

MODELING SPECIES GEOGRAPHIC DISTRIBUTIONS IN AQUATIC ECOSYSTEMS
USING A DENSITY-BASED CLUSTERING ALGORITHM

Mariana Castaneda-Guzman

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in
partial fulfillment of the requirements for the degree of:

Master of Science
In
Fisheries and Wildlife Sciences

Luis E. Escobar
Nicole Abaid
Emmanuel A. Frimpong

11 August 2022
Blacksburg, Virginia, United States

Keywords: Ecological niche modeling, *Vibrio cholerae*, climate change, infectious disease,
remote sensing, suitability, DBSCAN.

MODELING SPECIES GEOGRAPHIC DISTRIBUTIONS IN AQUATIC ECOSYSTEMS USING A DENSITY-BASED CLUSTERING ALGORITHM

Mariana Castaneda-Guzman

ABSTRACT

Distributional ecology is a branch of ecology which aims to reconstruct and predict the geographic range of free-living and symbiotic organisms in terrestrial and aquatic ecosystems. More recently, distributional ecology has been used to map disease transmission risk. The implementation of distributional ecology for disease transmission has, however, been erroneous in many cases. The inaccurate representation of disease distribution is detrimental to effective control and prevention. Furthermore, ecological niche modeling experiments are generally developed and tested using data from terrestrial organisms, neglecting aquatic organisms in case studies. Both disease and aquatic systems are often data limited, and current modeling methods are often insufficient. There is, therefore, a need to develop data-driven models that perform accurately even when only limited amounts of data are available or when there is little to no knowledge of the species' natural history to be modeled. Here, I propose a data-driven ecological niche modeling method that requires presence-only data (i.e., absence, pseudoabsence, or background records are not needed for model calibration). My method is expected to reconstruct environmental conditions where data-limited aquatic organisms are more likely to be present, based on a density-based clustering algorithm as a proxy of the realized niche (i.e., abiotic, and biotic environmental conditions occupied by the organism). Supported by ecological theories and methods, my central hypothesis is that because density-based clustering machine-learning modeling prevents extrapolation and interpolation, it can robustly reconstruct the realized niche of a data-limited aquatic organism. First, I assembled a comprehensive dataset of abiotic (temperature) and biotic (phytoplankton) environmental conditions and presence reports using *Vibrio cholerae*, a well-understood aquatic bacterium species in coastal waters globally (Chapter 2). Second, using *V. cholerae* as a model system, I developed detailed parameterizations of density-based clustering models to determine the parameter values with the best capacities to reconstruct and predict the species' distribution in global seawaters (Chapter 3). Finally, I compared the performance of density-based clustering modeling against traditional, correlative machine-learning ecological niche modeling methods (Chapter 4). Density-based clustering models, when assessed based on model fit and prediction, had comparable performance to traditional 'data-hungry' machine-learning correlative methods used in modern applications of ecological niche modeling. Modeling the environmental and geographic ranges of *V. cholerae*, an aquatic organism of free-living and parasitic ecologies, is a novel approach itself in distributional ecology. Ecological niche modeling applications to pathogens, such as *V. cholerae*, provide an opportunity to further the knowledge of directly-transmitted emerging diseases for which only limited data are available. Density-based clustering ecological niche modeling is termed here as *Marble*, honoring a previous, experimental version of this analytical approach, and is expected to provide new opportunities to understand how an ecological niche modeling method influences estimates of the distribution of data-limited organisms of complex ecology. These are lessons applicable to novel, rare, and cryptic aquatic organisms, such as emerging diseases, endangered fishes, and elusive aquatic species.

MODELING SPECIES GEOGRAPHIC DISTRIBUTIONS IN AQUATIC ECOSYSTEMS USING A DENSITY-BASED CLUSTERING ALGORITHM

Mariana Castaneda-Guzman

GENERAL AUDIENCE ABSTRACT

Distributional ecology is a branch of ecology which aims to reconstruct and predict the geographic distribution of land and water organisms. In the case of diseases, a correct representation of their geographic distributions is key for successful management. Previous studies highlight the need to develop new models that perform accurately even when limited amounts of data are available and there is little to no knowledge of the organisms' ecology. This thesis proposes a data-driven method, originally termed *Marble*. *Marble* is expected to help reconstruct environmental conditions where data-limited aquatic organisms are more likely to be found. Supported by ecological theories and methods, my hypothesis is that because *Marble* prevents under- and over-fitting, this method will produce results which better fit the data. Using *V. cholerae*, an aquatic organism, as a model system, I compared the performance of *Marble* against other traditional modeling algorithms. I found that *Marble*, in terms of model fit, performed similarly to traditional methods used in distributional ecology. Modeling the ecology of *V. cholerae* is a new approach in and of itself in ecological modeling. Furthermore, modeling pathogens provides an opportunity to further the knowledge of directly transmitted diseases, and *Marble* is expected to provide opportunities to understand how algorithm selection can reconstruct (or not) the distribution of data-limited aquatic organisms of diverse ecologies.

DEDICATION

First and foremost, I would like to dedicate this thesis to my parents and my sisters, who have been there every step of the way. I would also like to dedicate this thesis to my family back in Guatemala, to the amazing friendships I have made along the way, and to my therapist, for keeping me sane. Last, but not least, to Luis, who has shaped me into the researcher I am today.

ACKNOWLEDGEMENTS

I would like to devote all acknowledgment and gratitude toward my thesis advisor, Luis E. Escobar, for all his guidance and support. This thesis would have not been possible without his mentorship and continuous support in both my professional and personal development. I am very grateful for help from my committee members, Nicole Abaid and Dr. Emmanuel A. Frimpong. Their patience, insight, and guidance have been invaluable. I would also not have been able to accomplish many of my analyses without the collaboration of Huijie Qiao, whose aid added invariably to the rigor of my thesis. I gratefully recognize that this work would not have been possible without extensive help in data collection from Natalie Brown. I would also like to acknowledge the support of my lab mates Paige Van de Vuurst, Diego Tovar-Soler, and Steven N. Winter. I am very grateful for all my co-authors, Katherine EL Worsley-Tonks, Roman Biek, Meggan E Craft, Daniel G Streicker, Lauren A White, Nicholas M Fountain-Jones, Gabriel Mantilla-Saltos, Kris A Murray, Robert Settlage, Leticia del Carmen Castillo Signor, Thomas Edwards, Luis E Cuevas, Yolanda Mencos, Agnes Matope, Emily R Adams, that have shared their expertise and love for science. Thank you to all funding sources, including Virginia Tech Department of Fish and Wildlife Conservation, Department of Wildlife Resources, National Science Foundation, and support from the Virginia Sea Grant (VASG) Professional Development Award. I would also like to thank the people at Virginia Sea Grant, Sam Lake for all the help and support in my proposal preparations, and Michelle Rodriguez, Scott Sandridge, Troy Hartley, Madeleine Jepsen, for believing in me. I am also very grateful to Andrew Dolloff, Serena Ciparis, Megan Kirchgessner, Edward Fox, Ottoniel Monterroso, George G. Donohue, that supported me in countless proposal submissions. Although an inevitably incomplete list, I want to thank many friends in the FiW, LAIGSA, Salsa Tech who made Blacksburg feel like home. Special thanks to my roommates, Rachael Green and Amber Litterer.

ATTRIBUTUION

Committee chair Dr. Luis E. Escobar (LEE) and committee members, Dr. Nicole Abaid (NA) and Dr. Emmanuel A. Frimpong (EF), and each contributed significantly to this thesis. LEE acted as a project design supervisor and facilitator and contributed to the analytical framework of the study and the writing of the thesis chapters for publication. EF and NA provided crucial guidance on the research direction of this work and provided advice on the interpretation of the results. LEE, NA, EF, all served as editors and provided key guidance on the written presentation of this work. MCG the student, acted as the lead author and researcher for this thesis.

TABLE OF CONTENT

TITLE PAGE.....	i
ABSTRACT.....	ii
GENERAL AUDIENCE ABSTRACT.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENTS.....	v
ATTRIBUTION.....	vi
LIST OF FIGURES.....	ix
CHAPTER 1: INTRODUCTION.....	1
REFERENCES.....	11
CHAPTER 2:.....	18
ABSTRACT.....	18
INTRODUCTION.....	18
METHODS.....	23
RESULTS.....	28
DISCUSSION.....	28
REFERENCES.....	31
CHAPTER 3:.....	51
ABSTRACT.....	51
INTRODUCTION.....	52
METHODS.....	55
RESULTS.....	62
DISCUSSION.....	66
REFERENCES.....	69
CHAPTER 4:.....	83
ABSTRACT.....	83
INTRODUCTION.....	83
METHODS.....	84
RESULTS.....	86
DISCUSSION.....	87
REFERENCES.....	90
CHAPTER 5: CONCLUSIONS.....	99

REFERENCES.....	101
-----------------	-----

CHAPTER 1

LIST OF FIGURE AND TABLES

Figure 1. Example of Geographic (G) and Environmental (E) space. A) Geographic records of species occurrences (red circles; **G**-space). B) representation of environmental space (**E**-space) with a 2-dimensional plot of precipitation versus temperature. Blue circles represent all possible environmental combinations. Red circles represent the environmental combinations of the occurrence points. Figure from Escobar & Craft (2016).

Figure 2. Ecological Niche Model workflow diagram. Step 1, data preparation, requires two types of input data: species' occurrence records and environmental or biological predictors. Step 2: niche modeling, during this step the ecological niche model is selected, calibrated, and executed. To calibrate the model and select the parameter, a selected subset of occurrences records (i.e., calibration data) was used. In step 3, model projection and evaluation, the calibrated model was used to project into the study area and evaluated with the testing set (i.e., the unused subset of occurrences, those not included in the calibration set). Finally, step 4, model transferability, is the section to predict over a different geographic or temporal scale. Model workflow based on Peterson et al., 2011.

CHAPTER 2

LIST OF FIGURE AND TABLES

Table 1. Data specifications for MODIS remotely-sensed data. Original satellite-based imagery was collected by the MODIS instrument, part of the NASA Earth Observing System, and downloaded through the NASA'S ERDP server at a temporal resolution of monthly composite, from 2003 to 2020 and at a 4 km spatial resolution as NetCDF files.

Figure 1. Study area and geolocation of *Vibrio cholerae* in coastal areas. Exclusive economic zone around the world (grey) was used to limit the satellite-derived data of seawater conditions in coastal areas. Literature occurrence records of *Vibrio cholerae* (green points).

Figure 2. Database workflow diagram. (a) Remotely sensed data were downloaded from the NASA ERDDAP server in the form of NetCDF files. (b) Data were then transformed into a raster object. (c) Data were then cropped and masked to the exclusive economic zone and imported as GeoTIFF. (d) Data were analyzed to include statistical analyses and exported as raster files.

Figure 3. Data masking and cropping. Example of masking and cropping a raster. a) Raster from original NetCDF. b) Economic Exclusive Zone (solid lines). c) Raster after crop and mask.

Figure 4. Sea surface temperature correlation between MODIS and Sentinel-3 data during the year 2020. Correlation for mean, minimum, and maximum, between both sensors results in a high positive correlation with a Pearson correlation coefficient of $r > 0.96$ for all scenarios.

Figure 5. Chlorophyll-*a* correlation between MODIS and SeaWiFS data during the year 2010. Correlation for mean, minimum, and maximum, between both sensors results in a high positive correlation with a Pearson correlation coefficient of $r > 0.51$ for all scenarios.

Figure 6. Sea surface temperature mean monthly values from 2003–2020. (a) Temperate zone monthly averages between the years 2003–2020 (east coast of the United States). (b) Subtropical zone monthly averages between the years 2003–2020 (coast of Chile).

Figure 7. Chlorophyll-*a* mean monthly values from 2003–2020. (a) Tropical zone monthly averages between the years 2003–2020 (coast of Ecuador and Colombia). (b) Subtropical zone monthly averages between the years 2003–2020 (coast of Chile).

CHAPTER 3

LIST OF FIGURE AND TABLES

Figure 1. Clustering example. a) cluster in 2-D, called an area; b) cluster in 3-D, called a volume. Data points that fall closer to each other most of the time will be part of a cluster. Note that the clusters may vary in shape. When plotting environmental values in an XY or XYZ plot, this pertains to E- space (i.e., environmental space).

Figure 2. DBSCAN point labeling example. This figure shows one example of the labeling of points in DBSCAN. In this figure, only one core point (red) and one border point (green) are labeled as illustrative examples. Note brown points may also be core or border points. The radius of the neighborhood ϵ is denoted by the dotted lines. Outliers or noise points are shown in purple.

Figure 3. Basic concepts of *Marble* algorithm. a) shows examples of two classes of points: core (q) and border (p) points; b) illustrates the concept of direct density-reachability; (c) illustrates the concept of density-reachability; d) illustrates the concept of density-connectivity. Original figure from Qiao et al. (2015).

Figure 4. DBSCAN algorithmic steps. 1) arbitrary select a point p ; 2) retrieve all points density-reachable from p based on ϵ and $minPts$; 3) If p is a core point, a cluster is started; 4) each point p in the cluster will broadcast out their own perimeter looking to find new members to join the cluster. 5) If p is a border point, no points are density-reachable from p , and DBSCAN visits the next point of the database, and 6) continue the process until all the points have been processed.

Figure 5. Analytical framework of *Vibrio cholerae* bacteria ENM. a) *Vibrio cholerae* records from coastal seawaters during the last two decades were collected, curated, standardized, and coupled with satellite-derived coastal water conditions at 4 km resolution for the years from 2003 to 2020. b) Data integration accounting by locality and date of the *V. cholerae* record was developed using machine-learning clustering algorithms to build hypervolume models of *V. cholerae* environmental suitability for growth and reproduction.

Figure 6. Methods of the implementation *Marble* to reconstruct the realized niche of *Vibrio cholerae*. A) environmental and biological predictors were normalized and transformed using PCA. B) The selection of parameters to use in *Marble* was selected following literature recommendations for $minPts = 2 * dim$ (Sander et al., 1998, Ester et al., 1996), and ϵ (y-intercept of the red horizontal line) was calculated using the “elbow method” by calculating distances with KNN using an omission threshold of 5% (red vertical line). C and D illustrate *Marble* clusters in E-space c) in 2-d and d) in 3-d. E) show the resulting projection of *Marble* from E to G-space, retaining the clusters. F) resulting rasters after binarization (i.e., labeling as suitable and unsuitable, ignoring cluster number).

