in our knowledge about the dynamics of domoic acid production.[25]

Given that there is currently no official sampling program for *Pseudo-nitzschia* species in shellfish harvesting waters of Massachusetts, we suggest that a model such as

ours could be used as a rough guide to identify times when more frequent sampling should occur. Any type of response-based sampling would be more protective of human health than the current *status quo*. A limitation to using our model as it stands is that of the six parameters in the selected model three are based on macronutrient measurements from water column samples, another is the chlorophyll *a* measurement taken concurrently at the station. In other words, over half of the model parameters are currently measured *in situ*, with concurrent sampling for *Pseudo-nitzschia*. Two of the model parameters are based on physical measurements taken at other places. Precipitation observations are made at a land-based station to develop the variable 'prcp.day.before', and 'watertemp.44013' is based on water temperature measurements at buoy 44013 transmitted to shore with extremely high frequency. If remotely sensed proxies for the macronutrient levels and chlorophyll measurements become available at a useful level of resolution this could improve model development and forecasting. At the time of writing no such measurements were available with both spatial-temporal resolution and sufficient historical depth. However we expect that this will change in the future.

The results of our model selection and testing generate questions that could be refined into hypothesis. For example, *P. delicatissima* complex has been observed in Massachusetts Bay in all seasons so we chose to make a single annual model, but would seasonal partitioning of the data lead to different models or improved performance? Such an approach has been applied on the west coast of the U.S. where researchers developed separate models for annual, spring, and fall-winter conditions.[3] In that study the authors

observed that seasonality was a factor in model refinement, with only two predictor variables (chlorophyll *a* and silicic acid) included across all three models.[3] Another potential question relates to the temporal coverage of the data used to generate our model. The MWRA has 20+ years of sampling data, but is the limited annual coverage of 6-10 samples per year sufficient to capture the range of environmental conditions that may influence *P. delicatissima* complex, or total *Pseudo-nitzschia,* abundance? Our model was developed with samples that spanned the change in nutrient releases that accompanied the opening of the Deer Island Wastewater Treatment Plant in September 2000. Our training dataset consisted of the most recent 25% of samples from each station (roughly years 2011 to 2014), this might have influenced our model performance since maximum nitrogen levels around Station F23 decreased after the opening of the Deer Island Wastewater Treatment Plant.[7] In summary, the results of this model generation, selection, and cross-validation test exercise lead us to conclude the following: 1) at the alternate prediction point (0.3) model has some predictive (hindcast) value and could potentially be used in public health protection efforts as long as the high false positive rate does not result in costly managerial response actions without field sampling, 2) we suggest that more field sampling in Massachusetts Bay is required to develop a more accurate prediction model, and 3) there is tentative support for the influence of a small suite of variables on *P. delicatissima* complex abundance in Massachusetts Bay which may warrant further investigation.

***Enterococcus* model.** Our model for predicting the presence of *Enterococcus* (defined as counts over 10 cells/100mL) at three ocean-facing marine beaches along the northern coast of Massachusetts Bay performed poorly. In some ways this is not surprising. *Enterococcus* levels at these three study beaches are generally low during the summer months, especially in more recent years (see Figure 5-12, above, in the *Enterococcus* data description section). The relatively rare cases where *Enterococcus* levels are higher than 10 cells/100mL at these sites might be the result of stochastic factors not considered in this model. In places where routine test results indicate infrequent exceedances, it raises the possibility that these events are not driven by steady inputs from fixed land-based sources. Rather, rare exceedances could be linked to currently unrecorded phenomenon such as the presence of flocks of birds congregating onshore or in the intertidal zone or contributions from other wildlife populations that may shed *Enterococcus* through fecal waste. At present local health officials are allowed to proactively close beaches based on rainfall events.[23] In some locations with combined sanitary and storm sewers this may be a sensible precaution. However at our three study beaches precipitation the day before (prcp.bos.day.before) was a weaker predictor variable than maximum temperature at Marblehead (tmax.mblhd) and chlorophyll *a* at Buoy A01 (chl.a01). The relationship between *Enterococcus* levels and prcp.bos.day.before is shown below in Figure 5-17.

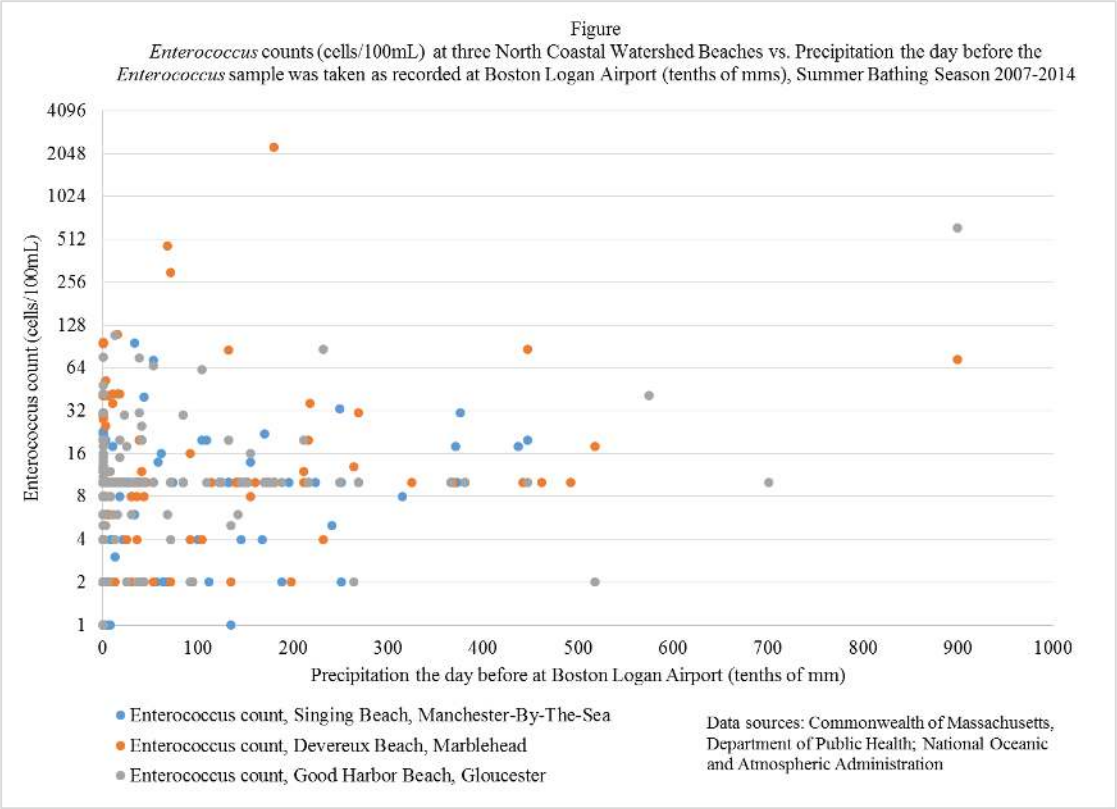

Figure 5-17. *Enterococcus* levels at three north coastal beaches vs. precipitation recorded at Boston Logan Airport on the previous day.

As shown in Figure 5-17 (above) it is possible to have high *Enterococcus* levels at the three study beaches after high, or low, levels of precipitation. Our poor model performance, and the known natural variability of this system, supports the rationale for direct water quality sampling as an appropriate strategy for water quality monitoring at recreational waters. At this point we are unable to make accurate location-specific predictions about *Enterococcus* abundance based on the available data for a limited suite of relevant variables. In summary, the results of this modeling generation and selection exercise lead us to conclude the following: 1) the most supported model from our

candidate set has a low level of predictive (hindcast) value at the default prediction point (0.5) and limited utility at a prediction point of 0.1, 2) direct field sampling as currently conducted is a more useful approach for assessing *Enterococcus* presence in recreational bathing waters, and 3) there is extremely tentative support for the relationship between chlorophyll a levels and *Enterococcus* levels in the same area which may warrant further direct investigation.

***P. delicatissima* complex and *Enterococcus* correlation.** In addition to modeling the presence/absence of *P. delicatissima* complex and *Enterococcus*, we were interested in identifying any possible relationship between the two since levels of *Enterococcus* and other fecal indicator bacteria are the current standards for recreational and shellfish-harvesting water quality.[37; 38] This question was presented in Chapter 3:

- Does there appear to be any clear relationship between *Enterococcus* levels and *Pseudo-nitzschia* levels in Massachusetts Bay?

There are very few cases where sampling for *P. delicatissima* complex at Station F22 and Enterococcus at one of the three north coastal beaches occurred on the same day. For Marblehead Devereux Beach there were 7 cases, for Manchester-By-The-Sea Singing Beach there were 9 cases, and Gloucester Good Harbor Beach there were 7 cases. The Spearman's rank correlation coefficient test results for *Enterococcus* levels at each beach and corresponding *P. delicatissima* complex counts at Station F22 are as follows:

- Station F22 and Marblehead Devereux Beach: -0.224

- Station F22 and Manchester-By-The-Sea Singing Beach: -0.230
- Station F22 and Gloucester Good Harbor Beach: 0.167

Based on the available monitoring data at Station F22 (for *P. delicatissima* complex) and three proximal north coastal beaches (for *Enterococcus*) there does not appear to be any relationship between the abundance levels of these organisms. We stress that this conclusion is based on field monitoring data alone, not on purpose-designed experiments under laboratory controlled conditions. Observations at these locations spanning the years 2007 to 2014 are shown below in Figure 5-18, note the log base 2 vertical scale. As shown in Figure 5-18 (below), it is possible to have high *Enterococcus* levels without any leading or lagging high *P. delicatissima* complex levels, and it is possible to have high *P. delicatissima* complex levels without any clear leading or lagging signal in the *Enterococcus* counts.

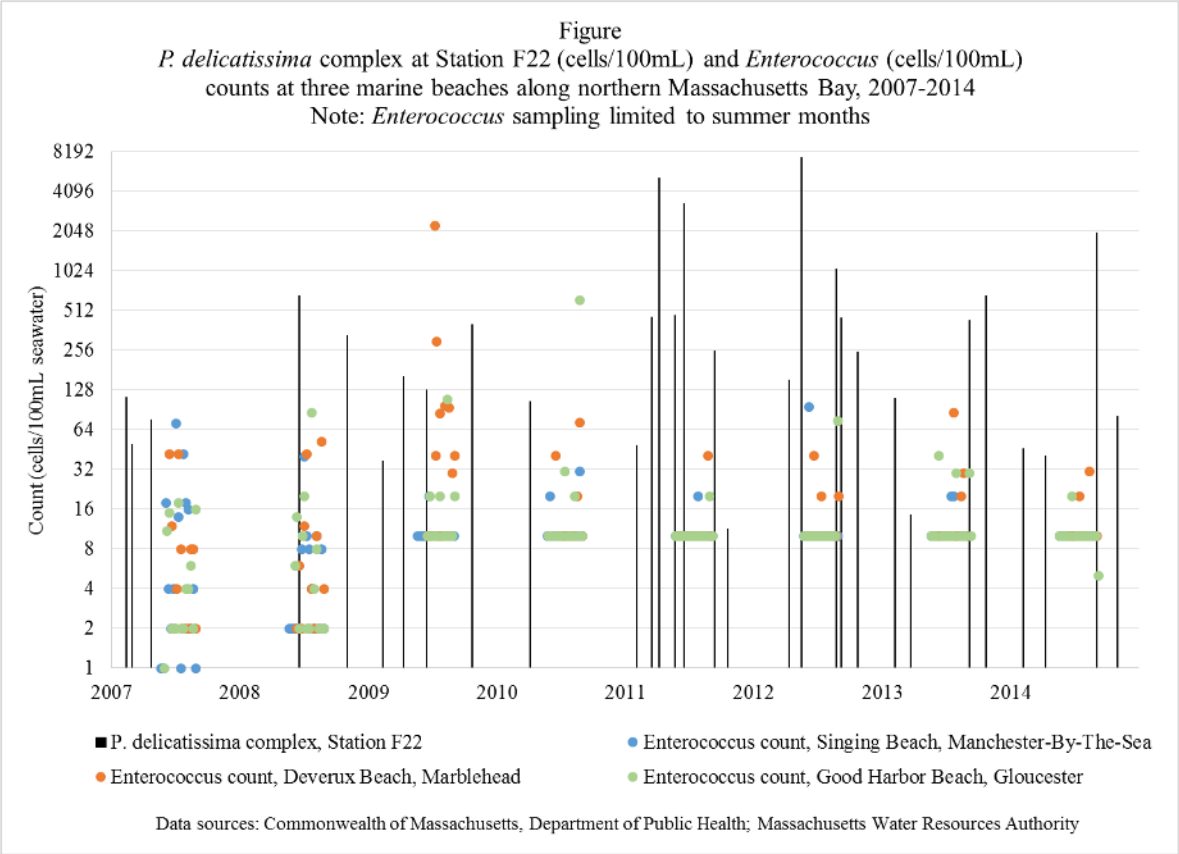

Figure 5-18. *P. delicatissima* complex at Station F22 and *Enterococcus* at three north coastal beaches, 2007-2014.
Sample values of 0 cells/100mL are not shown in this figure due to natural log vertical scale.

Neither population is sampled continuously, nor are there many instances of sample collection for both organisms on the same day, and we acknowledge this limitation of the data. A statistical investigation of correlation between the two organisms would have required extensive temporal interpolation of abundances, and significant assumptions about the doubling times, or die-off rates, of these organisms in this area. However, given our limited understanding of *P. delicatissima* complex baseline population levels

and reproduction rates, and the mixed literature results about *Enterococcus* sources in the wild we did not think such assumptions would lead to useful results.

At present *Enterococcus* levels are sampled from late May to early September, leaving only four or five potential opportunities for sample overlap in any given year under the current MWRA monthly monitoring schedule. It is possible that higher resolution sampling would reveal more common factors that could be subject to examination, or at least a better understanding of *P. delicatissima* complex bloom dynamics. The only clear commonality between *P. delicatissima* complex and *Enterococcus* abundance is that chlorophyll *a* is a predictive parameter in both of our selected models. Such an observation may be useful for future hypothesis development.

Suggestions for future work. The long-term goal of this work is to develop simultaneous forecasts for multiple marine-sourced risks. The final element of our Phase 3 evaluation of outputs is to revisit the question raised in the previous chapter:

- Are there any field measurements for which public data do not readily exist which scientific literature suggests would likely increase the predictive ability of these models?

Based on our evaluation of the assembled data we have multiple suggestions, some of which have already been mentioned (e.g., remote sensing measurements of macronutrients). For *Pseudo-nitzschia* species in Massachusetts Bay increased temporal resolution sampling would be useful for potentially identifying seasonal influences on

presence/absence, or even bloom size levels. Future work could benefit from a purpose-designed multi-year study with high temporal resolution that would have a better chance of capturing the subtleties of nutrient dynamics as they relate to *P. delicatissima* complex (or total *Pseudo-nitzschia*) abundance. Future work could also benefit from a broader effort to acquire and compile nutrient data (or other multi-purpose data) that might have scientific value for understanding the risk potential from multiple marine-based organisms present in the system, as was suggested in Chapter 2.

We could not identify any publicly available datasets for trace metals in Massachusetts Bay. Iron, copper, and lithium have been suggested as influences on either *Pseudo-nitzschia* species abundance and/or domoic acid production. Sampling for those metals, combined with mesocosm growth experiments under conditions mimicking Massachusetts Bay, could potentially improve our understanding of *Pseudo-nitzschia* responses to regional conditions and inform a predictive model.

Our model used *Pseudo-nitzschia* observations from the northern end of Massachusetts Bay, but shellfish harvesting is more common in the southern part of the Bay. Collecting *Pseudo-nitzschia* samples from the southern part of Massachusetts Bay could help identify the spatial extent of *Pseudo-nitzschia* presence which (the MWRA sampling ended for *Pseudo-nitzschia* at Station F02 ended in 2010 as far as we know). We note that there is a new, non-regulatory, pilot program in Massachusetts to engage volunteers in collecting and analyzing plankton samples for potentially toxigenic

species.[39] We hope that the results of this and any subsequent program are collected and made available to the general public through the MA-DMF.

Predictive modeling efforts for *Enterococcus* could benefit from increased information capture at the time of sampling. For example, observations of wildlife present in the area might indicate if high *Enterococcus* levels are from non-human sources. In addition, recording physical variables such as water temperature at the time and location of sampling might reveal subtle differences that contribute to *Enterococcus* persistence. Some locations with persistently poor water quality have initiated microbial source tracking efforts to identify *Enterococcus* to the species level and then match that to known host organisms which may be present upstream. Although this might not be necessary for the three beaches used in our study it is an important scientific development which can be used in other situations. Another type of data which would be useful is bather attendance counts for public beaches on every day of the summer bathing season, not just bather presence at the time of sampling. At present these are not collected or published for Massachusetts beaches in a coordinated way. Given that direct shedding of *Enterococcus* by bathers has been suggested to impact water quality[23] this information would fundamentally improve our understanding of that potential loading source at Massachusetts beaches.