Figure 7. Epsilon (ϵ) distribution based on omission threshold and *minPts*. (A) The panel grid shows the different distribution of ϵ separated by sensitivity threshold. The x-axis represents the number of *minPts*. Variation in resulting ϵ increases as the sensitivity threshold (i.e., the confidence in the data; one minus omission threshold) increases. Noting, however, how the values follow a logistic curve, approaching some equilibrium value for ϵ as *minPts* are increased. (C) Illustrates the distribution on the resulting ϵ using KNN as *minPts* is vary across a range of sensitivity thresholds (i.e., one minus omission threshold). Colored lines show the likely distribution of values grouped by *minPts*. Similarly, (B) Boxplot illustrates the distribution of the resulting ϵ when all *minPts* values are grouped by a specific sensitivity threshold. Both figures indicate that as the confidence in the input data grows, values for ϵ will increase, as well as if *minPts* increases.

Figure 8. Sensitivity versus commission rate output from cross-validation using evaluation data. This figure illustrates a variation of ROC evaluation, where the x-axis corresponds to the proportion of area predicted as suitable instead of the commission rate (i.e., false positive rate). Although proportion of suitable predicted area has a 1:1 ratio when compared to commission rate (Peterson et al., 2008). The independent evaluation data for each cross-validation fold (i.e., five folds) grouped by *minPts* was used to develop the curves. (B) is the zoom-in of the figure in panel (A). In panel (B) each color represents a different *minPts* value. Lines show the distribution grouped by *minPts* of the original data points (colored dots). Each colored dot represents one of the 330 model replications from the parametrization matrix, colored by the *minPts* value. Overall, model outputs from all *minPts* show that they predict better than by random choice.

Figure 9. Evaluation of *Marble* in G-space based on selected user-defined sensitivity threshold. The panel figure illustrates the performance of *Marble* through different value parameters of *minPts* and sensitivity threshold (i.e., 1 – omission threshold). Panel (A) shows the variation of omission rate through different thresholds of sensitivity. Boxplots show a negative correlation between variable omission rate and sensitivity threshold. Panel (B) illustrates the change in the predicted area through different sensitivity thresholds grouped by values of *minPts*. Panel (C) illustrates the differences in time complexity grouped by *minPts*. Legend for both Panel B and C is at the top of the figure.

CHAPTER 4

LIST OF FIGURE AND TABLES

Figure 1. Presence points for *Quercus alba*. Figure illustrate the geolocation of the available presence records for *Quercus alba* (black crosses) commonly used for research and teaching purposes in ecological niche modeling and available in the *hypervolume* package in R. *Quercus alba* is mainly found the Southeast United States and a few observations have been recorded in the Canada territory in the western border with the United States. This is a good study model of a species in a terrestrial ecosystem and suitable to compare ecological niche modeling methods.

Figure 2. Continuous and binary models of *Quercus alba* in the study area (omission threshold=5%). (A) illustrates the continuous model projections for *Q. alba* potential distribution. Binary models (B) were generated based on a 5% omission threshold from calibration occurrences. BRT, boosted regression trees; GAM, generalized additive model; GLM, generalized linear models; MAXENT, and maximum entropy; and SVM, support vector machines.

Figure 3. Model performance evaluation in G-space., Boosted regression trees (BRT), generalized additive model (GAM), generalized linear models (GLM), and maximum entropy (MAXENT), support vector machines (SVM), and clustering DBSCAN (MARBLE) models plotted in terms of their omission error rate (i.e., how well it predicted the evaluation presence points) and sensitivity (i.e., 1-omission rate). Figure (A) colored shapes distinguish between methods and represent the mean values among the different omission threshold for binarization (i.e., 0%-10% omission threshold). Vertical black lines represent the standard error of commission error (y-axis) and standard error in proportion of area predicted (x-axis). Figure (B) is a modification of the ROC curve (Peterson et al., 2008), anything above the 1:1 black diagonal line is considered to predict better than by random expectations. All methods fall above the 1:1 line, with BRT and SVM being the models with the lowest predictive performance and Maxent and *Marble* being the best performing models balancing accurate prediction and reduced extrapolation.

Table 1. Model evaluation results based on omission thresholds. Table shows the results obtained from the transformed binary maps based on the corresponding sensitivity threshold or 1-omission threshold (first column). CBP, Cumulative binomial probability; BRT, boosted

regression trees; GAM, generalized additive model; GLM, generalized linear models; MAXENT, and maximum entropy; and SVM, support vector machines.

CHAPTER 1

INTRODUCTION

Distributional ecology is a branch of ecology which aims to reconstruct and predict the geographic ranges of free-living (e.g., plants) and parasitic (e.g., nematodes) organisms in terrestrial and aquatic ecosystems (Escobar & Morand, 2020; Peterson & Soberón, 2011). Recent applications of distributional ecology include mapping disease transmission risk (Peterson, 2014). Nevertheless, the inaccurate representation of disease distribution is detrimental to effective control and prevention (Feng et al., 2019; Qiao, Soberón, et al., 2015). There is, therefore, an urgent need for the development of computational methods to determine promptly, thoroughly, and precisely why a pathogen is present in a geographical area but absent in others. Computational advances facilitate immediate understanding of the origin, spread, and posterior establishment of a species, which will move spatial forecast forward (Escobar, 2020a; Peterson, 2014; Peterson & Soberón, 2011).

Many modeling techniques have been developed to estimate the distribution of species, including ecological niche modeling. Ecological niche modeling has three main objectives: reconstruct the spatial distribution of the species, estimate ranges of environmental tolerances, and predict future distributions in different areas or during different periods (Escobar, 2020a; Escobar & Craft, 2016). Traditional ecological niche modeling requires species presence and absence data for model calibration (Peterson & Soberón, 2011). Because absence data are generally unavailable or of arguable quality, researchers often generate simulated absence data to fulfill algorithms' requirements to determine suitable conditions of species (Escobar & Craft, 2016; Peterson & Soberón, 2011).

Modern ecological niche models require substantial amounts of species presence and absence data. When these data are limited, traditional ecological niche models, such as regression models,

tend to extrapolate (i.e., predict outside the range of values observed) and interpolate (i.e., predictions within the range of values observed) with different levels of magnitude that could be detrimental to the accuracy of predictions (Escobar & Craft, 2016). Thus, there is a need to develop data-driven ecological niche models that perform accurately even when only limited amounts of data are available (e.g., no absence data available) and there is little to no knowledge of the biology of the species (e.g., rare or poorly studied species). Data limitations for poorly understood aquatic organisms are common, for example, for water-borne pathogens with limited understanding of their biology or with limited surveillance in aquatic ecosystems (Melo-Merino et al., 2020). Here, I propose a data-driven ecological niche modeling method that requires presence-only data (i.e., absence, pseudoabsence, or background records are not needed for modeling calibration) (Qiao, Lin, et al., 2015). This thesis links ecological niche theory with machine-learning applications for a new proposal of an ecological niche modeling method to add to the toolkit of ecologists. Considering that ecological niche modeling methods have been largely biased toward free-living, terrestrial organisms (Elith et al., 2006), I assessed the global distribution of a pathogen in an aquatic ecosystem.

Supported by ecological theories and methods, my central hypothesis is that because density-based clustering machine-learning modeling prevents extrapolation and interpolation, this method will more robustly fit the data of organisms with limited records to predict the organism's actual distribution better than random (Qiao, Escobar, & Peterson, 2017; Qiao, Lin, et al., 2015). It is expected that this modeling framework, when assessed based on model fit and prediction, will perform as well as traditional 'data-hungry' machine-learning correlative methods used in modern ecological niche modeling (Franklin, 2010).

First, I procured, prepared, and processed remotely sensed imagery of oceanographic conditions. More specifically, monthly composites of remotely sensed imagery of sea surface temperature (SST) and Chlorophyll-*a* (Chlo-*a*) from 2003 to 2020. Each image was constrained to the exclusive economic zone (i.e., zone where the coastal nations have jurisdiction over natural resources; United Nations Convention on the Law of the Sea, 1982) and annual, monthly, and yearly statistics (i.e., minimum, maximum, mean, range, and standard deviation) were calculated. The resulting processed imagery was compiled and shared as a publicly accessible database (Chapter 2; Castaneda-Guzman et al., 2021). I, then, built upon a previously proposed method in ecological niche modeling using density-based clustering. I tested this modeling method through a parameterization sensitivity analysis to determine the best combination of parameter values to reconstruct the range of an aquatic organism (Chapter 3). Finally, I compared the proposed method against traditional ecological niche models, including Maximum Entropy models (Maxent), Generalized Linear Model (GLM), Random Forest (RF), Boosted Regression Trees (BRT) and Generalize Adaptive Models (GAM; Chapter 4). Finally, I synthesized major findings, their relevance, and future research directions (Chapter 5).

Ecological niche theory

Ecological niche modeling (ENM) is an analytical approach to understanding the geographic distribution of organisms and investigating relationships between organisms and their biotic and abiotic surrounding (Peterson et al., 2011). ENM is utilized to predict where, when, and in which conditions populations of a specific organism can persist (Escobar, 2020; Escobar & Morand, 2020; Peterson et al., 2011). ENM methods have evolved primarily due to the increased availability of large quantities of data, the development of theoretical frameworks to interpret models, and the

development of computational tools for more accurate analyses (Escobar, 2020b; Peterson et al., 2011). Computational ENM advances can be grouped based on their analytical principles into three main categories: correlative models, which are based on regression models between reports of the geographic distribution of species and environmental variables (Araújo & Guisan, 2006; Franklin, 2010); mechanistic models that use physiological information to fit mathematical models (Kearney & Porter, 2009); and process-oriented models, which estimate distributions of species in terms of dispersal capability of species and (biotic) interactions with other species (Peterson et al., 2015).

All classes of ecological niche models utilize species presence data. Presence data are point localities (usually defined with latitude and longitude) that specify what is known about the species' geographic distribution (Peterson & Soberón, 2012b). In most cases, presence data comprise records of where the species has been observed to be present (i.e., presence-only data). Nevertheless, in some rare cases, records of places where sampling has occurred, but the species has not been documented are also available. When both types of data are available, presence and absence data, the data are termed presence-absence data. The environmental datasets characterize variation in environmental variables across the study area. Environmental data are generally acquired from weather-stations interpolation (e.g., temperature, rainfall), on-ground surveys interpolation (e.g., soil characteristics), or, as used here, from remote sensing imagery that requires little to no interpolation (Jiménez & Soberón, 2021; Peterson & Soberón, 2011).

Differences among correlative ENM algorithms include (i) the type of geographic species-occurrence data that the method requires (i.e., presence-only, presence-absence, presence-pseudoabsence); (ii) the underlying statistical approach (e.g., regression methods, classification procedures, machine-learning procedures, Bayesian statistics); (iii) the format of output (e.g., model output in continuous, binary, or ordinal units); (iv) capacity to generate higher model

complexity versus relatively simpler models of species responses to particular environmental variables; and (v) ability to incorporate diverse formats of environmental variables (e.g., categorical, continuous, binary) (Peterson et al., 2011; Peterson & Soberón, 2012).

ENM to map species distributions relies on ecological niche theory. The concept of ecological niches surfaces from integrating two fields, historical and ecological biogeography (Peterson et al., 2011). Historical biogeography attempts to reconstruct the history of areas and their biotas at a coarse scale (i.e., continental and global extents). Ecological biogeography focuses on spatial patterns such as the composition and functioning of ecological communities at a fine scale (i.e., regional extent). By linking these two disciplines together, niche-theory pioneers like Grinnell, Hutchinson and Austin contributed to the understating of the distribution of species in terms of their ecological requirements across multiple scales of time and geography (Peterson et al., 2011; Peterson & Soberón, 2012). The work on Grinnell, Hutchison, and Austin became thriving research areas, including what came to be incorrectly termed “species distribution modeling” (SDM; Araújo & Guisan, 2006; Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000; Hirzel et al., 2002). Because species distribution could be mapped using a series of techniques, such as hand-made maps based on expert opinion, minimum convex polygons, kernel densities, and other approaches, the focus of thesis would be on the use of ENM to map species distributions accounting for species’ environmental requirements (Peterson et al., 2011).

Modeling spatial distributions is not the same as modeling environmental requirements. That is, spatial distribution models are modeled in geographic space (**G**-space; Figure 1a), while ecological niches are modeled in environmental space (**E**-space; Figure 1b). The term ecological “niche” was first used in 1917 by Joseph Grinnell to refer to the species’ climatic and environmental factors that restrict the distribution of a species (referred to in modern ecology as

the *Grinnellian niche*; Grinnell, 1917). Later in 1927, Charles S. Elton defined niche as the functional role of a species on a community (i.e., its local effect; *Eltonian niche*; Elton, 1927). The Grinnellian and Eltonian niches are rooted in different ideas, as one stresses environmental requirements and geographic dimensions, while the other ignores the environment and focuses on population parameters and behavior (Chase & Leibold, 2004; Escobar & Craft, 2016). Given this disparity, in 1957, G. Evelyn Hutchinson redefined “ecological niche” by adding the distinction between the fundamental and realized niche (*Hutchisonian niche*; Hutchinson, 1957). Hutchinson defined “niche” as a hypervolume of environmental variables, where the fundamental niche is the entire set of conditions under which an organism population can survive and reproduce indefinitely, and the realized niche is the set of conditions actually used by organism populations accounting for interactions with other species (e.g., predation and competition) have been considered (Hutchinson, 1957). More recently, Soberón & Peterson (2005) used the term ecological niche to refer to the environmental conditions in which a species can maintain a population in the long term without the need for immigration. Here, the definition of ecological niche is the combination of three factors: (1) the presence of environmental conditions under which a species can establish, survive, and reproduce; (2) the biotic environment determined by the presence of species interactions with a symbiont, prey, competitor or predator species that determine if a target species can persist in a determined locality, and (3) the area that is accessible to the species through movement or dispersal capacity (Peterson-Soberón Niche; Peterson et al., 2011; Soberón & Peterson, 2005).

In a classical article published in *Biodiversity Informatics* Soberón & Peterson (2005) noted the importance of including dispersal limitation and movement potential of species to reach suitable areas in the design and interpretation of ecological niche modeling studies. This led to the

formulation of the **BAM** framework: **B**, biotic; **A**, abiotic; **M**, movement (Escobar & Craft, 2016; Peterson et al., 2011) which is a central component in model design and interpretation (Barve et al., 2011; Saupe et al., 2012). In the **BAM** framework, **B**, relating mainly to biotic interactions, represents the geographic regions where the interactions with other organisms are favorable for the target species. **A**, relating mainly to abiotic interactions, represents regions in **G**-space where scenopoetic conditions, i.e., variables that do not interact with others and change very slowly, allow growth rates to be positive. Finally, **M**, relating to the movement of individuals of a species, corresponds to the geographic regions that have been accessible to the species within a given time span. The intersection of these three components would be considered the theoretical occupied distributional area of the species (Escobar & Craft, 2016; Peterson et al., 2011). Modern ecological niche modeling algorithms have been conceived to analyze abiotic variables (**A**; e.g., temperature), with little effort to develop analytical tools suitable to analyze both abiotic and biotic variables (**B**; Araújo & Rozenfeld, 2014; Simões & Peterson, 2018). Similarly, common correlative methods applied to ecological niche modeling generate disparate results under different assumptions of the **M** parameter (Barve et al., 2011).

Another important concept in niche theory is the ‘Hutchinson’s duality’, which describes the relationship between an ecological niche (i.e., environmental dimensions; **E**-space) and its geographic manifestation (i.e., geographic dimension; **G**-space) (R. K. Colwell & Rangel, 2009). Hutchinson duality matters in ENM because this is the basic property of geography that allows for the prediction of regions in geography based on environment combinations. Without Hutchinson duality, it would be impossible to predict distributions based on ecological niches or calculate ecological niches based on geographic distributions (Escobar et al., 2018). In general, by defining a **G**-space (composed of grid cells or pixels) at a given resolution, and at a particular point in time,

and by intersecting that geographic area with digital data layers of environmental data (e.g., climate, soil, topography), it is difficult to extract subsets of existing **E**-space. For each point in **G**, there is one and only one point in **E**-space, but for each point in **E**, there could be one or more points in **G**. Areas of distribution are subsets of **G**-space and ecological niches are subsets of **E**-space. If one characterizes a region in **E**, it is possible to map it to **G**-space, and *vice versa* (R. K. Colwell & Rangel, 2009; Escobar & Craft, 2016; Peterson et al., 2011)

Another key idea is that there is a function from **E**-space to fitness, where different population of a species along its ecological niche (**E**-space) would correspond to different fitness values. Hutchinson called the fundamental ecological niche the subset of all points in **E**-space where a species displays a positive fitness without interacting species. The realized ecological niche is then the subset of **E** with positive fitness accounting for biotic interaction. For Hutchinson, biotic interactions were negative interactions, but this postulate must be revisited. Hutchinson was well aware that environmental conditions inside the fundamental ecological niche of a given species may not be available in the study areas of periods of interest. Thus, the subset of the fundamental ecological niche that can be actually mapped due to the availability of conditions is termed the ‘existing niche’ (Peterson et al., 2011; Peterson & Soberón, 2012b).

A decade ago, Soberón and Nakamura (2009) coined the ‘Eltonian Noise Hypothesis’ concept, which proposed that the biotic interactions constitute a non-significant effect on the distributional potential of species at a coarse scale. This hypothesis, however, must be revisited for non-free-living organisms for which the presence of another species, e.g., host or vectors, is a limiting factor when modeling species’ geographic distributions. The Eltonian Noise Hypothesis also refers to biological processes and represents a simplification of modeling methodologies, considering the difficulties of including biotic interactions in the process (Araújo & Rozenfeld, 2014; Preuss &

Padial, 2021; Simões & Peterson, 2018). For some species, interactions with other organisms may not play a dominant role at the coarse resolution typically used for modeling, and ecological niche models for free-living organisms have demonstrated a good performance in describing species distribution based on **A** and **M** from the **BAM** diagram (Lira-Noriega et al., 2013).

According to Peterson et al. (2011) and Zurell et al. (2020), ENM can be developed in four stages: (a) data preparation, (b) model calibration, (c) model projection and evaluation, and (d) model transferability (also outlined in Figure 2). The first stage (a) in building an ecological niche model is to collect, collate, process, and standardize the species occurrence and environmental data necessary for model input (Chapter 2).

Furthermore, stage (b) in building an ecological niche model includes the implementation of a statistical algorithm to characterize the species responses as a function of environmental data (Chapter 3) (Zurell et al., 2020). In other words, algorithms identify associations between environmental conditions and species' occurrence. Depending on the algorithm employed, the values can represent probabilities, suitability indexes, or likelihoods (Jiménez & Soberón, 2021; Peterson & Soberón, 2012b; Soberón & Nakamura, 2009). The goal is to build the species' ecological niche in a multi-dimensional **E**-space, i.e., integrate many predictor variables since species are likely to respond to a combination of multiple environmental factors. At this stage of the modeling process, one task is to tune the algorithm to identify the parameters that provide optimal results, generally measured in terms of model fit to the calibration data (Escobar et al., 2018; Jiménez & Soberón, 2021; Qiao, Escobar, & Peterson, 2017). The **E**-space could consider abiotic or biotic variables, but the algorithm should be suitable to analyze such types of variables.

With that, the next stage (c) is to map (i.e., project) the prediction from **E**-space to **G**-space and to evaluate the extent to which the model correctly predicts independent evaluation data

(Escobar, 2020b; Peterson et al., 2011; Qiao, Soberón, et al., 2015). Evaluation data can be generated with techniques like data-splitting; evaluation metrics are discussed further in Chapter 4. Lastly, stage four (d) comes when models are used to measure the likelihood of a species occurring along new regions and time periods with regard to environmental suitability (Peterson et al., 2011; Zurell et al., 2020).

Ecological niche modeling in aquatic environments

Early implementations of ENM focused on mapping the distribution of organisms in terrestrial ecosystems, which has dominated the literature (L. M. Robinson et al., 2011). Their application to aquatic environments such as the ocean has been, however, less frequent, but it has been gaining some traction, especially in recent years (Melo-Merino et al., 2020; N. M. Robinson et al., 2017). Aquatic ENM may be conceptually and methodologically challenging, owing to the physical and biological characteristics of aquatic environments, species and data availability (Melo-Merino et al., 2020; L. M. Robinson et al., 2011). Thus, the first step for aquatic ENM is clearly understating seasonal and geographic changes in the aquatic environment (e.g., current speed, thermocline, water depth), which could be less pronounced or absent in terrestrial ecosystems.

Limited availability of environmental variables is a major limitation on comprehensive ENM applications to aquatic ecosystems. An array of oceanographic variables, such as SST and Chlo-*a*, have been crucial in biogeographic studies to reconstruct aquatic environmental phenomena, like the emergence of water-borne diseases, such as *V. cholerae* (Escobar et al., 2015a; E. K. Lipp et al., 2002; Vezzulli et al., 2016), algae blooms (Grimes et al., 2014; Shen et al., 2012a; Wei et al.,

2008), El Niño and La Niña dynamics (Hayashi et al., 2020), and coral bleaching (Hughes et al., 2018).

Efforts summarizing the state of the field of ENM in aquatic environments show that bacteria are the least studied group, with only two publications in the literature to date (2/328 articles reviewed, Melo-Merino et al., 2020). Melo-Merino et al. (2020) developed a review of ENM applications to marine ecosystems and found that many studies were based on correlative techniques to answer ecological or biographic questions about mechanisms underlying geographic ranges, climate change, and conservation strategies. From these correlative techniques, Maxent, a machine-learning approach, was most used, followed by statistical approaches such as GAMs and GLMs. Melo-Merino et al. determined that the groups most studied in aquatic environments were species with commercial importance (e.g., fish and mollusks), and species with conservation importance (e.g., mammals).

Ecological niche modeling for aquatic ecosystems is a developing field in its early stages. Modeling non-commercial aquatic organisms is the next frontier due to the limited data available, the scattered reports, and the urgent need for an understanding of their ecology (Melo-Merino et al., 2020). For instance, modeling bacteria of aquatic ecosystems, especially in coastal areas, may provide new insights regarding ENM performance for non-vertebrate aquatic organisms, which could help refine the conceptual and analytical frameworks of ENM (Escobar et al., 2015b)

Reference

- Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), 1677–1688. <https://doi.org/10.1111/j.1365-2699.2006.01584.x>
- Araújo, M. B., & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, 37(5), 406–415. <https://doi.org/10.1111/j.1600-0587.2013.00643.x>

- Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11), 1810–1819. <https://doi.org/10.1016/j.ecolmodel.2011.02.011>
- Castaneda-Guzman, M., Mantilla-Saltos, G., Murray, K. A., & Escobar, L. E. (2021). A database of global coastal conditions. *Scientific Data*, In review.
- Chase, J. M., & Leibold, M. A. (2004). Ecological niches: Linking classical and contemporary approaches. *Biodiversity and Conservation*, 13(9), 1791–1793. <https://doi.org/10.1023/B:BIOC.0000029366.24837.fc>
- Colwell, R. K., & Rangel, T. F. (2009). Hutchinson’s duality: The once and future niche. *Proceedings of the National Academy of Sciences*, 106(Supplement_2), 19651–19658. <https://doi.org/10.1073/pnas.0901650106>
- Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., ... E. Zimmermann, N. (2006). Novel methods improve prediction of species’ distributions from occurrence data. *Ecography*, 29(2), 129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Elton, C. S. (1927). *Animal ecology*. Sidgwick and Jackson.
- Escobar, L. E. (2020a). Ecological niche modeling: An introduction for veterinarians and epidemiologists. *Frontiers in Veterinary Science*, 7(October), 1–15. <https://doi.org/10.3389/fvets.2020.519059>
- Escobar, L. E. (2020b). Ecological niche modeling: An introduction for veterinarians and epidemiologists. *Frontiers in Veterinary Science*, 7. <https://doi.org/10.3389/fvets.2020.519059>
- Escobar, L. E., & Craft, M. E. (2016). Advances and limitations of disease biogeography using ecological niche modeling. *Frontiers in Microbiology*, 07. <https://doi.org/10.3389/fmicb.2016.01174>
- Escobar, L. E., & Morand, S. (2020). Disease ecology and biogeography. *Frontiers in Veterinary Science, Special Issue*. <https://www.frontiersin.org/research-topics/12035/disease-ecology-and-biogeography>
- Escobar, L. E., Qiao, H., Cabello, J., & Peterson, A. T. (2018). Ecological niche modeling re-examined: A case study with the Darwin’s fox. *Ecology and Evolution*, 8(10), 4757–4770. <https://doi.org/10.1002/ece3.4014>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015a). A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015b). A global map of suitability for coastal *Vibrio cholerae* under

- current and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution*, 3, 1382–1395. <https://doi.org/10.1038/s41559-019-0972-5>
- Franklin, J. (2010). *Mapping species distributions*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511810602>
- Grimes, J. D., Ford, T. E., Colwell, R. R., Baker-Austin, C., Martinez-Urtaza, J., Subramaniam, A., & Capone, D. G. (2014). Viewing marine bacteria, their activity and response to environmental drivers from orbit: satellite remote sensing of bacteria. *Microbial Ecology*, 67(3), 489–500. <https://doi.org/10.1007/s00248-013-0363-4>
- Grinnell, J. (1917). The niche-relationships of the california thrasher. *The Auk*, 34(4), 427–433. <https://doi.org/10.2307/4072271>
- Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9), 993–1009. <https://doi.org/10.1111/j.1461-0248.2005.00792.x>
- Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2–3), 147–186. [https://doi.org/10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9)
- Hayashi, M., Jin, F., & Stuecker, M. F. (2020). Dynamics for El Niño-La Niña asymmetry constrain Equatorial-Pacific warming pattern. *Nature Communications*, 11(1), 1–10. <https://doi.org/10.1038/s41467-020-17983-y>
- Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7), 2027. <https://doi.org/10.2307/3071784>
- Hughes, T. P., Anderson, K. D., Connolly, S. R., Heron, S. F., Kerry, J. T., Lough, J. M., Baird, A. H., Baum, J. K., Berumen, M. L., Bridge, T. C., Claar, D. C., Eakin, C. M., Gilmour, J. P., Graham, N. A. J., Harrison, H., Hobbs, J. A., Hoey, A. S., Hoogenboom, M., Lowe, R. J., ... Schoepf, V. (2018). *Spatial and temporal patterns of mass bleaching of corals in the Anthropocene*. 359(6371), 80–83. <https://doi.org/10.1126/science.aan8048>
- Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0), 415–427. <https://doi.org/10.1101/SQB.1957.022.01.039>
- Jiménez, L., & Soberón, J. (2021). *Estimating the fundamental niche: accounting for the uneven availability of existing climates*.
- Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12(4), 334–350. <https://doi.org/10.1111/j.1461-0248.2008.01277.x>
- Lipp, E. K., Huq, A., & Colwell, R. R. (2002). Effects of global climate on infectious disease: the cholera model. *Clinical Microbiology Reviews*, 15(4), 757–770. <https://doi.org/10.1128/CMR.15.4.757-770.2002>

- Lira-Noriega, A., Soberón, J., & Miller, C. P. (2013). Process-based and correlative modeling of desert mistletoe distribution: a multiscale approach. *Ecosphere*, 4(8), art99. <https://doi.org/10.1890/ES13-00155.1>
- Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, 415, 108837. <https://doi.org/10.1016/j.ecolmodel.2019.108837>
- Peterson, A. T. (2014). *Mapping disease transmission risk enriching models using biogeography and ecology*. Johns Hopkins University Press.
- Peterson, A. T., Papeş, M., & Soberón, J. (2015). Mechanistic and correlative models of ecological niches. *European Journal of Ecology*, 1(2), 28–38. <https://doi.org/10.1515/eje-2015-0014>
- Peterson, A. T., & Soberón, J. (2012a). Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza a Conservacao*, 10(2), 102–107. <https://doi.org/10.4322/natcon.2012.019>
- Peterson, A. T., & Soberón, J. (2012b). Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação*, 10(2), 102–107.
- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press. <https://doi.org/10.1515/9781400840670>
- Preuss, G., & Padial, A. A. (2021). Increasing reality of species distribution models of consumers by including its food resources. *Neotropical Biology and Conservation*, 16(3), 411–425. <https://doi.org/10.3897/neotropical.16.e64892>
- Qiao, H., Escobar, L. E., & Peterson, A. T. (2017). Accessible areas in ecological niche comparisons of invasive species: Recognized but still overlooked. *Scientific Reports*, 7(1), 1213. <https://doi.org/10.1038/s41598-017-01313-2>
- Qiao, H., Lin, C., Jiang, Z., & Ji, L. (2015). Marble algorithm: A solution to estimating ecological niches from presence-only records. *Scientific Reports*, 5(1), 14232. <https://doi.org/10.1038/srep14232>
- Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, 6(10), 1126–1136. <https://doi.org/10.1111/2041-210X.12397>
- Robinson, L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., Possingham, H. P., & Richardson, A. J. (2011). Pushing the limits in marine species distribution modelling: Lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6), 789–802. <https://doi.org/10.1111/j.1466-8238.2010.00636.x>
- Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E., & Lundquist, C. J. (2017). A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, 4. <https://doi.org/10.3389/fmars.2017.00421>