The currently available data for *Enterococcus* and *Pseudo-nitzschia* spp. in Massachusetts Bay is not collected for the purpose of developing predictive models or exploring potential relationships between different types of marine-sourced risks. There

is little overlap on sampling dates, and we found no mention in any public sources of plans for coordination in future. Such coordination might add important scientific value to data already being collected for routine monitoring purposes. For other marine-sourced risks known to exist in Massachusetts Bay (e.g. *Vibrio parahaemolyticus*, anthropogenic antibiotics, human enteric viruses) there are no equivalent long-term monitoring programs. However, the Massachusetts Division of Marine Fisheries has initiated a limited sampling program for *Vibrio parahaemolyticus*[38], but results are not yet published in the same way as *Enterococcus* counts for recreational water quality. We suggest that all state and federal monitoring programs with close ties to public health should publish their data online in a timely fashion (no more than a 1-year time lag) as this would facilitate data discovery and identify potential opportunities for coordination and collaboration.

Summary Conclusion.

In this chapter we used an information-theoretic approach to develop a suite of candidate models that we tested against each other to find the one with the most support based on an information criteria measure. These models were developed *a priori*, informed by our understanding of previous modeling efforts, the biology of *Pseudo-nitzschia* spp. diatoms and *Enterococcus* bacteria, and our knowledge of the Massachusetts Bay area. We used Akaike's Information Criteria (AIC) to identify which model in our candidate set had the most support, and was thus estimated to be the closest

estimation of reality among the candidate models in the set. The publicly available data used to develop our models was divided into two parts: 75% of the total cases in the dataset were used for model selection, and the final 25% of cases were reserved for testing the selected model.

The selected model for the presence/absence of *P. delicatissima* complex performed poorly-to-adequately when tested. The model over-predicted the presence of *P. delicatissima* complex, with a false positive rate of 0.8 at the alternate prediction point. Such over-prediction may have public health value if the model were used to guide managerial responses that started with low-cost water sampling efforts to confirm the presence of potentially toxigenic organisms. The selected model for the presence/absence of *Enterococcus* at three north coastal beaches performed poorly when tested. It displayed 0.07 mean sensitivity and a high mean possible false negative rate (0.82) at the default prediction point of 0.5. An alternate prediction point of 0.1 raised the mean sensitivity (to 0.69) and decreased the mean false negative rate (0.20), but such a low prediction point is so conservative it results in a model with very little utility.

In addition, we discerned no relationship between the presence or absence of *Enterococcus* in recreational water samples and the presence or absence of *P. delicatissima* complex in water samples taken further offshore. These results support the argument for continued direct sampling for microbial risk factors at recreational bathing waters or shellfish harvesting waters. In light of these results we suggest that a purpose-designed high-temporal-resolution sampling effort for *Pseudo-nitzschia* species in

Massachusetts Bay could dramatically improve our understanding of the dynamics of this potentially toxigenic organism in the region. Ongoing water quality monitoring efforts, experimental results, and remote sensing outputs (which continue to improve in resolution) will all have value in understanding marine-sourced risks to human health.

Literature Cited.

1. Adl, S. M., Simpson, A. G., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V. 2012. The revised classification of eukaryotes. J. Eukaryot. Microbiol. 59: 429-514.

2. Kaczmarska, I., Martin, J. L., Ehrman, J. M., LeGresley, M. M. 2007. *Pseudo-nitzschia* species population dynamics in the Quoddy Region, Bay of Fundy. Harmful Algae. 6: 861-874.

3. Lane, J. Q., Raimondi, P. T., Kudela, R. M. 2009. Development of a logistic regression model for the prediction of toxigenic *Pseudo-nitzschia* blooms in Monterey Bay, California. Mar. Ecol. Prog. Ser. 383: 37-51.

4. Anderson, C. R., Sapiano, M. R. P., Prasad, M., Long, W., Tango, P. J., Brown, C. W., Murtugudde, R. 2010. Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the Chesapeake Bay. J. Mar. Syst. 83: 127-140.

5. Johnson, C. N., Bowers, J. C., Griffitt, K. J., Molina, V., Clostio, R. W., Pei, S., Laws, E., Paranjpye, R. N., Strom, M. S., Chen, A., Hasan, N. A., Huq, A., Noriea, N. F.,3rd, Grimes, D. J., Colwell, R. R. 2012. Ecology of *Vibrio parahaemolyticus* and *Vibrio vulnificus* in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washington (United States). Appl. Environ. Microbiol. 78: 7249-7257.

6. National Oceanic and Atmospheric Administration. 2015. National Data Buoy Center: Station BHBM3 - 8443970 - Boston, MA. U.S. Department of Commerce. Stennis Space Center, MS. http://www.ndbc.noaa.gov/station_page.php?station=bhbm3 (Accessed May 25, 2015).

7. Massachusetts Water Resources Authority. 2015. Boston Harbor and Massachusetts Bay: Water Quality Data
. Massachusetts Water Resources Authority. Boston, MA. http://www.mwra.state.ma.us/harbor/html/wq_data.htm (Accessed May 17, 2015).

8. Northeastern Regional Association of Coastal and Ocean Observing Systems. 2014. NERACOOS: Data & Tools. The Gulf of Maine Research Institute. Portland, ME. http://neracoos.org/datatools (Accessed May 25, 2015).

9. Downes-Tettmar, N., Rowland, S., Miller, P., Llewellyn, C. 2013. Seasonal variation in *Pseudo-nitzschia* spp. and domoic acid production in the Western English Channel. Continental Shelf Research. 53: 40-49.

10. Burnham, K. P., Anderson, D. R. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd Edition). Springer. New York, N.Y.

11. Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. Proceeding of the Second International Symposium on Information Theory. : 267-281.

12. Grueber, C. E., Nakagawa, S., Laws, R. J., Jamieson, I. G. 2011. Multimodel inference in ecology and evolution: challenges and solutions. J. Evol. Biol. 24: 699-711.

13. Massachusetts Water Resources Authority. 2012. Water Column Monitoring in Massachusetts Bay: 1992 - 2006.

14. National Oceanic and Atmospheric Administration. 2015. National Data Buoy Center: Historical NDBC Data. U.S. Department of Commerce. Stennis Space Center, MS. http://www.ndbc.noaa.gov/historical_data.shtml (Accessed November 12, 2015).

15. National Oceanic and Atmospheric Information, National Centers for Environmental Information. 2015. National Centers for Environmental Information: Climate Data Online Search. U.S. Department of Commerce. Silver Spring, M.D. http://www.ncdc.noaa.gov/cdo-web/confirmation (Accessed December 31, 2014).

16. Libby, P. S., Geyer, W. R., Keller, A. A., Mansfield, A. D., Turner, J. T., Borkman, D., Oviatt, C. A. 2006. 2004 Annual Water Column Monitoring Report. Report ENQUAD 2006-15. Massachusetts Water Resources Authority. Boston, M.A. 1-177.

17. Massachusetts Water Resources Authority. . 2015. pseudonitz_1992-2014.xlsx [MS Excel file]. M. Kress.

18. Libby, P. S., Fitzpatrick, M. R., Buhl, R. L., Lescarbeau, G. R., Leo, W. S., Borkman, D. G., Turner, J. T. 2014. Quality assurance project plan (QAPP) for water column monitoring 2014-2016: Tasks 4-7 and 10. Report 2014-01. Massachusetts Water Resources Authority. Boston, M.A. 1-67.

19. Massachusetts Water Resources Authority. 2009. The Deer Island Sewage Treatment Plant. Massachusetts Water Resources Authority. Boston, M.A. http://www.mwra.com/03sewer/html/sewditp.htm (Accessed October 18, 2015).

20. Angus, T. H. 2015. Examining the Toxicity, Exposure, and Regulatory Approach to Potential Human Health Risks of the Algal Toxin Domoic Acid. Doctor of Philosophy thesis, University of Massachusetts Boston, Boston, MA.

21. Jiang, M., Zhou, M. 2008. The Massachusetts Bay Hydrodynamic Model: 2005 Simulation. Report 2008-12. Massachusetts Water Resources Authority. Boston, M.A. 1-63.

22. Commonwealth of Massachusetts, Department of Public Health. 2014. Marine Beaches in Massachusetts, Annual Beach Reports. Commonwealth of Massachusetts. Boston, MA. http://mass.digitalhealthdepartment.com/public_21/index.cfm (Accessed 01/24, 2015).

23. Commonwealth of Massachusetts, Department of Public Health. 2015. Marine and Freshwater Beach Testing in Massachusetts, Annual Report: 2014 Season. Commonwealth of Massachusetts. Boston, M.A. 1-151.

24. U.S. Census Bureau. 2014. 2010 Census Interactive Population Search: Massachusetts. U.S. Department of Commerce. Washington, D.C. http://www.census.gov/2010census/popmap/ipmtext.php?fl=24 .

25. Lelong, A., Hégaret, H., Soudant, P., Bates, S. 2012. *Pseudo-nitzschia* species, domoic acid and amnesic shellfish poisoning: revisiting previous paradigms. 51: 168-216.

26. Chatterjee, S., Simonoff, J. S. 2013. Wiley Handbooks in Applied Statistics : Handbook of Regression Analysis. John Wiley & Sons. Somerset, N.J.

27. Mazerolle, M. J. 2015. Package: AICcmodavg: Model selection and multimodel inference based on (Q) AIC (c). 2.0-3.

28. RStudio Inc. 2009. R Studio, Version 0.97.551.

29. Wei, T. 2013. Package: corrplot. 0.73.

30. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. Nature. 457: 1012-1014.

31. Loureiro, S., Jauzein, C., Garcés, E., Collos, Y., Camp, J., Vaqué, D. 2009. The significance of organic nutrients in the nutrition of *Pseudo-nitzschia delicatissima* (Bacillariophyceae). J. Plankton Res. 31: 399-410.

32. Sarthou, G., Timmermans, K. R., Blain, S., Tréguer, P. 2005. Growth physiology and fate of diatoms in the ocean: a review. J. Sea Res. 53: 25-42.

33. 314 CMR 4.00: Massachusetts Surface Water Quality Standards. 2013. 314: 01-06.

34. Mote, B. L., Turner, J. W., Lipp, E. K. 2012. Persistence and growth of the fecal indicator bacteria enterococci in detritus and natural estuarine plankton communities. Appl. Environ. Microbiol. 78: 2569-2577.

35. Fisher, K., Phillips, C. 2009. The ecology, epidemiology and virulence of *Enterococcus*. Microbiology. 155: 1749-1757.

36. Wonnacott, T. H., Wonnacott, R. J. 1990. Introductory Statistics, 5th Edition. Wiley. New York, NY.

37. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Public Health Protection: Shellfish Sanitation. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/public-health-protection.html#ssanitation (Accessed December 27, 2015).

38. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Massachusetts Marine Fisheries: 2014 Annual Report. Commonwealth of Massachusetts. Boston, M.A. 1-123.

39. Commonwealth of Massachusetts, Division of Marine Fisheries. 2015. Public Health Protection: Shellfish Sanitation. Commonwealth of Massachusetts. Boston, M.A. http://www.mass.gov/eea/agencies/dfg/dmf/programs-and-projects/public-health-protection.html (Accessed October 19, 2015).

CHAPTER 6

CONCLUSION

This conclusory section brings together information from the previous chapters to summarize the findings of this dissertation as they relate to the study area of Massachusetts Bay and the neighboring coastal watersheds. The previous chapters dealt with 1) frameworks to understand and organize both problem-framing and response actions; 2) the human demographics of the study area followed by a review of the state of knowledge about five marine-sourced risks that exist in the area and may disproportionately affect the study population; 3) a method for investigating interdisciplinary science questions and a discussion of the practice of data science; and 4) development and testing of predictive models for two potentially harmful marine-sourced risks using publicly available data.

Organizing Frameworks.

Chapter 1 introduced two frameworks that can be used to organize, understand, and communicate information about environmental or health problems, and shape possible solutions to those problems and evaluation measures. Those two frameworks are the Driver-Pressure-State-Impact-Response (DPSIR) and the Driver-Pressure-State-

Exposure-Effect-Action (DPSEEA) frameworks. The DPSEEA framework is a version of the DPSIR that has been tailored for use in the public health and medical communities, but both frameworks contain the same essential structure.

The DPSIR framework, being more general, has seen wide use in a variety of research and policy assessment efforts, eleven such applications were described in Chapter 1. These applications included identifying and framing environmental challenges which may be unique to coastal megacities around the world[1]; understanding historical influences on development practices in South Africa[2; 3]; linking upstream influences with downstream impacts on bathing beach water quality in Venice, Italy[3]; identifying the forces influencing coastal wetland loss in Xiamen, China[4]; and contrasting environmental management challenges in three coastal cities in different parts of South America.[5] In addition to the DPSIR applications, we presented summaries of two DPSEEA applications where policy makers are developing solutions to interlinked challenges that involve the natural environment, physical infrastructure, and health. One example is from São Paulo, Brazil and the other is from Scotland, both examples demonstrate the flexibility of the DPSEEA framework when addressing location specific health challenges.

The work in Chapter 1 demonstrated that the DPSIR and DPSEEA frameworks are useful for organizing information in a flexible and question-specific way across multiple environmental and public health issues. The structure of these frameworks allows people to see where alternative solutions might fit within a suite of possible

response/action choices. By specifying causal relationships in advance, users can also identify evaluation measures that will allow policy makers to measure the success of an implemented *response/action*. The framework structure also allows for identification of data gaps which might need to be remedied before committing to a *response/action*. Transparency and accountability are important components of policy-making and the use of DPSIR and DPSEEA frameworks supports those principles. Through these integrating frameworks researchers and policy-makers can identify which actions within complex systems can best support environmental and human health.

Human Population Demographics and Marine-sourced Risks in Massachusetts Bay.

Chapter 2 introduced the study area, Massachusetts Bay and the neighboring coastal watersheds, which includes the coastal city of Boston and much of the surrounding metropolitan area. Proximity to the sea lends itself to the potential for greater physical interaction with the ocean or locally harvested seafood, both avenues for exposure to marine-sourced risks. Characteristics such as age distribution can influence a population's overall susceptibility to infectious agents[6] or toxins, so we used data from the U.S. Census Bureau to examine human population demographics and dynamics in Massachusetts Bay coastal watersheds at the year 2000 and year 2010 timepoints.[7; 8] Using coastal watersheds as the spatial unit of interest, this chapter presented original estimates of watershed populations along with important population characteristics such as median income, and percent of population over age 65. Between 2000 and 2010 total population increase in the coastal watersheds around Massachusetts Bay was less than 3

percent, similar to the state as a whole. The Cape Cod watershed was the only Massachusetts Bay watershed to see a total decrease in the resident population in the same time period. The results showed that the Cape Cod watershed had both the highest percent of residents over age 65, and the lowest median income among all watersheds, indicating a potentially greater population-level susceptibility to marine-sourced risks.

Chapter 2 also described five categories of marine-sourced risk (enteric bacteria, indigenous marine bacteria, enteric viruses, natural marine toxins, and anthropogenic compounds) and then identified a specific example from each category known to exist in the Massachusetts Bay area. The specific risks chosen were *Enterococcus* bacteria, *Vibrio parahaemolyticus* bacteria, Hepatitis A Virus, *Pseudo-nitzschia* genus diatoms which can produce the toxin Domoic Acid, and anthropogenic antibiotics. Each of these risks is associated with a reportable illness in Massachusetts.[9-11] In Chapter 2 we reviewed existing epidemiological data for these risks (at the national or state level if available). Since epidemiological data for seafood-borne and recreational water-borne risks is known to be incomplete[12-14] we also reviewed the biology of these risks in the natural environment. The weight of evidence of existing epidemiological data combined with the known natural history of these risks strongly suggests that each can be present with varying abundance in Massachusetts Bay. Abundance of these risks is likely a product of both environmental variability and human-driven influences. This chapter closed with a matrix showing known influences on each example risk. Influences that affect more than one type of risk are considered high value data types that would be

useful in multi-risk modeling efforts. High value influences identified in this chapter included both environmental influences (e.g., water temperature, sunlight, rainfall or other freshwater input) and socio-economic influences (e.g., composition and volume of anthropogenic nutrient releases, wastewater treatment type, and local human population). If this same approach were used in a different location users could customize the marine-sourced risks of interest, the influences identified, and the subsequent data needs.

Interdisciplinary Data Science.

The purpose of Chapter 3 was to present an overview of the data landscape that currently exists for interdisciplinary environmental health researchers, and to provide a generalized workflow that others could use to organize and plan interdisciplinary work. In addition, Chapter 3 discussed the interrelated topics of big data, crowdsourced data, data science, and the associated challenges and opportunities of evolving data sources.[15-18] Big data refers to millions, or billions, of records of a certain type, common examples include financial transaction records, electronic medical records, and social network-derived datasets.[19-22] The rise of big data has necessitated the development of new analytical and technological tools to manage and query these datasets. Crowdsourced data can be generated purposely (e.g., voluntary contributions to research efforts[23], aggregated commentary on the same topic) or anonymously (e.g., geo-referenced location data from mobile phones, online search queries[24]). Crowdsourced data may become big data, and data scientists may combine crowdsourced data, big data, and traditional

scientific data from multiple disciplines to ask new questions. Others are already using crowdsourced data to try to detect regional patterns of certain infectious diseases[25], but success so far has been mixed.[15] Although the modeling work in this dissertation does not use crowdsourced data or big data, we expect that future environmental health work may be able to take advantage of these sources.