- Saupe, E. E., Barve, V., Myers, C. E., Soberón, J., Barve, N., Hensz, C. M., Peterson, A. T., Owens, H. L., & Lira-Noriega, A. (2012). Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling*, 237–238, 11–22. <https://doi.org/10.1016/j.ecolmodel.2012.04.001>
- Shen, L., Xu, H., & Guo, X. (2012). Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework. *Sensors*, 12(6), 7778–7803. <https://doi.org/10.3390/s120607778>
- Simões, M. V. P., & Peterson, A. T. (2018). Importance of biotic predictors in estimation of potential invasive areas: The example of the tortoise beetle *Eurypedus nigrosignatus*, in Hispaniola. *PeerJ*, 2018(12). <https://doi.org/10.7717/peerj.6052>
- Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 106(Supplement_2), 19644–19650. <https://doi.org/10.1073/pnas.0901637106>
- Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2(0). <https://doi.org/10.17161/bi.v2i0.4>
- United Nations Convention on the Law of the Sea*, 1833 U.N.T.S. 397 (1982) (testimony of United Nations). https://www.un.org/depts/los/convention_agreements/convention_overview_convention.htm
- Vezzulli, L., Grande, C., Reid, P. C., Hélaouët, P., Edwards, M., Höfle, M. G., Brettar, I., Colwell, R. R., & Pruzzo, C. (2016). Climate influence on *Vibrio* and associated human diseases during the past half-century in the coastal North Atlantic. *Proceedings of the National Academy of Sciences*, 113(34), E5062–E5071. <https://doi.org/10.1073/pnas.1609157113>
- Wei, G. F., Tang, D. L., & Wang, S. (2008). Distribution of chlorophyll and harmful algal blooms (HABs): A review on space based studies in the coastal environments of Chinese marginal seas. *Advances in Space Research*, 41(1), 12–19. <https://doi.org/10.1016/j.asr.2007.01.037>
- Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., ... Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, 43(9), 1261–1277. <https://doi.org/10.1111/ecog.04960>

FIGURE AND TABLES

CHAPTER 1

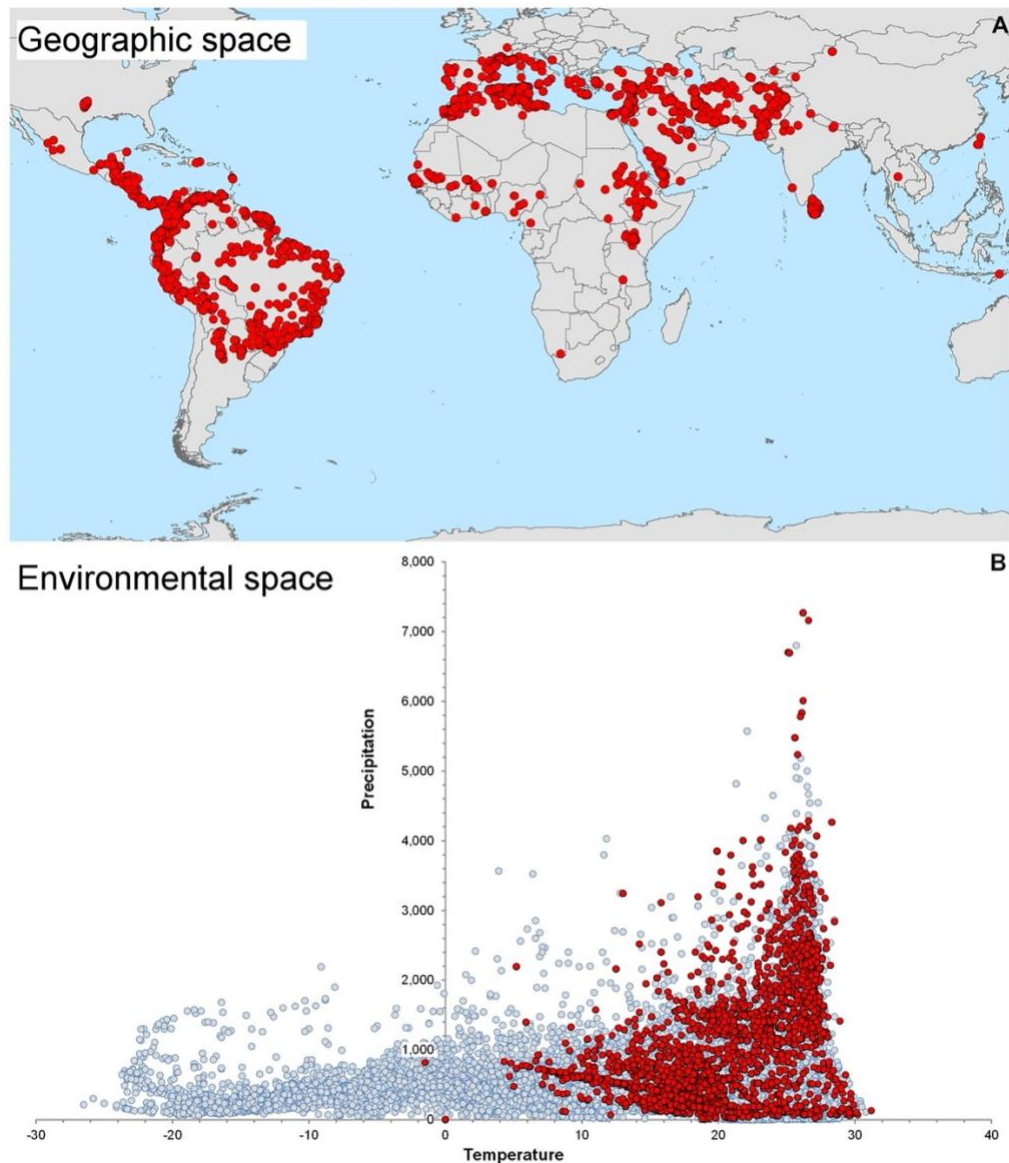


Figure 1. Example of Geographic (G) and Environmental (E) space. A) Geographic records of species occurrences (red circles; G-space). B) representation of environmental space (E-space) with a 2-dimensional plot of precipitation versus temperature. Blue circles represent all possible environmental combinations. Red circles represent the environmental combinations of the occurrence points. Note the clustering of the occurrence points in E-space, despite the broad geographic spread in G-space (Hutchinson's duality). Figure from Escobar & Craft (2016).

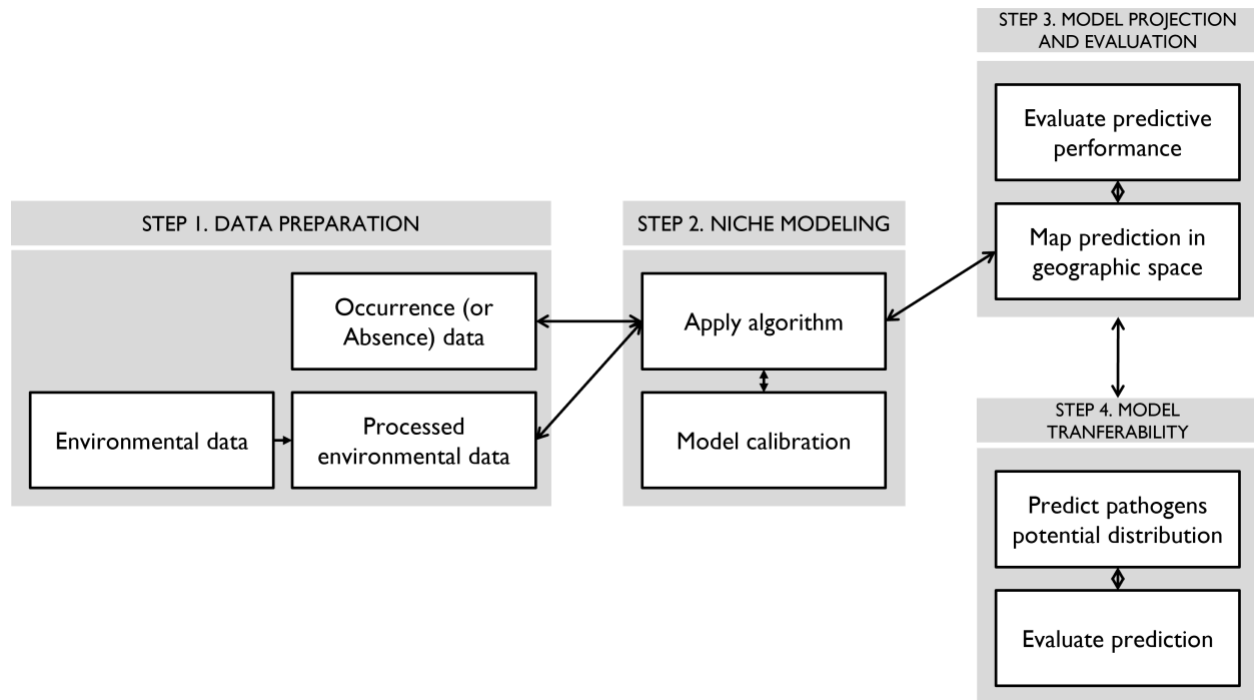


Figure 2. Ecological Niche Model workflow diagram. Step 1, data preparation, requires two types of input data: species' occurrence records and environmental or biological predictors. Step 2: niche modeling, during this step the ecological niche model is selected, calibrated and executed. To calibrate the model and select the parameter, a subset of occurrences records (i.e., calibration data) was selected. In step 3, model projection and evaluation, I used the calibrated model to project into the study area and evaluated with the testing set (i.e., the unused subset of occurrences, those not included in the calibration set). Finally, step 4, model transferability, is the section to predict over a different geographic or temporal scale. Model workflow based on Peterson et al., 2011.

CHAPTER 2

A DATABASE OF VIBRIO AND GLOBAL COSTAL CONDITIONS

Published as: Castaneda-Guzman, M., Mantilla-Saltos, G., Murray, K.A., Settlage, R., Escobar, L.E. A database of global coastal conditions. *Sci Data* 8, 304 (2021).

Abstract

Remotely sensed data, coupled with field data and ecological modeling has the potential for predicting bacterial response to environmental changes. Aquatic ecosystems have been studied with less intensity than terrestrial ecosystems due, in part, to data limitations. Data on sea surface temperature (SST) and Chlorophyll-*a* (Chlo-*a*) can provide quantitative information of environmental conditions in coastal regions at high spatial and temporal resolutions. Chlo-*a* and SST have also been consistently correlated with harmful algal blooms and *Vibrio cholerae* occurrence in seawaters around the world, which makes them ideal for ecological modeling. Using the exclusive economic zone of coastal regions as the study area monthly and annual statistics of SST and Chlo-*a* globally for 2003 to 2020 and *Vibrio cholerae* occurrence data from the literature were compiled. Data may be of interest to researchers in the areas of ecology, oceanography, biogeography, fisheries, and global change. Target applications of the database include environmental monitoring of biodiversity and marine microorganisms, and environmental anomalies.

Introduction

Commonly known effects of climate change on coastal areas include sea-level rise, coral bleaching, and coastal flooding. Climate change linked to rising sea temperatures can also impact

coastal ecosystem health. Nevertheless, coastal ecosystem health remains largely neglected in climate change assessments (Laffoley & Baxter, 2016). The deficit of targeted analysis on the effect of climate change on coastal ecosystem health, which is linked to fisheries and aquaculture, can impact food security and local economies (Barange et al., 2018).

Changes in seawater temperature increase the occurrence of water-borne diseases, such as Vibriosis, which pose a threat to human health (Escobar et al., 2015c; Shen et al., 2012b). This is the case of Cholera, a water-borne disease caused by *Vibrio cholerae* bacteria, which causes severe disease and even mortality. Cholera lends itself to the study of the role of climate in infectious disease. This disease has a historical context linking it to specific seasons and biogeographical zones (R. R. Colwell, 1996; Escobar et al., 2015a; E. K. Lipp et al., 2002). *Vibrio spp.*, including *V. cholerae*, can be found in virtually any coastal water body, especially in the tropics and subtropics (World Health Organization, 2013). *Vibrio spp.* form associations with phytoplankton (Main et al., 2015) and increase occurrence in warmer waters (E. K. Lipp et al., 2002). In a changing climate, the geographic range of these pathogens may also change, potentially resulting in increased exposure and risk of infection for humans. The risk of infection by *Vibrio* bacteria increases by ingesting raw or undercooked shellfish. Ingesting infected shellfish, beyond digestive problems, can also cause skin infections (e.g., *V. vulnificus*), which in extreme cases can require amputation (Center for Disease Control and Prevention (CDC), 2019). Given the close connection among sea temperature, phytoplankton, and *Vibrio* bacteria, the study of cholera offers an excellent example of the effect of climate on infectious diseases and pathogens (E. K. Lipp et al., 2002). Unfortunately, the assessment of coastal ecosystem health across large areas, can prove to be highly challenging and costly (Melo-Merino et al., 2020; N. M. Robinson et al., 2017).

Recently, ecologically based models have been developed which define the role of environmentally, weather, and climate related variables in outbreaks of cholera. Understanding the factors that influence the emergence of past cholera epidemics may help inform monitoring and prevention of future outbreaks. Using environmental data from satellite imagery, ENM can be employed to predict global *V. cholerae* presence and identify high-risk areas under current and future climate scenarios (R. R. Colwell, 1996; E. K. Lipp et al., 2002).

Remote sensing, referring to the acquisition of information about the Earth's surface through satellite imagery, has become a powerful tool for monitoring the environment and predicting risks associated with environmental changes (Carter & Paulson, 1979; Horning et al., 2010; Li et al., 2020; Nurdin et al., 2013), e.g., predicting human health risks associated with microscopic organisms (Grimes et al., 2014). In a plethora of applications, remotely sensed data have been used to detect landscape change (Dewan & Yamaguchi, 2009; K. Green et al., 1994; Singh, 1989; D. Ward et al., 2000), assess biodiversity (Nagendra, 2001), monitor carbon emissions (Liu, 1997; Rosenqvist et al., 2003), predict infectious diseases (R. R. Colwell, 1996; Escobar et al., 2015a; Watts et al., 2019), and track marine coasts (Alesheikh et al., 2007). Marine ecosystems, however, have been studied with less intensity than terrestrial ecosystems due, in part, to data limitations. Remote sensing can help detect environmental signals across space and time, including effects of sea surface temperature (SST), precipitation change, floods, harmful algal blooms, and environmental signatures associated with *Vibrio* emergence in coastal areas (Horning et al., 2010; Schowengerdt, 2012). Remote sensing, coupled with field data and ecological modeling, has the potential for predicting bacterial response to environmental change (Horning et al., 2010). A limitation in the use of global-level remotely sensed data is how time-consuming it proves to be,

given that sometimes complex data compilation, curation, standardization, and storage may require high-performance computational facilities (E. P. Green et al., 1996; Specter & Gayle, 1990).

Nevertheless, a significant benefit of satellite-derived information is the historical archives of data (R. R. Colwell, 1996; Li et al., 2020; Rosenqvist et al., 2003). Technological advances and innovative design have resulted in new generations of satellite sensors that monitor marine environments, such as the Moderate Resolution Imaging Spectroradiometer (MODIS). MODIS sensors are part of the National Aeronautics and Space Administration's Earth Observing System onboard the Terra and Aqua satellites and were designed to provide measurements of global dynamics of terrestrial, freshwater, and marine ecosystems (Esaias et al., 1998; Kilpatrick et al., 2015; NASA, 2021). MODIS provides the longest standing observational time series, given that both the Aqua and Terra satellites have been in orbit since the early 2000s, and it provides a larger set of marine variables for potential evaluation at the same spatial and temporal scale (NASA, 2021). Nevertheless, there are other enhanced satellite instruments (C. J. Donlon et al., 2002), such as the Along Track Scanning Radiometer (Minnett, 1990), Suomi National Polar-orbiting Partnership (Hillger et al., 2013), Visible Infrared Imaging Radiometer Suite (Minnett et al., 2014; O'Brien, 2020), and Sentinel (Drusch et al., 2012), which offer opportunities for future multi-sensor marine variables.

Out of the possible marine variables derived from observations of MODIS, sea surface temperature (SST) and Chlorophyll-*a* (Chlo-*a*) have the potential to increase the understanding of abiotic (e.g., temperature) and biotic (e.g., primary productivity) ocean conditions (Esaias et al., 1998; Nurdin et al., 2013). SST measured by MODIS infrared radiometers is also referred to as the skin temperature of the ocean. This is because the radiance measured by infrared radiometers originates in the surface thermal skin layer of the ocean and not the water below as measured by

in situ thermometers (C. Donlon et al., 2007). SST provides fundamental information on the global climate systems, and it is an essential parameter in weather prediction (NOAA, 2020). Chlo-*a* is a proxy for understanding fluctuations in algae and pigmented bacteria as it can elucidate photosynthetic activity in coastal systems (Esaias et al., 1998; Nurdin et al., 2013; Wei et al., 2008). The near-surface concentration of Chlo-*a* is calculated using an empirical relationship derived from *in situ* measurements, and the implementation of the standard O'Reilly band ratio OC_x (e.g., OC_{3M}, for the MODIS sensor) algorithm merged with the color index algorithm of Hu et al. (Hu et al., 2012a; O'Reilly et al., 1998b). SST and Chlo-*a* have been crucial in studies to reconstruct environmental phenomena, such as *Vibrio cholerae* emergence, algae blooms (Grimes et al., 2014; Shen et al., 2012a; Wei et al., 2008), El Niño and La Niña dynamics (Hayashi et al., 2020), and coral bleaching. There is a strong link between coastal ecosystem change with harmful algal and *Vibrio cholerae* outbreaks. Thus, ecosystem health studies have been used to anticipate algal bloom and cholera in coastal areas. Preliminary work by Dr. Luis Escobar at the Department of Fish and Wildlife Conservation, Virginia Tech, has revealed signals of *Vibrio cholerae* emergence in the coastal areas of the Mid-Atlantic region, including Virginia, in the past (Murray et al., 2020; Watts et al., 2019) and future (Escobar et al., 2015a) decades, suggesting a potential risk for public health. Assessing rapid changes in ecosystem health due to climatic variation may facilitate early and effective mitigation and adaptation plans.

Satellite-derived data have many limitations given their sensitivity to absorption of solar isolation, heat exchange with the atmosphere, and sub-surface turbulence. Nevertheless, since these conditions are known and common, validation and uncertainty are estimated relative to *in situ* buoys to correct final datasets (Minnett, 2010; Minnett et al., 2004; Minnett & Corlett, 2012). Satellite-derived data provide an opportunity to analyze large study areas during extended periods,

at the cost of limiting the information to surface level. Complementary approaches may include the addition of more oceanic and atmospheric observations like bathymetry, wind direction, and wind speed (Horning et al., 2010). I compiled remotely sensed data of monthly SST and Chlo-*a* from the exclusive economic zone (EEZ) of coastal areas globally for a 18-year period (2003-2020) and generate summary statistics at yearly and monthly composites. I compiled a dataset of occurrence data of *Vibrio cholerae* from the literature.

Methods

This section describes the procedures used to generate the individual data records that will comprise the SST and Chlo-*a* and the *Vibrio cholerae* databases. Data retrieval and analysis that will be performed during the development of the database will be executed using the statistical software R (R Core Team, 2022) and from the literature, respectively.

First, I assembled a comprehensive dataset of *V. cholerae* occurrence data for the last two decades. Each *V. cholerae* record will be carefully curated following standardized data-cleaning protocols to reduce bias and errors (Cobos et al., 2018; Escobar et al., 2017a) and records will be linked to seawater data of the site and date of *V. cholerae* sample collection. Records to quantify the environmental tolerances of the species will be recovered from about 50 journal articles (Atlantic et al., 2010; Barbera et al., 2004; Batabyal et al., 2016; Binsztein et al., 2004; Bliem et al., 2018; Dalusi et al., 2015; de Menezes et al., 2014; De Menezes et al., 2017; Dheenana et al., 2014; Di et al., 2017; Escobar et al., 2015b; Esteves et al., 2015b, 2015a; L. Fang et al., 2019; Fernández-Delgado et al., 2017; Fri et al., 2017; Gardade & Khandeparker, 2017; Gdoura et al., 2016; Grothen et al., 2017; Gyraite et al., 2019; Hackbusch et al., 2020; Izumiya et al., 2017; Khamesipour et al., 2014; Kim & Lee, 2014; Kokashvili et al., 2015; E. K. E. Lipp et al., 2003; López et al., 2010;

Louis et al., 2003; Machado & Bordalo, 2016; Main et al., 2015; Matteucci et al., 2015; Meena et al., 2019; Meyer et al., 2016; Ming et al., 2020; Mukhopadhyay et al., 1998; Neogi et al., 2018; Orozco et al., 1996; Pal et al., 2006; Pascual et al., 2000; Perkins et al., 2014; Sack et al., 2003; Siboni et al., 2016; Silva et al., 2019; Sneha et al., 2016; Sulca et al., 2018; Y. Y. Wong et al., 2019; Xu et al., 2015; Yue et al., 2014; Zaw et al., 2019; Zhang et al., 2015). *Vibrio cholerae* occurrence data will include reports from coastal regions, corroborated in laboratory facilities, and geolocated following the protocol described by Escobar et al., (2015b) . To account for geolocation uncertainty and water displacement, I will collect seawater conditions in immediate pixels neighboring each *V. cholerae*. All cells with duplicate environmental values will be removed to reduce bias (Figure 1)

The SST and Chlo-a databases will be developed in four stages: (a) data procurement, (b) preparation, (c) processing, and (d) analysis. The first two stages will be associated with input data, while the third stage will be applied specific methods to construct the core of each database. The fourth stage will include the statistical analyses of the data. The methodological stages are summarized in Figure 2 and described in detail below.

Data Procurement.

The database is based on satellite observations derived from the MODIS satellite. The Terra and Aqua satellites have been orbiting around the Earth since their launch in 1999 and 2002, respectively, obtaining data of Earth's surface every one to two days at three spatial resolutions (250, 500, 1000 m) and 36 spectral bands (from 0.405 to 14.385 μm). From the available atmospheric and oceanic observations made available from NASA's Aqua Spacecraft, Sea Surface

Temperature (SST) in °C and Chlorophyll-*a* (Chlo-*a*) in mg·m⁻³ were selected since they summarize major physical and biological phenomena. SST and Chlo-*a* are available at a temporal resolution of 1-day, 8-day, and monthly composites and a spatial resolution of ~4 km (**Error! Not a valid bookmark self-reference.**).

SST and Chlo-*a*, among other environmental variables, can be accessed through National Oceanic and Atmospheric Administration's (NOAA) Coastal Watch Environmental Research Division (ERD) Environmental Research Division Data Access Protocol (ERDDAP) data server, also known as the NOAA's Coastal Watch. NOAA's Coastal Watch is a program that provides timely access to near-real-time satellite data to monitor, restore, and manage coastal ocean resources, and the ERDDAP Data Server supports manual downloads through a web application and remote downloads from any computer program (e.g., MATLAB, R, JSONP, Python) of both gridded and tabular data (NOAA, 2021).

Data Downloading.

The remote request to the ERDDAP Data Server relies on the creation of specially formed URLs to query the server for a specific database. A URL consists of a root, a target, and a constraint expression (NOAA, 2021). To procure the inputs needed to assemble this database especially formed URLs will be created through a programming algorithm in R.

The root or base URLs that provided the location of the gridded database were obtained from the ERDDAP griddap documentation webpage (<https://coastwatch.pfeg.noaa.gov/erddap/griddap/documentation.html>) and remained constant in all requests for a specific database.

The target is the equivalent to the unique identifier or data set ID previously assigned by the ERDDAP (<https://coastwatch.pfeg.noaa.gov/erddap/griddap>), in conjunction with a specific data file type extension, for this study *.nc* was selected producing NetCDF-3 binary files with COARDS/CF/ACDD metadata. NetCDF, Network Common Data Form, files are recommended when using software tools to analyze geospatial data as they provide multi-dimensional scientific data in a standardized manner (<https://coastwatch.pfeg.noaa.gov/erddap/griddap/documentation.html>)(Stanford, 2021; UCAR Community Programs, 2021).

The constraint expression (or query) help define the parameters, which correspond to the study period and spatial coverage. Regarding the first parameter, the study period comprises all available observations from the MODIS instrument aboard the Aqua satellite (i.e., monthly composites from 2003 to 2019). The spatial coverage will be defined by the minimum and maximum latitude (i.e., 89.98°S to 89.98°N) and longitude (i.e., 179.98°W to 179.98°E) from the original satellite image for global coverage.

Data Preparation

Data within the NetCDF files will be imported into R using the *RNetCDF* package(Michna & Woods, 2019). A NetCDF object contains a list of at least four attributes: time, longitude, latitude, and the values of the variable being measured (i.e., SST and Chlo-*a*). The attribute corresponding to the specific variable being measured will be extracted from the NetCDF object and transformed into a raster object using the *RNetCDF* and *raster* packages in R(Hijmans, 2020). A raster object

consists of a matrix of cells (i.e., pixels) organized into rows and columns where each cell contains a value representing information (i.e., temperature and pigmentation) and the metadata corresponding to spatial information of object(ArcGIS, 2021).

As the last piece of the data preparation process, the extent of the raster will be verified to match that of the original satellite data. Extent will be set to latitude and longitude of 89.98°S to 89.98°N and 179.98°W to 179.98°E, respectively. The coordinate reference system (CRS) will be defined to be relative to the WGS84 datum for easy manipulation by the end user.

Data Processing

A significant feature of the SST and Chlo-*a* databases will be the addition of the segmentation by the world's exclusive economic zone (EEZ). EEZ is a marine zone within 200 nautical miles from a country's coastline where each country claims jurisdiction for economic activities(*United Nations Convention on the Law of the Sea*, 1982). Given the oceanographic nature of the data, focusing on the 200-mile buffer of EEZ will provide a more comprehensive explanation of oceanic changes, with the potential to promote the development of ocean planning initiatives directly influencing human settlements on the coasts. To represent the EEZ, a geospatial vector file in shapefile format, already constructed, delimits a buffer of ~200 miles off coastlines globally.

The EEZ regions will be defined using the functions crop and mask from the raster(Hijmans, 2020) package. The function mask will allow to place the area of interest (i.e., the EEZ) on top of each monthly raster, assigning no value to cells outside of the area of interest, while the function crop ensured that each raster matched the extent of that of the area of interest (Figure 3).

Statistical Analysis

Complementary to the core database, data were treated as an m by n matrix, where m represents the years and n represents the months and stacked in two distinct ways (1) in yearly composites and (2) monthly composites.

$$\sum_{j=1}^n x_{ij} \text{ for } i = 1, \dots, m \quad (1)$$

$$\sum_{i=1}^m x_{ij} \text{ for } j = 1, \dots, n \quad (2)$$

two stacks by using *stack* function in the *raster* package were created (Hijmans, 2020). The mean, range, maximum, minimum, and standard deviation values were estimated for annual and monthly SST and Chlo-*a*. I obtained a total of 90 rasters for the yearly composites (18 years, five different statistics) and 60 rasters for the monthly summaries (12 months, five different statistics).

Data Records

Final data are provided in the form of GeoTIFFs for the EEZs boundaries and statistical analysis results (Castaneda-Guzman, Mantilla-Saltos, Murray, Settlage, et al., 2021a). Data can be downloaded based on annually, monthly, or as summary composites of the 18-years period. Data can also be updated using the code included in the Auxiliary Material in Figshare (Castaneda-Guzman, Mantilla-Saltos, Murray, Settlage, et al., 2021b).

Technical Validation

Remotely sensed environmental observations from the MODIS instrument, including SST and Chlo-*a*, have been validated profusely by the scientific community against a number of models and *in situ* measurements (H. Fang et al., 2012; Gentemann, 2014; Hao et al., 2017; Hoge et al., 2003; Hosoda et al., 2007; Remer, 2002; Sims et al., 2006; Tilstone et al., 2013) and used in a diverse set of studies (Chen & Quan, 2013; Escobar et al., 2015a; Golder et al., 2021; Kilpatrick et al., 2015; Ma et al., 2021; Miles & He, 2010; Minnett et al., 2002; Moradi & Kabiri, 2015; Qin et al., 2014; Saulquin et al., 2011; Watts et al., 2019, 2020). For instance, validation of the SST observation uses accurate ship-based infrared radiometers and differing and moored buoys with thermometers a meter of depth (Hao et al., 2017; Hosoda et al., 2007; Minnett et al., 2004). NASA's standard processing and distribution of the SST products are performed using software developed by the Ocean Biology Processing Group (NASA, 2021). SST products are validated internally by NASA using a collocated matchup database of *in situ* observations that are collected within 30 minutes of an overpass and 10 km of a pixel. MODIS SST observations represent the thermal skin layer of the ocean, which is <1 mm thick and is cooler than the underlying water due to vertical heat flux (Hanafin & Minnett, 2002; E. W. Wong & Minnett, 2018). At night or when wind speeds are greater than ~6 m/s, the relationship between the skin temperature and the subsurface are nearly equal. It is under these conditions that validation and uncertainty estimates relative to sub-surface *in situ* buoys are typically reported (Esaias et al., 1998; Minnett et al., 2004). The estimation vs. observation relationship, however, can be very variable under conditions of low wind speeds and reduced sub-surface turbulence (C. J. Donlon et al., 2002; B. Ward, 2006). Furthermore, NASA MODIS uses a collection of cloud classification algorithms to indicate when a pixel corresponds to clear sky conditions (i.e., no cloud coverage). The most recent cloud-classification method is the Alternating Decision Tress (Kilpatrick et al., 2001). Other SST

observations validations tests include a regional ice test, where reflectance threshold are determined using the Sentiner-2 MSI calibrated reflectance(Hollstein et al., 2016) and correction of dust contamination(Luo et al., 2019).