Chapter 3 presented three examples of how researchers have combined epidemiological and medical data with remote-sensing data to gain new insights into diseases. Those three cases were 1) Rift Valley Fever in the Horn of Africa[26; 27], 2) cholera in Bangladesh[27], and 3) Kawasaki disease in Japan.[28] Rift Valley Fever is caused by a virus spread by mosquitoes[26; 27], cholera is caused by bacteria spread through fecal-contaminated food or water[27], and the cause of Kawasaki disease is unknown but suspected to be an inhaled natural substance such as an aerosolized fungus or bacteria.[28] Through our generalized workflow process we showed that investigations of different diseases can follow the same general workflow even when the research products are very different. In all cases the first phase of interdisciplinary work is to review the existing scientific literature and identify potentially relevant data sets. The second phase of interdisciplinary work is to produce outputs, such outputs may include disease maps, historical timelines, prediction maps, or predictive models. The third phase of this interdisciplinary work is to evaluate the outputs from the second phase and assess their utility. The evaluation phase asks the following questions:

- Is there a useful level of predictive value in the output product?

- Is further field sampling needed to better understand the system?
- Do these outputs further the development of theory?

As with any exploratory endeavor, it is possible that initial attempts to reveal linkages between environmental factors and human health may raise more questions than they answer.

This dissertation asked questions about marine-sourced risks in Massachusetts Bay. To that end we included a list of datasets relevant to investigating marine-sourced risks in Massachusetts Bay at the end of Chapter 3. This list included multi-year monitoring datasets for *Pseudo-nitzschia* species diatoms and *Enterococcus* bacteria in different parts of Massachusetts Bay, we were unable to locate comparable datasets for anthropogenic antibiotics, Hepatitis A Virus, or *Vibrio parahaemolyticus*. We used the identified data sets of environmental variables, socio-economic variables, and marine-sourced risks to develop probabilistic models, discussed in Chapter 4.

Marine-sourced Risk Models.

The purpose of Chapter 4 was to use publicly available data to develop probabilistic predictive models for the presence / absence of the diatom *Pseudo-nitzschia delicatissima* complex and the bacteria *Enterococcus*, both of which have been shown to be present in Massachusetts Bay at different times. We used an information-theoretic approach to select the most supported model from a suite of logistic regression models developed *a priori*. Each model represented a unique hypothesis to explain the

presence/absence *P. delicatissima* complex or *Enterococcus*, each model was developed using our understanding of the biology of these organisms, our knowledge of the Massachusetts Bay system, and the results of previous modeling efforts from other locations (where available).

We used public, but unpublished, data from the Massachusetts Water Resources Authority containing *Pseudo-nitzschia delicatissima* complex counts to develop a dichotomous presence/absence response variable for *P. delicatissima* complex.[29; 30] Potential input variables were developed using macronutrient measurements from sampling stations in Massachusetts Bay along with weather, riverflow, and oceanographic records from other public sources.[31-33] Similarly, to develop the *Enterococcus* presence/absence response variable we used public data from bathing beach water quality testing, published by the Commonwealth of Massachusetts Department of Public Health, Bureau of Environmental Health.[34] Weather records, oceanographic observations, and census records from other public sources were used to develop the predictor variables for *Enterococcus* abundance.[31-33] The total dataset containing response and predictor variables was used to generate and identify the most supported model. Cross-validation testing involved removing one year of data, fitting the model on the remaining year, and then using the fitted model to make predictions for the missing year. For the *P. delicatissima* complex there were 19 years of available data, and thus an ensemble of 19 cross-validation experiments. For Enterococcus we used 7 years of data and thus had 7 cross-validation experiments.

We tested the hindcast performance of each predictive model against its respective training dataset and measured performance in four areas: sensitivity, specificity, false positive rate, and false negative rate. This allowed us to see how closely the cross-validation results of hindcast probabilistic predictions matched observed responses. The *P. delicatissima* complex predictive model was biased towards over-prediction of diatom presence, there was a high false positive rate when tested on the training dataset. There was a very low false negative rate. Although the model was not highly accurate in predicting either presence or absences, the bias towards over-prediction suggests that such as model could be used to guide low-cost response efforts such as increased field sampling to detect the presence of *P. delicatissima* complex or any *Pseudo-nitzschia* species in Massachusetts Bay. Potential improvements in model utility could be achieved through remote sensing of predictor (input) variables such as macronutrients; at present macronutrients are measured through direct field sampling.

The *Enterococcus* predictive model was biased towards under-prediction, it had a high false negative rate. Overall the model had poor performance, suggesting that the current method of direct field sampling has more relevance to public health protection at the three beaches which we used at data sources for *Enterococcus* response. One potential limitation in our model development is the low frequency of high *Enterococcus* counts at these beaches, especially in more recent years. In addition to the predictive models we examined the data for any signs of a relationship between *P. delicatissima* complex abundance at offshore sampling stations and *Enterococcus* levels at three north

coastal bathing beaches. Our data for this comparison was limited to cases where sampling for both organisms occurred on the same day, Marblehead Devereux Beach had 7 cases, Manchester-By-The-Sea had 9 cases, and Gloucester Good Harbor Beach had 7 cases. The Spearman's rank correlation coefficient test results indicated no statistically significant relationship between the presence of these two organisms at their sampling locations.

Summary Conclusion.

The purpose of this dissertation was to 1) discuss the utility and applicability of the DPSIR and DPSEEA organizing frameworks, 2) examine human demographics in coastal watersheds around Massachusetts Bay and identify marine-sourced risks that may affect those populations through a review of epidemiological and biological data for five different kinds of risk, 3) discuss the current opportunities and challenges for interdisciplinary environmental health science research, and 4) develop probabilistic predictive models for two marine-sourced risks known to exist in Massachusetts Bay. Long-term data collected for other purposes shows that the potentially toxigenic diatom *Pseudo-nitzschia delicatissima* complex has repeatedly been present in Massachusetts Bay during all seasons and at varying abundance over the past twenty years. Other regions where molluscan shellfish are regularly harvested for human consumption have implemented direct monitoring for *Pseudo-nitzschia* species, including Washington State[35] and Great Britain.[36] Such an approach may be warranted in Massachusetts Bay until the accuracy of predictive models reaches a satisfactory level. At present there may

be unrecognized consumption of the neurotoxin domoic acid produced by *Pseudo-nitzschia* genus diatoms via shellfish harvested in Massachusetts Bay, and a growing body of research suggests that consumption of any amount of domoic acid may be harmful to mammals.[37-41] This suggests that there is a potentially under-appreciated public health risk receiving very little attention at present in Massachusetts.

While we recognize the value of direct sampling to monitor for marine-sourced risks, sampling for all known human health risks that exist in the nearshore coastal environment may simply be beyond the scope of public health authorities. In such cases a multi-risk predictive modeling effort built upon existing data and a thorough understanding of local system dynamics may help guide public health protection efforts. Predictive modeling, combined with direct sampling as needed and follow-up action by public health authorities, has the potential to reduce exposure to marine-sourced risks that may harm humans who interact with, or consume raw seafood harvested from, coastal ocean waters.

Literature Cited.

1. Sekovski I., A. Newton & W. C. Dennison. 2012. Megacities in the coastal zone: Using a driver-pressure-state-impact-response framework to address complex environmental problems. Estuar. Coast. Shelf Sci. 96: 48-59.

2. Palmer B. J., T. R. Hill, G. K. Mcgregor *et al*. 2011. An Assessment of Coastal Development and Land Use Change Using the DPSIR Framework: Case Studies from the Eastern Cape, South Africa. Coast. Manage. 39: 158-174.

3. Pastres R. & C. Solidoro. 2012. Monitoring and modeling for investigating driver/pressure–state/impact relationships in coastal ecosystems: Examples from the Lagoon of Venice. Estuar. Coast. Shelf Sci. 96: 22-30.

4. Lin T., X. Z. Xue & C. Y. Lu. 2007. Analysis of coastal wetland changes using the "DPSIR" model: a case study in Xiamen, China. Coast. Manage. 35: 289-303.

5. Bowen R. E., M. Kress, G. Morris *et al*. 2014. Integrating Frameworks to Assess Human Health and Well-Being in Marine Environmental Systems. *In* Oceans and Human Health: Implications for Society and Well-being. Bowen R. E., M. H. Depledge, C. P. Carlarne *et al*, Eds.: 23-45. John Wiley & Sons, Ltd. Oxford, England.