MODIS Chlo-*a* observations are derived from the O'Reilly OC3M algorithm and the Hu color index (Hu et al., 2012b; O'Reilly et al., 1998a). The algorithm is calculated using an empirical relationship from *in situ* measurements and remote sensing reflectance in the blue-to-green region of the visible spectrum. Level 3 MODIS data may provide biased minima and maxima values during errors in the observation that, for example, has some cloud contamination or sunlight affecting the value captured by the sensor. Due to potential atmospheric contamination some regions could have a limited number of observations from which to estimate the monthly values, which increases uncertainty. There is an estimated $\pm 35\%$ nominal uncertainty related to the OC3M algorithm use to derive the global Chlo-*a* product. Nevertheless, error could increase in optically complex waters like those present in coastal areas (Moore et al., 2009; Pieri et al., 2015).

I performed a data validation procedure comparing MODIS observation of SST and Chlo-*a* against gold-standard sensors. More specifically, I compared MODIS data against SST data from Sentinel-3 (Tilstone et al., 2021) during the year 2020. I found that data from MODIS and Sentinel-3 were statistically indistinguishable with a Pearson correlation coefficient of $r=0.99$ for the annual mean, minimum, and maximum composites ($R^2=0.99$, $p<0.05$; Figure 4). Additionally, Chlo-*a* data were evaluated by comparing MODIS data against SeaWiFS(O'Reilly et al., 1998a) observations for the year 2010, when the SeaWiFS satellite ended operations. I found that MODIS Chlo-*a* data were significantly correlated with SeaWiFS Chlo-*a* data but with less strength than for SST evaluations. More specifically, correlation was $r=0.83$ ($R^2=0.67$, $p<0.05$) for the mean, $r=0.71$ ($R^2=0.53$, $p<0.05$) for the maximum, and $r=0.76$ ($R^2=0.52$, $p<0.05$) for the minimum Chlo-*a*

composites (Figure 5). Together, these results suggest that MODIS data have a robust representation of environmental conditions in global coastal waters, at least when compared against gold-standard datasets of SST and Chlo-*a*.

Usage Notes

The proposed use of this dataset is for coarse-scale, regional or global-level studies of coastal environmental conditions. Fine-scale assessments of SST and Chlo-*a* are warranted to improve accuracy and detail of these variables for local-level applications. The data can be used to identify anomalies for SST and Chlo-*a* at local, regional, and global levels. The example demonstrates SST and Chlo-*a* data explorations in tropical and temperate localities, identifying patterns along time (Figure 6). Areas in the mid-Atlantic region of the United States show an increase in mean SST during the month of June to October (Figure 6a), while areas in the subtropics of the Americas (i.e., Ecuador and Colombia) reveal cooler temperatures during the same period (Figure 6b). Additional exploration of the data in tropical and subtropical zones of different latitude reveal that Chlo-*a* increases from September to December (Figure 7b). Contrarily, in the tropics, Chlo-*a* concentration increases between March and May (Figure 7a).

Code Availability

Code in R language to recreate the database and the figures in the Usage Notes is available on Figshare (Castaneda-Guzman, Mantilla-Saltos, Murray, Settlage, et al., 2021a).

References

- Alesheikh, A. A., Ghorbanali, A., & Nouri, N. (2007). Coastline change detection using remote sensing. *International Journal of Environmental Science & Technology*, 4(1), 61–66. <https://doi.org/10.1007/BF03325962>
- ArcGIS. (2021). *What is a raster data?* <https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm>
- Atlantic, S., Martinelli Filho, J. E., Lopes, R. M., Rivera, I. N. G., & Colwell, R. R. (2010). *Vibrio cholerae* O1 detection in estuarine and coastal zooplankton. *Journal of Plankton Research*, 33, 51–62. <https://doi.org/10.1093/plankt/fbq093>
- Barange, M., Bahri, T., Beveridge, M., Cochrane, K., Funge-Smith, S., & Poulain, F. (2018). *Impacts of climate change on fisheries and aquaculture—Synthesis of current knowledge, adaptation and mitigation options.*
- Barbera, A. La, Zerpa, A., Grau, C., Barbera, A. La, Zerpa, A., Silva, S., & Gallardo, O. (2004). Aislamiento de *Vibrio spp.* y evaluación de la condición sanitaria de los moluscos bivalvos Arca zebra y Perna perna procedentes de la costa nororiental del Edo, Sucre, Venezuela. *FCU-LUZ*, 14, 513–521.
- Batabyal, P., Mookerjee, S., Einsporn, M. H., Lara, R. J., & Palit, A. (2016). Environmental drivers on seasonal abundance of riverine-estuarine *V. cholerae* in the Indian Sundarban mangrove. *Ecological Indicators*, 69, 59–65. <https://doi.org/10.1016/j.ecolind.2016.04.004>
- Binsztein, N., Costagliola, M., Pichel, M., Jurquiza, V., Ramirez, F. C., Akselman, R., Vacchino, M., Huq, A., & Colwell, R. R. (2004). Viable but nonculturable *Vibrio cholerae* O1 in the aquatic environment of Argentina. *Appl Environ Microbiol*, 70, 7481–7486. <https://doi.org/10.1128/AEM.70.12.7481-7486.2004>
- Bliem, R., Reischer, G., Linke, R., Farnleitner, A., & Kirschner, A. (2018). Spatiotemporal dynamics of *Vibrio cholerae* in turbid alkaline lakes as determined by quantitative PCR. *Applied and Environmental Microbiology*, 84, 1–14. <https://doi.org/10.1128/AEM.00317-18>
- Carter, W. D., & Paulson, R. W. (1979). *Introduction to monitoring dynamic environmental phenomena of the world using satellite data collection systems.* Geological Survey. <https://doi.org/https://doi.org/10.3133/cir803>
- Castaneda-Guzman, M., Mantilla-Saltos, G., Murray, K. A., Settlage, R., & Escobar, L. E. (2021a). A database of global coastal conditions. *Figshare*. <https://doi.org/https://doi.org/10.6084/m9.figshare.c.5660263.v1>
- Castaneda-Guzman, M., Mantilla-Saltos, G., Murray, K. A., Settlage, R., & Escobar, L. E. (2021b). Methods and code. *Figshare*. <https://doi.org/https://doi.org/10.6084/m9.figshare.13708642.v4>
- Center for Disease Control and Prevention (CDC). (2019). *Vibrio species causing Vibriosis.* <https://www.cdc.gov/vibrio/vibrio-oysters.html>
- Chen, J., & Quan, W. (2013). An improved algorithm for retrieving chlorophyll-*a* from the Yellow River Estuary using MODIS imagery. *Environmental Monitoring and Assessment*, 185(3), 2243–2255. <https://doi.org/10.1007/s10661-012-2705-y>

- Cobos, M. E., Jiménez, L., Nuñez-Penichet, C., Romero-Alvarez, D., & Simoes, M. (2018). Sample data and training modules for cleaning biodiversity information. *Biodiversity Informatics*, 13, 49–50. <https://doi.org/10.17161/bi.v13i0.7600>
- Colwell, R. R. (1996). Global climate and infectious disease: The cholera paradigm. *Science*, 274(5295), 2025–2031. <https://doi.org/10.1126/science.274.5295.2025>
- Dalusi, L., Lyimo, T. J., Lugomela, C., Hosea, K. M. M., & Sjöling, S. (2015). Toxigenic *Vibrio cholerae* identified in estuaries of Tanzania using PCR techniques. *FEMS Microbiology Letters*, 362, fnv009. <https://doi.org/10.1093/femsle/fnv009>
- de Menezes, F. G. R., Neves, S. da S., de Sousa, O. V., Vila-Nova, C. M. V. M., Maggioni, R., Theophilo, G. N. D., Hofer, E., & Vieira, R. H. S. dos F. (2014). Detection of virulence genes in environmental strains of *Vibrio cholerae* from estuaries in northeastern Brazil. *Revista Do Instituto de Medicina Tropical de São Paulo*, 56, 427–432. <https://doi.org/10.1590/S0036-46652014000500010>
- De Menezes, F. G. R., Rodriguez, M. T. T., de Carvalho, F. C. T. T., Rebouças, R. H., Costa, R. A., De Sousa, O. V., Hofer, E., Vieira, R. H. S. F. S. F., Rofriguez, M. T. T., de Carvalho, F. C. T. T., Rebouças, R. H., Costa, R. A., De Sousa, O. V., Hofer, E., & Vieira, R. H. S. F. S. F. (2017). Pathogenic *Vibrio* species isolated from estuarine environments (Ceará, Brazil) - antimicrobial resistance and virulence potential profiles. *Anais Da Academia Brasileira de Ciências*, 89, 1175–1188. <https://doi.org/10.1590/0001-3765201720160191>
- Dewan, A. M., & Yamaguchi, Y. (2009). Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Applied Geography*, 29(3), 390–401. <https://doi.org/10.1016/j.apgeog.2008.12.005>
- Dheenan, P. S., Jha, D. K., Vinithkumar, N. V., Ponmalar, A. A., Venkateshwaran, P., & Kirubakaran, R. (2014). Spatial variation of physicochemical and bacteriological parameters elucidation with GIS in Rangat Bay, Middle Andaman, India. *Journal of Sea Research*, 85, 534–541. <https://doi.org/10.1016/j.seares.2013.09.001>
- Di, D. Y. W., Lee, A., Jang, J., Han, D., & Hur, H. G. (2017). Season-specific occurrence of potentially pathogenic *Vibrio spp.* on the southern coast of South Korea. *Applied and Environmental Microbiology*, 83, 1–13. <https://doi.org/10.1128/AEM.02680-16>
- Donlon, C. J., Minnett, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J., Ward, B., & Murray, M. J. (2002). Toward improved validation of satellite SST measurements for climate research. *Journal of Climaternal*, 15(February), 353–369. [https://doi.org/10.1175/1520-0442\(2002\)015<0353](https://doi.org/10.1175/1520-0442(2002)015<0353)
- Donlon, C., Robinson, I., Casey, K. S., Vazquez-Cuervo, J., Armstrong, E., Arino, O., Gentemann, C., May, D., LeBorgne, P., Piollé, J., Barton, I., Beggs, H., Poulter, D. J. S., Merchant, C. J., Bingham, A., Heinz, S., Harris, A., Wick, G., Emery, B., ... Rayner, N. (2007). The global ocean data assimilation experiment high-resolution sea surface temperature pilot project. *Bulletin of the American Meteorological Society*, 88(8), 1197–1214. <https://doi.org/10.1175/BAMS-88-8-1197>
- Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini,

- P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, 120, 25–36. <https://doi.org/10.1016/j.rse.2011.11.026>
- Esaias, W. E., Abbott, M. R., Barton, I., Brown, O. B., Campbell, J. W., Carder, K. L., Clark, D. K., Evans, R. H., Hoge, F. E., Gordon, H. R., Balch, W. M., Letelier, R., & Minnett, P. J. (1998). An overview of MODIS capabilities for ocean science observations. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4), 1250–1265. <https://doi.org/10.1109/36.701076>
- Escobar, L. E., Qiao, H., Lee, C., & Phelps, N. B. D. (2017). Novel methods in disease biogeography: A case study with Heterosporosis. *Frontiers in Veterinary Science*, 4(JUL). <https://doi.org/10.3389/fvets.2017.00105>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015a). A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015b). A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015c). A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Esteves, K., Hervio-Heath, D., Mosser, T., Rodier, C., Tournoud, M., Jumas-Bilak, E., Colwell, R. R., & Monfort, P. (2015a). *Vibrio cholerae* during freshwater flash Floods in french mediterranean coastal lagoons. *Frontiers in Microbiology*, 81, 7600–7609. <https://doi.org/10.1128/AEM.01848-15.Editor>
- Esteves, K., Hervio-Heath, D., Mosser, T., Rodier, C., Tournoud, M.-G., Jumas-Bilak, E., Colwell, R. R., & Monfort, P. (2015b). Rapid proliferation of *Vibrio parahaemolyticus*, *Vibrio vulnificus*, and *Vibrio cholerae* during freshwater flash floods in french mediterranean coastal lagoons. *Applied and Environmental Microbiology*, 81, 7600–7609. <https://doi.org/10.1128/AEM.01848-15>
- Fang, H., Wei, S., & Liang, S. (2012). Validation of MODIS and CYCLOPES LAI products using global field measurement data. *Remote Sensing of Environment*, 119, 43–54. <https://doi.org/10.1016/j.rse.2011.12.006>
- Fang, L., Ginn, A. M., Harper, J., Kane, A. S., & Wright, A. C. (2019). Survey and genetic characterization of *Vibrio cholerae* in Apalachicola Bay, Florida (2012–2014). *Journal of Applied Microbiology*, 126, 1265–1277. <https://doi.org/10.1111/jam.14199>
- Fernández-Delgado, M., Suárez, P., Giner, S., Sanz, V., Peña, J., Sánchez, D., & García-Amado, M. A. (2017). Occurrence and virulence properties of *Vibrio* and *Salini vibrio* isolates from

- tropical lagoons of the southern Caribbean Sea. *Antonie van Leeuwenhoek*, 110, 833–841. <https://doi.org/10.1007/s10482-017-0856-0>
- Fri, J., Ndip, R. N., Njom, H. A., & Clarke, A. M. (2017). Occurrence of virulence genes associated with human pathogenic vibrios isolated from two commercial Dusky Kob (*Argyrosomus japonicus*) farms and kareiga estuary in the Eastern Cape Province, South Africa. In *International Journal of Environmental Research and Public Health* (Vol. 14). <https://doi.org/10.3390/ijerph14101111>
- Gardade, L., & Khandeparker, L. (2017). Spatio-temporal variations in pathogenic bacteria in the surface sediments of the Zuari Estuary, Goa, India. *Current Science*, 113(9), 1729. <https://doi.org/10.18520/cs/v113/i09/1729-1738>
- Gdoura, M., Sellami, H., Nasfi, H., Trabelsi, R., Mansour, S., Attia, T., Nsaibia, S., Vallaey, T., Gdoura, R., & Siala, M. (2016). Molecular detection of the three major pathogenic *vibrio* species from seafood products and sediments in Tunisia using real-Time PCR. *Journal of Food Protection*, 79(12), 2086–2094. <https://doi.org/10.4315/0362-028X.JFP-16-205>
- Gentemann, C. L. (2014). Three way validation of MODIS and AMSR-E sea surface temperatures. *Journal of Geophysical Research: Oceans*, 119(4), 2583–2598. <https://doi.org/10.1002/2013JC009716>
- Golder, M. R., Shuva, M. S. H., Rouf, M. A., Uddin, M. M., Bristy, S. K., & Bir, J. (2021). Chlorophyll-*a*, SST and particulate organic carbon in response to the cyclone Amphan in the Bay of Bengal. *Journal of Earth System Science*, 130(3), 157. <https://doi.org/10.1007/s12040-021-01668-1>
- Green, E. P., Mumby, P. J., Edwards, A. J., & Clark, C. D. (1996). A review of remote sensing for the assessment and management of tropical coastal resources. *Coastal Management*, 24(1), 1–40. <https://doi.org/10.1080/08920759609362279>
- Green, K., Kempka, D., & Lackey, L. (1994). Using remote sensing to detect and monitor land-cover and land-use change. *Photogrammetric Engineering & Remote Sensing*, 60(3), 331–337.
- Grimes, J. D., Ford, T. E., Colwell, R. R., Baker-Austin, C., Martinez-Urtaza, J., Subramaniam, A., & Capone, D. G. (2014). Viewing marine bacteria, their activity and response to environmental drivers from orbit: satellite remote sensing of bacteria. *Microbial Ecology*, 67(3), 489–500. <https://doi.org/10.1007/s00248-013-0363-4>
- Grothen, D. C., Zach, S. J., & Davis, P. H. (2017). Detection of intestinal pathogens in river, shore, and drinking water in Lima, Peru. *Journal of Genomics*, 5, 4–11. <https://doi.org/10.7150/jgen.18378>
- Gyraite, G., Katarzyte, M., & Schernewski, G. (2019). First findings of potentially human pathogenic bacteria *Vibrio* in the south-eastern Baltic Sea coastal and transitional bathing waters. *Marine Pollution Bulletin*, 149(September), 110546. <https://doi.org/10.1016/j.marpolbul.2019.110546>

- Hackbusch, S., Wichels, A., Gimenez, L., Döpke, H., & Gerdt, G. (2020). Potentially human pathogenic *Vibrio spp.* in a coastal transect: Occurrence and multiple virulence factors. *Science of the Total Environment*, 707. <https://doi.org/10.1016/j.scitotenv.2019.136113>
- Hanafin, J. A., & Minnett, P. J. (2002). Thermal profiling of the sea surface skin layer using FTIR measurements. In *Gas Transfer at Water Surfaces* (pp. 161–166). Blackwell Publishing Ltd. <https://doi.org/10.1029/GM127p0161>
- Hao, Y., Cui, T., Singh, V. P., Zhang, J., Yu, R., & Zhang, Z. (2017). Validation of MODIS sea surface temperature product in the coastal waters of the Yellow Sea. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5), 1667–1680. <https://doi.org/10.1109/JSTARS.2017.2651951>
- Hayashi, M., Jin, F., & Stuecker, M. F. (2020). Dynamics for El Niño-La Niña asymmetry constrain Equatorial-Pacific warming pattern. *Nature Communications*, 11(1), 1–10. <https://doi.org/10.1038/s41467-020-17983-y>
- Hijmans, R. J. (2020). *raster: Geographic data analysis and modeling* (R package version 3.1-5). <https://cran.r-project.org/package=raster>
- Hillger, D., Kopp, T., Lee, T., Lindsey, D., Seaman, C., Miller, S., Solbrig, J., Kidder, S., Bachmeier, S., Jasmin, T., & Rink, T. (2013). First-light imagery from Suomi NPP VIIRS. *Bulletin of the American Meteorological Society*, 94(7), 1019–1029. <https://doi.org/10.1175/BAMS-D-12-00097.1>
- Hoge, F. E., Lyon, P. E., Swift, R. N., Yungel, J. K., Abbott, M. R., Letelier, R. M., & Esaias, W. E. (2003). Validation of Terra-MODIS phytoplankton chlorophyll fluorescence line height I Initial airborne lidar results. *Applied Optics*, 42(15), 2767. <https://doi.org/10.1364/AO.42.002767>
- Hollstein, A., Segl, K., Guanter, L., Brell, M., & Enesco, M. (2016). Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sensing*, 8(8), 666. <https://doi.org/10.3390/rs8080666>
- Horning, N., Robinson, J. a, Sterling, E. J., Turner, W., & Spector, S. (2010). Remote sensing for ecology and conservation. In *Techniques in Ecology & Conservation Series*. Oxford University Press.
- Hosoda, K., Murakami, H., Sakaida, F., & Kawamura, H. (2007). Algorithm and validation of sea surface temperature observation using MODIS sensors aboard terra and aqua in the western North Pacific. *Journal of Oceanography*, 63(2), 267–280. <https://doi.org/10.1007/s10872-007-0027-4>
- Hu, C., Lee, Z., & Franz, B. (2012a). Chlorophyll-*a* algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, 117, C01011. <https://doi.org/10.1029/2011JC007395>
- Hu, C., Lee, Z., & Franz, B. (2012b). Chlorophyll-*a* algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, 117, C01011. <https://doi.org/10.1029/2011JC007395>

- Izumiya, H., Furukawa, M., Ogata, K., Isobe, J., Watanabe, S., Sasaki, M., Ichinose, K., Arakawa, E., Morita, M., Kurane, I., & Ohnishi, M. (2017). A double-quadratic model for predicting *Vibrio* species in water environments of Japan. *Archives of Microbiology*, *199*, 1293–1302. <https://doi.org/10.1007/s00203-017-1402-1>
- Khamesipour, R. M., Rahimi, E., & Khodadoostan, A. (2014). Occurrence of *Vibrio* spp., *Aeromonas hydrophila*, *Escherichia coli* and *Campylobacter* spp. in crayfish (*Astacus leptodactylus*) from Iran. *Iranian Journal of Fisheries Sciences*, *13*, 944–954.
- Kilpatrick, K. A., Podestá, G. P., & Evans, R. (2001). Overview of the NOAA/NASA advanced very high resolution radiometer Pathfinder algorithm for sea surface temperature and associated matchup database. *Journal of Geophysical Research: Oceans*, *106*(C5), 9179–9197. <https://doi.org/10.1029/1999JC000065>
- Kilpatrick, K. A., Podestá, G., Walsh, S., Williams, E., Halliwell, V., Szczodrak, M., Brown, O. B., Minnett, P. J., & Evans, R. (2015). A decade of sea surface temperature from MODIS. *Remote Sensing of Environment*, *165*, 27–41. <https://doi.org/10.1016/j.rse.2015.04.023>
- Kim, J. Y., & Lee, J. L. (2014). Multipurpose assessment for the quantification of *Vibrio* spp. and total bacteria in fish and seawater using multiplex real-time polymerase chain reaction. In *Journal of the Science of Food and Agriculture* (Vol. 94, Issue 13, pp. 2807–2817). <https://doi.org/10.1002/jsfa.6699>
- Kokashvili, T., Whitehouse, C. A., Tskhvediani, A., Grim, C. J., Elbakidze, T., Mitaishvili, N., Janelidze, N., Jaiani, E., Haley, B. J., Lashkhi, N., Huq, A., Colwell, R. R., & Tediashvili, M. (2015). Occurrence and Diversity of Clinically Important *Vibrio* Species in the Aquatic Environment of Georgia. *Frontiers in Public Health*, *3*(October), 1–12. <https://doi.org/10.3389/fpubh.2015.00232>
- Laffoley, D., & Baxter, J. M. (2016). *Explaining ocean warming: Causes, scale, effects and consequences* (D. Laffoley & J. M. Baxter, Eds.). IUCN, International Union for Conservation of Nature. <https://doi.org/10.2305/IUCN.CH.2016.08.en>
- Li, J., Pei, Y., Zhao, S., Xiao, R., Sang, X., & Zhang, C. (2020). A review of remote sensing for environmental monitoring in China. *Remote Sensing*, *12*(7), 1130. <https://doi.org/10.3390/rs12071130>
- Lipp, E. K. E., Rivera, I. N. G. I., Gil, A. I. A., Espeland, E. M., Choopun, N., Louis, V. R., Russek-Cohen, E., Huq, A., & Colwell, R. R. (2003). Direct detection of *Vibrio cholerae* and *ctxA* in Peruvian coastal water and plankton by PCR. *Applied Environ Mircrobiol*, *69*(6), 3676. <https://doi.org/10.1128/AEM.69.6.3676>
- Lipp, E. K., Huq, A., & Colwell, R. R. (2002). Effects of global climate on infectious disease: the cholera model. *Clinical Microbiology Reviews*, *15*(4), 757–770. <https://doi.org/10.1128/CMR.15.4.757-770.2002>
- Liu, J. (1997). A process-based boreal ecosystem productivity simulator using remote sensing inputs. *Remote Sensing of Environment*, *62*(2), 158–175. [https://doi.org/10.1016/S0034-4257\(97\)00089-8](https://doi.org/10.1016/S0034-4257(97)00089-8)

- López, L., Manjarrez, G., Herrera, L., Montes, A., Olascuaga, Y., & Ortega, R. (2010). Estudio piloto para el aislamiento de *Vibrio spp.* en ostras (*Crassostera rhizophorae*) capturadas en la Ciénaga de la Virgen, Cartagena, Colombia. *RSPYN*, *11*, 1–6.
- Louis, V. R., Russek-Cohen, E., Choopun, N., Rivera, I. N. G., Gangle, B., Jiang, S. C., Rubin, A., Patz, J. A., Huq, A., Colwell, R. R., Louis, R., & Rubin, A. (2003). Predictability of *Vibrio cholerae* in Chesapeake Bay. *Appl Environ Microbiol*, *69*, 2773–2785. <https://doi.org/10.1128/AEM.69.5.2773-2785.2003>
- Luo, B., Minnett, P. J., Gentemann, C., & Szczodrak, G. (2019). Improving satellite retrieved night-time infrared sea surface temperatures in aerosol contaminated regions. *Remote Sensing of Environment*, *223*, 8–20. <https://doi.org/10.1016/j.rse.2019.01.009>
- Ma, S., Zhang, X., Ding, C., Han, W., & Lu, Y. (2021). Comparison of the Spatiotemporal Variation of Chl-*a* in the East China Sea and Bohai Sea based on long time series satellite data. *2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 1–6. <https://doi.org/10.1109/Agro-Geoinformatics50104.2021.9530337>
- Machado, A., & Bordalo, A. A. (2016). Detection and quantification of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and *Vibrio vulnificus* in Coastal Waters of Guinea-Bissau (West Africa). *EcoHealth*, *13*, 339–349. <https://doi.org/10.1007/s10393-016-1104-1>
- Main, C. R., Salvitti, L. R., Whereat, E. B., & Coyne, K. J. (2015). Community-level and species-specific associations between phytoplankton and particle-associated *Vibrio* species in Delaware’s Inland Bays. *Applied and Environmental Microbiology*, *81*(17), 5703–5713. <https://doi.org/10.1128/AEM.00580-15>
- Matteucci, G., Schippa, S., Di Lallo, G., Migliore, L., & Thaller, M. C. (2015). Species diversity, spatial distribution, and virulence associated genes of culturable vibrios in a brackish coastal Mediterranean environment. *Annals of Microbiology*, *65*(4), 2311–2321. <https://doi.org/10.1007/s13213-015-1073-6>
- Meena, B., Anburajan, L., Sathish, T., Das, A. K., Vinithkumar, N. V., Kirubakaran, R., & Dharani, G. (2019). Studies on diversity of *Vibrio sp.* and the prevalence of hapA, tcpI, st, rtxA&C, acfB, hlyA, ctxA, ompU and toxR genes in environmental strains of *Vibrio cholerae* from Port Blair bays of South Andaman, India. In *Marine Pollution Bulletin* (pp. 105–116). <https://doi.org/10.1016/j.marpolbul.2019.05.011>
- Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, *415*, 108837. <https://doi.org/10.1016/j.ecolmodel.2019.108837>
- Meyer, J. L., Gunasekera, S. P., Scott, R. M., Paul, V. J., & Teplitski, M. (2016). Microbiome shifts and the inhibition of quorum sensing by Black Band Disease cyanobacteria. *ISME Journal*, *10*(5), 1204–1216. <https://doi.org/10.1038/ismej.2015.184>
- Michna, P., & Woods, M. (2019). *RNetCDF: Interface to “NetCDF” Datasets* (R package version 2.1-1).

- Miles, T. N., & He, R. (2010). Temporal and spatial variability of Chl-*a* and SST on the South Atlantic Bight: Revisiting with cloud-free reconstructions of MODIS satellite imagery. *Continental Shelf Research*, 30(18), 1951–1962. <https://doi.org/10.1016/j.csr.2010.08.016>
- Ming, H., Ma, Y., Gu, Y., Su, J., Guo, J., Li, J., Li, X., Jin, Y., & Fan, J. (2020). Enterococci may not present the pollution of most enteric pathogenic bacteria in recreational seawaters of Xinghai bathing Beach, China. *Ecological Indicators*, 110, 105938. <https://doi.org/10.1016/j.ecolind.2019.105938>
- Minnett, P. J. (1990). Satellite infrared scanning radiometers — AVHRR and ATSR/M. In *Microwave Remote Sensing for Oceanographic and Marine Weather-Forecast Models* (pp. 141–163). Springer Netherlands. https://doi.org/10.1007/978-94-009-0509-2_7
- Minnett, P. J. (2010). The validation of sea surface temperature retrievals from spaceborne infrared radiometers. In *Oceanography from Space* (pp. 229–247). Springer Netherlands. https://doi.org/10.1007/978-90-481-8681-5_14
- Minnett, P. J., Brown, O. B., Evans, R. H., Key, E. L., Kearns, E. J., Kilpatrick, K., Kumar, A., Maillet, K. A., & Szczodrak, G. (2004). Sea-surface temperature measurements from the moderate-resolution imaging spectroradiometer (MODIS) on aqua and terra. *IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004*, 7(10), 4576–4579. <https://doi.org/10.1109/IGARSS.2004.1370173>
- Minnett, P. J., & Corlett, G. K. (2012). A pathway to generating climate data records of sea-surface temperature from satellite measurements. *Deep Sea Research Part II: Topical Studies in Oceanography*, 77–80, 44–51. <https://doi.org/10.1016/j.dsr2.2012.04.003>
- Minnett, P. J., Evans, R. H., Kearns, E. J., & Brown, O. B. (2002). Sea-surface temperature measured by the Moderate Resolution Imaging Spectroradiometer (MODIS). *IEEE International Geoscience and Remote Sensing Symposium*, 2, 1177–1179. <https://doi.org/10.1109/IGARSS.2002.1025872>
- Minnett, P. J., Evans, R. H., Podestá, G. P., & Kilpatrick, K. A. (2014). Sea-surface temperature from Suomi-NPP VIIRS: algorithm development and uncertainty estimation. In W. W. Hou & R. A. Arnone (Eds.), *SPIE 9111, Ocean Sensing and Monitoring VI* (p. 91110C). <https://doi.org/10.1117/12.2053184>
- Moore, T. S., Campbell, J. W., & Dowell, M. D. (2009). A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sensing of Environment*, 113(11), 2424–2430. <https://doi.org/10.1016/j.rse.2009.07.016>
- Moradi, M., & Kabiri, K. (2015). Spatio-temporal variability of SST and Chlorophyll-*a* from MODIS data in the Persian Gulf. *Marine Pollution Bulletin*, 98(1–2), 14–25. <https://doi.org/10.1016/j.marpolbul.2015.07.018>
- Mukhopadhyay, A. K. A., Basu, A., Garg, P., Bag, P. K., Ghosh, A., Bhattacharya, S. K., Takeda, Y., & Nair, G. B. (1998). Molecular epidemiology of reemergent *Vibrio cholerae* O139 Bengal in India. *Journal of Clinical Microbiology*, 36, 2149–2152.