6. Centers for Disease Control and Prevention. 2015. People at High Risk of Developing Flu–Related Complications. Centers for Disease Control and Prevention. http://www.cdc.gov/flu/about/disease/high_risk.htm (Accessed June 22, 2015).

7. U.S. Census Bureau. 2015. Maps & Data: TIGER Products. U.S. Department of Commerce. http://www.census.gov/geo/maps-data/data/tiger.html (Accessed July 2, 2015).

8. U.S. Census Bureau. 2014. 2010 Census Interactive Population Search: Massachusetts. U.S. Department of Commerce. http://www.census.gov/2010census/popmap/ipmtext.php?fl=24 .

9. Commonwealth of Massachusetts, Department of Public Health. 2013. Communicable and Other Infectious Diseases Reportable in Massachusetts by Healthcare Providers (PDF File). : 1-2.

10. Commonwealth of Massachusetts, Department of Public Health. 2013. Communicable and Other Infectious Diseases Reportable in Massachusetts To Local Boards of Health (PDF File). : 1.

11. Commonwealth of Massachusetts, Department of Public Health. 2013. Communicable and Other Infectious Diseases Reportable in Massachusetts by Clinical Laboratories (PDF File). : 1.

12. Centers for Disease Control and Prevention. 2015. National Enteric Disease Surveillance: COVIS Annual Summary, 2013. : 1-13.

13. Centers for Disease Control and Prevention. 2015. National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS): Human Isolates Final Report, 2013. : 1-81.

14. Hlavsa M. C., V. A. Roberts, A. M. Kahler *et al*. 2015. Outbreaks of illness associated with recreational water—United States, 2011–2012. MMWR. 64: 668-672.

15. Lazer D. M., R. Kennedy, G. King *et al*. 2014. The parable of Google Flu: traps in big data analysis. Science. 343: 1203-1205.

16. Madrigal A. C. 2014. In Defense of Google Flu Trends. The Atlantic.

17. Dryad. 2015. Dryad: Frequently Asked Questions. Dryad. http://datadryad.org/pages/faq (Accessed October 10, 2015).

18. IBM. 2015. What is a Data Scientist? IBM. http://www-01.ibm.com/software/data/infosphere/data-scientist/ (Accessed July 12, 2015).

19. Twitter Inc. 2015. About Twitter, Inc. Twitter, Inc. https://about.twitter.com/company (Accessed March 11, 2015).

20. Facebook. 2015. Facebook: Homepage. Facebook. https://www.facebook.com/?_rdr (Accessed March 9, 2015).

21. Instagram. 2015. Instagram. Instagram. https://instagram.com/# (Accessed March 9, 2015).

22. IBM. 2015. The Four V's of Big Data. IBM. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg (Accessed July 12, 2015).

23. University of Washington. The Science Behind FoldIt. University of Washington. http://fold.it/portal/info/about (Accessed March 9, 2015).

24. Google Inc. 2015. Google Trends. Google, Inc. https://www.google.com/trends/ (Accessed July 13, 2015).

25. Ginsberg J., M. H. Mohebbi, R. S. Patel *et al*. 2009. Detecting influenza epidemics using search engine query data. Nature. 457: 1012-1014.

26. Anyamba A., J. P. Chretien, J. Small *et al*. 2009. Prediction of a Rift Valley fever outbreak. Proc. Natl. Acad. Sci. U. S. A. 106: 955-959.

27. Koelle K., X. Rodó, M. Pascual *et al*. 2005. Refractory periods and climate forcing in cholera dynamics. Nature. 436: 696-700.

28. Rodó X., R. Curcoll, M. Robinson *et al*. 2014. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proc. Natl. Acad. Sci. U. S. A. 111: 7952-7957.

29. Massachusetts Water Resources Authority. 2015. pseudonitz_1992-2014.xlsx [MS Excel file].

30. Massachusetts Water Resources Authority. 2015. 1995-2015_F22_F23.xlsx [MS Excel file]. Excel file with requested MWRA data.

31. Northeastern Regional Association of Coastal and Ocean Observing Systems. 2014. NERACOOS: Data & Tools. The Gulf of Maine Research Institute. http://neracoos.org/datatools (Accessed May 25, 2015).

32. National Oceanic and Atmospheric Administration. 2015. National Data Buoy Center: Historical NDBC Data. U.S. Department of Commerce. http://www.ndbc.noaa.gov/historical_data.shtml (Accessed November 12, 2015).

33. National Oceanic and Atmospheric Information and National Centers for Environmental Information. 2015. National Centers for Environmental Information: Climate Data Online Search. U.S. Department of Commerce. http://www.ncdc.noaa.gov/cdo-web/confirmation (Accessed December 31, 2014).

34. Commonwealth of Massachusetts, Department of Public Health. 2014. Marine Beaches in Massachusetts, Annual Beach Reports. Commonwealth of Massachusetts. http://mass.digitalhealthdepartment.com/public_21/index.cfm (Accessed 01/24, 2015).

35. Trainer V. L. & M. Suddleson. 2005. Monitoring Approaches for Early Warning of Domoic Acid Events in Washington State. Oceanography. 18: 228-237.

36. Center for Environment, Fisheries, and Aquaculture Science. 2007. Marine Microbial Communities in UK Waters from Phylogenic Studies to Remote Studies.

37. Angus T. H. 2015. Examining the Toxicity, Exposure, and Regulatory Approach to Potential Human Health Risks of the Algal Toxin Domoic Acid. Doctor of Philosophy thesis, University of Massachusetts Boston, Boston, MA.

38. Robbins M. A., C. L. Ryan, A. L. Marriott *et al*. 2013. Temporal Memory Dysfunction and Alterations in Tyrosine Hydroxylase Immunoreactivity in Adult Rats Following Neonatal Exposure to Domoic Acid. Methods. 24: 25.

39. Tiedeken J. A. & J. S. Ramsdell. 2013. Persistent neurological damage associated with spontaneous recurrent seizures and atypical aggressive behavior of domoic acid epileptic disease. Toxicol. Sci. 133: 133-143.

40. Goldstein T., J. A. Mazet, T. S. Zabka *et al*. 2008. Novel symptomatology and changing epidemiology of domoic acid toxicosis in California sea lions (Zalophus californianus): an increasing risk to marine mammal health. Proc. Biol. Sci. 275: 267-276.

41. Kirkley K. S., J. E. Madl, C. Duncan *et al*. 2014. Domoic acid-induced seizures in California sea lions (Zalophus californianus) are associated with neuroinflammatory brain injury. Aquatic Toxicology. 156: 259-268.

The original R Studio software code for the 16 candidate set models for *P. delicatissima* complex.

```
Cand.set<- list()
Cand.set[[1]]<-glm(formula = binary0 ~ watertemp.44013 + sio4+ nh4 + no3 + po4 + zoo.ln + prcp.day.before + sal.station, family = binomial(logit), data = delicat.data)
Cand.set[[2]]<-glm(formula = binary0 ~ watertemp.44013 + prcp.5day.total + po4 +sal.station, family = binomial(logit), data = delicat.data)
Cand.set[[3]]<-glm(formula = binary0 ~ watertemp.44013 + po4 + no2+no3 + month, family = binomial(logit), data = delicat.data)
Cand.set[[4]]<-glm(formula = binary0 ~ watertemp.44013 + po4 +sal.station + sio4 + river.1wkavg + month, family = binomial(logit), data = delicat.data)
Cand.set[[5]]<-glm(formula = binary0 ~ watertemp.44013 + sio4 + chl.station, family = binomial(logit), data = delicat.data)
Cand.set[[6]]<-glm(formula = binary0 ~ sio4 + chl.station + river.30davg + no3, family = binomial(logit), data = delicat.data)
Cand.set[[7]]<-glm(formula = binary0 ~ no2+no3 + nn.p, family = binomial(logit), data = delicat.data)
Cand.set[[8]]<-glm(formula = binary0 ~ no2 + sio4 + prcp.day.before, family = binomial(logit), data = delicat.data)
Cand.set[[9]]<-glm(formula = binary0 ~ tdn + si.no3 + prcp.day.before, family = binomial(logit), data = delicat.data)
Cand.set[[10]]<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 + prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data = delicat.data)
Cand.set[[11]]<-glm(formula = binary0 ~ watertemp.44013 + sio4 + chl.station + no2+no3 + si.no3 + prcp.day.before + DON + nh4, family = binomial(logit), data = delicat.data)
Cand.set[[12]]<-glm(formula = binary0 ~ zoo.ln + sio4 + po4 + nh4 + prcp.day.before, family = binomial(logit), data = delicat.data)
Cand.set[[13]]<-glm(formula = binary0 ~ watertemp.44013 + sio4 + chl.station + nh4 + po4 + prcp.day.before, family = binomial(logit), data = delicat.data)
Cand.set[[14]]<-glm(formula = binary0 ~ watertemp.44013 + po4 + no2+no3 + month + latitude + longitude, family = binomial(logit), data = delicat.data)
Cand.set[[15]]<-glm(formula = binary0 ~ tdn + si.no3 + prcp.day.before + latitude + longitude, family = binomial(logit), data = delicat.data)
```

Cand.set[[16]]<-glm(formula = binary0 ~ latitude + longitude, family = binomial(logit), data = delicat.data)

R Studio output for AICc test of 16 candidate models for *P. delicatissima* complex

#Model selection based on AICc results

aictab(Cand.set, modnames = NULL, second.ord = TRUE, nobs = NULL, sort = TRUE)

Model selection based on AICc :

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Mod10	8	263.17	0	0.98	0.98	-123.2
Mod11	10	270.97	7.8	0.02	1	-124.89
Mod5	4	288.08	24.91	0	1	-139.94
Mod13	7	288.21	25.03	0	1	-136.82
Mod4	7	292.36	29.19	0	1	-138.9
Mod1	9	293.6	30.43	0	1	-137.34
Mod3	6	293.81	30.64	0	1	-140.7
Mod9	4	295.27	32.1	0	1	-143.54
Mod15	5	295.29	32.12	0	1	-142.51
Mod14	7	295.31	32.14	0	1	-140.38
Mod2	5	297.91	34.74	0	1	-143.81
Mod6	5	312.84	49.67	0	1	-151.28
Mod12	6	312.91	49.74	0	1	-150.26
Mod8	4	316.33	53.16	0	1	-154.07
Mod16	2	318.35	55.17	0	1	-157.15
Mod7	4	319.43	56.26	0	1	-155.62

The original R Studio software output with summary information for Model 10 of *P. delicatissima* complex presence/absence is shown below. Model 10 was the most supported model in the candidate set.

> mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +

prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =

delicat.data)

> summary(mod10)

Call:
glm(formula = binary0 ~ watertemp.44013 + tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit),
data = delicat.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7788	-1.1235	-0.3695	1.0860	1.9061

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.35E+02	1.10E+02	-2.129	0.03327	*
watertemp.44013	5.05E-02	3.02E-02	1.674	0.09404	.
tdn	-3.62E-02	2.18E-02	-1.66	0.09694	.
si.no3	-2.52E-02	7.81E-03	-3.222	0.00127	**
sio4	-8.95E-03	5.17E-02	-0.173	0.86251	
prcp.day.before	-3.30E-03	2.60E-03	-1.265	0.20575	
chl.station	6.71E-02	6.57E-02	1.02	0.30773	
latitude	5.55E+00	2.60E+00	2.136	0.03269	*
longitude	NA	NA	NA	NA	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 272.97 on 196 degrees of freedom
Residual deviance: 246.41 on 189 degrees of freedom
(32 observations deleted due to missingness)

AIC: 262.41
Number of Fisher Scoring iterations: 5

#R code for the Hosmer and Lemeshow goodness of fit test for Model 10 for *P. delicatissima* complex.

```
> hl<- hoslem.test(mod10$y, fitted(mod10), g=10)
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod10$y, fitted(mod10)
X-squared = 10.437, df = 8, p-value = 0.2357
```

Cross-validation for the *P. delicatissima* model involves leaving out one year of data, fitting the model on remaining years, then using that model generate predictive probabilities for the missing year.

 # Generate predictive probabilities for each year, write results to text files

```
load(" c:/Desktop/RStudio/2016-03-20 workspace.RData")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
delicat.data)

delicat_no95 <- read.csv(" c:/Desktop/delicatissima data by year/delicat_no95.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
delicat_no95)

mod10.predict <-predict(mod10, delicat_no95, type="response")

delicat_no95$predict.values<-predict(mod10, delicat_no95, type="response")

delicat95 <- read.csv(" c:/Desktop/delicatissima data by year/delicat95.csv")

delicat95$predict.values<-predict(mod10, delicat95, type="response")

delicat_no96 <- read.csv(" c:/Desktop/delicatissima data by year/delicat_no96.csv")

delicat96 <- read.csv(" c:/Desktop/delicatissima data by year/delicat96.csv")
```

```
mod10.predict <-predict(mod10, delicat_no96, type="response")

delicat96$predict.values<-predict(mod10, delicat96, type="response")

write.table(delicat96, " c:/Desktop/delicat96predict.txt", sep="\t")

delicat_no97 <- read.csv(" c:/Desktop/delicatissima data by year/delicat_no97.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
delicat_no97)

delicat_no97$predict.values<-predict(mod10, delicat_no97, type="response")

delicat97 <- read.csv(" c:/Desktop/delicatissima data by year/delicat97.csv")

delicat97$predict.values<-predict(mod10, delicat97, type="response")

write.table(delicat97, " c:/Desktop/delicat97predict.txt", sep="\t")

delicat_no98 <- read.csv(" c:/Desktop/delicatissima data by year/delicat_no98.csv")

delicat98 <- read.csv(" c:/Desktop/delicatissima data by year/delicat98.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
delicat_no98)

delicat98$predict.values<-predict(mod10, delicat98, type="response")

write.table(delicat98, " c:/Desktop/delicat98predict.txt", sep="\t")

delicat_no99 <- read.csv(" c:/Desktop/delicatissima data by year/delicat_no99.csv")

delicat99 <- read.csv(" c:/Desktop/delicatissima data by year/delicat99.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
delicat_no99)

delicat99$predict.values<-predict(mod10, delicat99, type="response")

write.table(delicat99, " c:/Desktop/delicat99predict.txt", sep="\t")

del_no2000 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2000.csv")

del_2000 <- read.csv(" c:/Desktop/delicatissima data by year/del_2000.csv")
```

```
mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2000)

del_2000$predict.values<-predict(mod10, del_2000, type="response")

write.table(del_2000, " c:/Desktop/delicat2000predict.txt", sep="\t")

del_no2001 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2001.csv")

del_2001 <- read.csv(" c:/Desktop/delicatissima data by year/del_2001.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2001)

del_2001$predict.values<-predict(mod10, del_2001, type="response")

write.table(del_2001, " c:/Desktop/delicat2001predict.txt", sep="\t")

del_no2002 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2002.csv")

del_2002 <- read.csv(" c:/Desktop/delicatissima data by year/del_2002.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2002)

del_2002$predict.values<-predict(mod10, del_2002, type="response")

write.table(del_2002, " c:/Desktop/delicat2002predict.txt", sep="\t")

del_no2003 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2003.csv")

del_2003 <- read.csv(" c:/Desktop/delicatissima data by year/del_2003.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2003)

del_2003$predict.values<-predict(mod10, del_2003, type="response")

write.table(del_2003, " c:/Desktop/delicat2003predict.txt", sep="\t")

del_no2004 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2004.csv")

del_2004 <- read.csv(" c:/Desktop/delicatissima data by year/del_2004.csv")
```

```
mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2004)

del_2004$predict.values<-predict(mod10, del_2004, type="response")

write.table(del_2004, " c:/Desktop/delicat2004predict.txt", sep="\t")

del_no2005 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2005.csv")

del_2005 <- read.csv(" c:/Desktop/delicatissima data by year/del_2005.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2005)

del_2005$predict.values<-predict(mod10, del_2005, type="response")

write.table(del_2005, " c:/Desktop/delicat2005predict.txt", sep="\t")

del_no2006 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2006.csv")

del_2006 <- read.csv(" c:/Desktop/delicatissima data by year/del_2006.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2006)

del_2006$predict.values<-predict(mod10, del_2006, type="response")

write.table(del_2006, " c:/Desktop/delicat2006predict.txt", sep="\t")

del_no2007 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2007.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2007)

del_2007 <- read.csv(" c:/Desktop/delicatissima data by year/del_2007.csv")

del_2007$predict.values<-predict(mod10, del_2007, type="response")

write.table(del_2007, " c:/Desktop/delicat2007predict.txt", sep="\t")

del_no2008 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2008.csv")

del_2008 <- read.csv(" c:/Desktop/delicatissima data by year/del_2008.csv")
```

```
mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2008)

del_2008$predict.values<-predict(mod10, del_2008, type="response")

write.table(del_2008, " c:/Desktop/delicat2008predict.txt", sep="\t")

del_no2009 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2009.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2009)

del_2009 <- read.csv(" c:/Desktop/delicatissima data by year/del_2009.csv")

del_2009$predict.values<-predict(mod10, del_2009, type="response")

write.table(del_2009, " c:/Desktop/delicat2009predict.txt", sep="\t")

del_no2010 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2010.csv")

del_2010 <- read.csv(" c:/Desktop/delicatissima data by year/del_2010.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2010)

del_2010$predict.values<-predict(mod10, del_2010, type="response")

write.table(del_2010, " c:/Desktop/delicat2010predict.txt", sep="\t")

del_no2011 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2011.csv")

del_2011 <- read.csv(" c:/Desktop/delicatissima data by year/del_2011.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2011)

del_2011$predict.values<-predict(mod10, del_2011, type="response")

write.table(del_2011, " c:/Desktop/delicat2011predict.txt", sep="\t")

del_no2012 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2012.csv")

del_2012 <- read.csv(" c:/Desktop/delicatissima data by year/del_2012.csv")
```

```
mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2012)

del_2012$predict.values<-predict(mod10, del_2012, type="response")

write.table(del_2012, " c:/Desktop/delicat2012predict.txt", sep="\t")

del_no2013 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2013.csv")

del_2013 <- read.csv(" c:/Desktop/delicatissima data by year/del_2013.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2013)

del_2013$predict.values<-predict(mod10, del_2013, type="response")

write.table(del_2013, " c:/Desktop/delicat2013predict.txt", sep="\t")

del_no2014 <- read.csv(" c:/Desktop/delicatissima data by year/del_no2014.csv")

del_2014 <- read.csv(" c:/Desktop/delicatissima data by year/del_2014.csv")

mod10<-glm(formula = binary0 ~ watertemp.44013 +tdn + si.no3 + sio4 +
prcp.day.before + chl.station + latitude + longitude, family = binomial(logit), data =
del_no2014)

del_2014$predict.values<-predict(mod10, del_2014, type="response")

write.table(del_2014, " c:/Desktop/delicat2014predict.txt", sep="\t")
```

#End of Delicatissima work.

This section of Appendix A includes the code relevant for *Enterococcus* predictive modeling.

```
  Cand.set<- list()
  Cand.set[[1]]<-glm(formula = entero.over10 ~ prcp.mblhd.day.before, family =
binomial(logit), data = entero_allyear)
  Cand.set[[2]]<-glm(formula = entero.over10 ~ human.pop.tract, family =
binomial(logit), data = entero_allyear)
  Cand.set[[3]]<-glm(formula = entero.over10 ~  dog.pop, family = binomial(logit), data =
entero_allyear)
```

```
 Cand.set[[4]]<-glm(formula = entero.over10 ~  turbidity.a01, family = binomial(logit),
data = entero_allyear)
 Cand.set[[5]]<-glm(formula = entero.over10 ~  watertemp.a01 , family =
binomial(logit), data = entero_allyear)
 Cand.set[[6]]<-glm(formula = entero.over10 ~ river.2wkavg, family = binomial(logit),
data = entero_allyear)
 Cand.set[[7]]<-glm(formula = entero.over10 ~ prcp.mblhd.2day , family =
binomial(logit), data = entero_allyear)
 Cand.set[[8]]<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + human.pop.tract + dog.pop , family =
binomial(logit), data = entero_allyear)
 Cand.set[[9]]<-glm(formula = entero.over10 ~ latitude + longitude , family =
binomial(logit), data = entero_allyear)
Cand.set[[10]]<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_allyear)

Cand.set[[11]]<-glm(formula = entero.over10 ~ tmax.marblehead + +prcp.mblhd.2day +
year + watertemp.a01 +river.1wkavg, family = binomial(logit), data = entero_allyear)

Cand.set[[12]]<-glm(formula = entero.over10 ~ latitude + year + watertemp.a01
+river.1wkavg, family = binomial(logit), data = entero_allyear)
```

#Identifying the most supported model for *Enterococcus* prediction using AIC

```
> aictab(Cand.set, modnames = NULL, second.ord = TRUE, nobs = NULL, sort =
TRUE)
```

Model selection based on AICc :

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Mod10	10	295.14	0	0.98	0.98	-137.24
Mod8	7	303.45	8.31	0.02	0.99	-144.56
Mod11	6	305.69	10.55	0.01	1	-146.72
Mod12	5	317.64	22.5	0	1	-153.73
Mod9	3	319.93	24.79	0	1	-156.93
Mod2	2	320.27	25.13	0	1	-158.12
Mod7	2	320.97	25.83	0	1	-158.47
Mod3	2	321.05	25.91	0	1	-158.51
Mod4	2	321.3	26.16	0	1	-158.63
Mod1	2	323.4	28.26	0	1	-159.68
Mod5	2	323.82	28.68	0	1	-159.89
Mod6	2	324.36	29.22	0	1	-160.16

The original R Studio output summary information for *Enterococcus* presence/absence Model 10 is shown below. Model 10 was the most supported model in the candidate set.

```
> mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude +
longitude +river.1wkavg, family = binomial(logit), data = entero_no2014)
> summary(mod10)

Call:
glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +
    prcp.mblhd.2day + year + watertemp.a01 + latitude + longitude +
    river.1wkavg, family = binomial(logit), data = entero_allyear)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.3840  -0.6159  -0.4452  -0.2978   2.4745

Coefficients:

                    Estimate      Std. Error    z value    Pr(>|z|)
(Intercept)        5.94E+03      2.77E+03        2.147     0.03177   *
chl.a01             4.09E-01      1.96E-01         2.09      0.0366   *
tmax.marblehead    -1.17E-02      4.67E-03       -2.496     0.01254   *
prcp.bos.day.before 4.09E-03      1.51E-03        2.702     0.00689   **
prcp.mblhd.2day    -2.28E-03      1.54E-03       -1.482     0.13826
year               -1.73E-01      7.59E-02       -2.284     0.02236   *
watertemp.a01       1.77E-01      8.42E-02        2.098     0.03591   *
latitude           -6.81E+01      3.27E+01       -2.087     0.03687   *
longitude           3.81E+01      1.93E+01        1.972     0.04856   *
river.1wkavg       -3.60E-05      3.40E-05       -1.057     0.29034
---
Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 317.27  on 340  degrees of freedom
```

Residual deviance: 274.47 on 331 degrees of freedom
(8 observations deleted due to missingness)
AIC: 294.47

Number of Fisher Scoring iterations: 5

#Code for *Enterococcus* model fitting with one year left out of model then predictive probability values generated for the left out year.
#

```
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2007)
summary(mod10)
mod10.predict <-predict(mod10, entero_no2007, type="response")
entero_no2007$predict.values<-predict(mod10, entero_no2007, type="response")
entero2007$predict.values<-predict(mod10, entero2007, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2008)
entero_no2008$predict.values<-predict(mod10, entero_no2008, type="response")
entero2008$predict.values<-predict(mod10, entero2008, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2009)
entero_no2009$predict.values<-predict(mod10, entero_no2009, type="response")
entero2009$predict.values<-predict(mod10, entero2009, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2010)
entero_no2010$predict.values<-predict(mod10, entero_no2010, type="response")
```

```
entero2010$predict.values<-predict(mod10, entero2010, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2011)
entero_no2011$predict.values<-predict(mod10, entero_no2011, type="response")
entero2011$predict.values<-predict(mod10, entero2011, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2012)
entero_no2012 <- read.csv(" c:/Desktop/entero_no2012.csv")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2012)
entero_no2012$predict.values<-predict(mod10, entero_no2012, type="response")
entero2012$predict.values<-predict(mod10, entero2012, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2013)
entero_no2013$predict.values<-predict(mod10, entero_no2013, type="response")
entero2013$predict.values<-predict(mod10, entero2013, type="response")
mod10<-glm(formula = entero.over10 ~ chl.a01 + tmax.marblehead +
prcp.bos.day.before +prcp.mblhd.2day + year + watertemp.a01+ latitude + longitude
+river.1wkavg, family = binomial(logit), data = entero_no2014)
entero_no2014$predict.values<-predict(mod10, entero_no2014, type="response")
entero2014$predict.values<-predict(mod10, entero2014, type="response")
```

```
#Code to save files with year-by-year predictive probability added
write.table(entero2007, "c:/Desktop/entero2007predict.txt", sep="\t")
write.table(entero2008, "c:/Desktop/entero2008predict.txt", sep="\t")
write.table(entero2009, "c:/Desktop/entero2009predict.txt", sep="\t")
write.table(entero2010, "c:/Desktop/entero2010predict.txt", sep="\t")
write.table(entero2011, "c:/Desktop/entero2011predict.txt", sep="\t")
write.table(entero2012, "c:/Desktop/entero2012predict.txt", sep="\t")
write.table(entero2013, "c:/Desktop/entero2013predict.txt", sep="\t")
write.table(entero2014, "c:/Desktop/entero2014predict.txt", sep="\t")
# End of Enterococcus work
```