- Murray, K. A., Escobar, L. E., Lowe, R., Rocklöv, J., Semenza, J. C., & Watts, N. (2020). Tracking infectious diseases in a warming world. *BMJ*, *371*(1), m3086. <https://doi.org/10.1136/bmj.m3086>
- Nagendra, H. (2001). Using remote sensing to assess biodiversity. *International Journal of Remote Sensing*, *22*(12), 2377–2400. <https://doi.org/10.1080/01431160117096>
- NASA. (2021). *MODIS (Moderate Resolution Imaging Spectroradiometer)*. <https://modis.gsfc.nasa.gov/about/>
- Neogi, S. B., Lara, R., Alam, M., Harder, J., Yamasaki, S., & Colwell, R. R. (2018). Environmental and hydroclimatic factors influencing *Vibrio* populations in the estuarine zone of the Bengal delta. *Environmental Monitoring and Assessment*, *190*, 565. <https://doi.org/10.1007/s10661-018-6925-7>
- NOAA. (2020). *Ocean Facts: Why do scientists measure sea surface temperature?* <https://oceanservice.noaa.gov/facts/sea-surface-temperature.html>
- NOAA. (2021). *National Oceanic and Atmospheric Administration (NOAA) Coastal Watch*. <https://coastwatch.pfeg.noaa.gov/erddapinfo/>
- Nurdin, S., Mustapha, M. A., & Lihan, T. (2013). The relationship between sea surface temperature and chlorophyll-*a* concentration in fisheries aggregation area in the archipelagic waters of spermonde using satellite images. *AIP Conference Proceedings*, *1571*, 466–472. <https://doi.org/10.1063/1.4858699>
- O'Brien, J. (2020). *From MODIS to VIIRS - Making the switch for air quality professionals*. NASA Earth Science/Applied Science. <https://appliedsciences.nasa.gov/our-impact/news/modis-viirs-making-switch-air-quality-professionals>
- O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., & McClain, C. (1998a). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans*, *103*(C11), 24937–24953. <https://doi.org/10.1029/98JC02160>
- O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., & McClain, C. (1998b). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans*, *103*(C11), 24937–24953. <https://doi.org/10.1029/98JC02160>
- Orozco, R., Castillo, S., Enríquez, E., Fernández, E., Morón, O., & Córdova, J. (1996). *Evaluación de la constaminación y calidad microbiológica del agua de mar en las bahías de Ferrol y Samanco* (Vol. 56). Instituto del Mar del Peru.
- Pal, B. B., Khuntia, H. K., Samal, S. K., Das, S. S., & Chhotray, G. P. (2006). Emergence of *Vibrio cholerae* O1 biotype El Tor serotype Inaba causing outbreaks of cholera in Orissa, India. *Jpn J Infect Dis*, *59*, 266–269.
- Pascual, M., Rodó, X., Ellner, S. P., Colwell, R. R., & Bouma, M. J. (2000). Cholera dynamics and El Niño-Southern Oscillation. *Science*, *289*, 1766–1769. <https://doi.org/10.1126/science.289.5485.1766>

- Perkins, T. L., Clements, K., Baas, J. H., Jago, C. F., Jones, D. L., Malham, S. K., & McDonald, J. E. (2014). Sediment composition influences spatial variation in the abundance of human pathogen indicator bacteria within an estuarine environment. *PLoS ONE*, *9*, e112951. <https://doi.org/10.1371/journal.pone.0112951>
- Pieri, M., Massi, L., Lazzara, L., Nuccio, C., Lapucci, C., & Maselli, F. (2015). Assessment of three algorithms for the operational estimation of CHL from MODIS data in the Western Mediterranean Sea. *European Journal of Remote Sensing*, *48*(1), 383–401. <https://doi.org/10.5721/EuJRS20154822>
- Qin, H., Chen, G., Wang, W., Wang, D., & Zeng, L. (2014). Validation and application of MODIS-derived SST in the South China Sea. *International Journal of Remote Sensing*, *35*(11–12), 4315–4328. <https://doi.org/10.1080/01431161.2014.916439>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (4.0.3). <https://www.r-project.org/>
- Remer, L. A. (2002). Validation of MODIS aerosol retrieval over ocean. *Geophysical Research Letters*, *29*(12), 8008. <https://doi.org/10.1029/2001GL013204>
- Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E., & Lundquist, C. J. (2017). A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, *4*. <https://doi.org/10.3389/fmars.2017.00421>
- Rosenqvist, Å., Milne, A., Lucas, R., Imhoff, M., & Dobson, C. (2003). A review of remote sensing technology in support of the Kyoto protocol. *Environmental Science & Policy*, *6*(5), 441–455. [https://doi.org/10.1016/S1462-9011\(03\)00070-4](https://doi.org/10.1016/S1462-9011(03)00070-4)
- Sack, R. B., Siddique, A. K., Longini, I. M., Nizam, A., Islam, M. S., Morris, J. G., Ali, A., Huq, A., Nair, G. B., Qadri, F., Faruque, S. M., Sack, D. A., & Colwell, R. R. (2003). A 4-year study of the epidemiology of *Vibrio cholerae* in four rural areas of Bangladesh. *The Journal of Infectious Diseases*, *212*(5), 96–101.
- Saulquin, B., Gohin, F., & Garrello, R. (2011). Regional objective analysis for merging high-resolution MERIS, MODIS/Aqua, and SeaWiFS chlorophyll-*a* data from 1998 to 2008 on the European Atlantic shelf. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(1), 143–154. <https://doi.org/10.1109/TGRS.2010.2052813>
- Schowengerdt, R. A. (2012). *Remote sensing: models and methods for image processing*. Elsevier. <https://doi.org/10.1016/C2009-0-21902-7>
- Shen, L., Xu, H., & Guo, X. (2012a). Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework. *Sensors*, *12*(6), 7778–7803. <https://doi.org/10.3390/s120607778>
- Shen, L., Xu, H., & Guo, X. (2012b). Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework. *Sensors*, *12*(6), 7778–7803. <https://doi.org/10.3390/s120607778>

- Siboni, N., Balaraju, V., Carney, R., Labbate, M., & Seymour, J. R. (2016). Spatiotemporal dynamics of *Vibrio* spp. within the Sydney harbour estuary. *Frontiers in Microbiology*, *7*, 460. <https://doi.org/10.3389/fmicb.2016.00460>
- Silva, M. M., Maldonado, G. C., Castro, R. O., de Sá Felizardo, J., Cardoso, R. P., dos Anjos, R. M., & de Araújo, F. V. (2019). Dispersal of potentially pathogenic bacteria by plastic debris in Guanabara Bay, RJ, Brazil. *Marine Pollution Bulletin*, *141*, 561–568. <https://doi.org/10.1016/j.marpolbul.2019.02.064>
- Sims, D. A., Rahman, A. F., Cordova, V. D., El-Masri, B. Z., Baldocchi, D. D., Flanagan, L. B., Goldstein, A. H., Hollinger, D. Y., Misson, L., Monson, R. K., Oechel, W. C., Schmid, H. P., Wofsy, S. C., & Xu, L. (2006). On the use of MODIS EVI to assess gross primary productivity of North American ecosystems. *Journal of Geophysical Research: Biogeosciences*, *111*(G4). <https://doi.org/10.1029/2006JG000162>
- Singh, A. (1989). Review article: Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, *10*(6), 989–1003. <https://doi.org/10.1080/01431168908903939>
- Sneha, K. G., Anas, A., Jayalakshmy, K. V., Jasmin, C., Das, P. V. V., Pai, S. S., Pappu, S., Nair, M., Muraleedharan, K. R., Sudheesh, K., & Nair, S. (2016). Distribution of multiple antibiotic resistant *Vibrio* spp across Palk Bay. *Regional Studies in Marine Science*, *3*, 242–250. <https://doi.org/10.1016/j.rsma.2015.11.004>
- Specter, C., & Gayle, D. (1990). Managing technology transfer for coastal zone development: Caribbean experts identify major issues. *International Journal of Remote Sensing*, *11*(10), 1729–1740. <https://doi.org/10.1080/01431169008955126>
- Stanford. (2021). *Best practices for file formats*. <https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats>
- Sulca, M. A., Orozco, R., & Alvarado, D. E. (2018). Antimicrobial resistance not related to 1,2,3 integrons and Superintegron in *Vibrio* spp. isolated from seawater sample of Lima (Peru). *Marine Pollution Bulletin*, *131*, 370–377. <https://doi.org/10.1016/j.marpolbul.2018.04.050>
- Tilstone, G. H., Lotliker, A. A., Miller, P. I., Ashraf, P. M., Kumar, T. S., Suresh, T., Ragavan, B. R. R., & Menon, H. B. (2013). Assessment of MODIS-Aqua chlorophyll-*a* algorithms in coastal and shelf waters of the Eastern Arabian Sea. *Continental Shelf Research*, *65*, 14–26. <https://doi.org/10.1016/j.csr.2013.06.003>
- Tilstone, G. H., Pardo, S., Dall’Olmo, G., Brewin, R. J. W., Nencioli, F., Dessailly, D., Kwiatkowska, E., Casal, T., & Donlon, C. (2021). Performance of ocean colour chlorophyll-*a* algorithms for Sentinel-3 OLCI, MODIS-Aqua and Suomi-VIIRS in open-ocean waters of the Atlantic. *Remote Sensing of Environment*, *260*, 112444. <https://doi.org/10.1016/j.rse.2021.112444>
- UCAR Community Programs. (2021). *Network Common Data Form (NetCDF)*. <https://www.unidata.ucar.edu/software/netcdf/>
- United Nations Convention on the Law of the Sea*, 1833 U.N.T.S. 397 (1982) (testimony of United Nations).

https://www.un.org/depts/los/convention_agreements/convention_overview_convention.htm

- Ward, B. (2006). Near-surface ocean temperature. *Journal of Geophysical Research*, *111*(C2), C02004. <https://doi.org/10.1029/2004JC002689>
- Ward, D., Phinn, S. R., & Murray, A. T. (2000). Monitoring growth in rapidly urbanizing areas using remotely sensed data. *Professional Geographer*, *52*(3), 371–386. <https://doi.org/10.1111/0033-0124.00232>
- Watts, N., Amann, M., Arnell, N., Ayeb-Karlsson, S., Beagley, J., Belesova, K., Boykoff, M., Byass, P., Cai, W., Campbell-Lendrum, D., Capstick, S., Chambers, J., Coleman, S., Dalin, C., Daly, M., Dasandi, N., Dasgupta, S., Davies, M., Di Napoli, C., ... Costello, A. (2020). The 2020 report of The Lancet Countdown on health and climate change: responding to converging crises. *The Lancet*, *6736*(19). [https://doi.org/10.1016/S0140-6736\(20\)32290-X](https://doi.org/10.1016/S0140-6736(20)32290-X)
- Watts, N., Amann, M., Arnell, N., Ayeb-Karlsson, S., Belesova, K., Boykoff, M., Byass, P., Cai, W., Campbell-Lendrum, D., Capstick, S., Chambers, J., Dalin, C., Daly, M., Dasandi, N., Davies, M., Drummond, P., Dubrow, R., Ebi, K. L., Eckelman, M., ... Montgomery, H. (2019). The 2019 report of The Lancet Countdown on health and climate change: Ensuring that the health of a child born today is not defined by a changing climate. *Lancet*, *394*(10211), 1836–1878. [https://doi.org/10.1016/S0140-6736\(19\)32596-6](https://doi.org/10.1016/S0140-6736(19)32596-6)
- Wei, G. F., Tang, D. L., & Wang, S. (2008). Distribution of chlorophyll and harmful algal blooms (HABs): A review on space based studies in the coastal environments of Chinese marginal seas. *Advances in Space Research*, *41*(1), 12–19. <https://doi.org/10.1016/j.asr.2007.01.037>
- Wong, E. W., & Minnett, P. J. (2018). The response of the ocean thermal skin layer to variations in incident infrared radiation. *Journal of Geophysical Research: Oceans*, *123*(4), 2475–2493. <https://doi.org/10.1002/2017JC013351>
- Wong, Y. Y., Lee, C. W., Bong, C. W., Lim, J. H., Narayanan, K., & Sim, E. U. H. (2019). Environmental control of *Vibrio* spp. abundance and community structure in tropical waters. *FEMS Microbiology Ecology*, *95*, fiz176. <https://doi.org/10.1093/femsec/fiz176>
- World Health Organization. (2013). Cholera annual report. *Wkly Epidemiol Rec*, 321–336.
- Xu, M., Kan, B., & Wang, D. (2015). Identifying environmental risk factors of cholera in a coastal area with geospatial technologies. *International Journal of Environmental Research and Public Health*, *12*(1), 354–370. <https://doi.org/10.3390/ijerph120100354>
- Yue, Y., Gong, J., Wang, D., Kan, B., Li, B., & Ke, C. (2014). Influence of climate factors on *Vibrio cholerae* dynamics in the Pearl River estuary, South China. *World Journal of Microbiology and Biotechnology*, *30*, 1797–1808. <https://doi.org/10.1007/s11274-014-1604-5>
- Zaw, M. T., Emran, N. A., Ibrahim, M. Y., Suleiman, M., Awang Mohd, T. A., Yusuff, A. S., Naing, K. S., Myint, T., Jikal, M., Salleh, M. A., & Lin, Z. (2019). Genetic diversity of toxigenic *Vibrio cholerae* O1 from Sabah, Malaysia 2015. *Journal of Microbiology, Immunology and Infection*, *52*(4), 563–570. <https://doi.org/10.1016/j.jmii.2018.01.003>

Zhang, X., Song, Y., Liu, D., Keesing, J. K., & Gong, J. (2015). Macroalgal blooms favor heterotrophic diazotrophic bacteria in nitrogen-rich and phosphorus-limited coastal surface waters in the Yellow Sea. *Estuarine, Coastal and Shelf Science*, 163, 75–81. <https://doi.org/10.1016/j.ecss.2014.12.015>

FIGURE AND TABLES
CHAPTER 2

Table 1. Data specifications for MODIS remotely-sensed data. Original satellite-based imagery was collected by the MODIS instrument, part of the NASA Earth Observing System, and downloaded through the NASA’S ERDP server at a temporal resolution of monthly composite, from 2003 to 2020 and at a 4 km spatial resolution as NetCDF files.

Database Title	Originator	Access	Dataset ID	Temporal range	Temporal resolution	Spatial resolution	Type	Format
SST, AQUA_MODIS, L3m.MO.SST.sst.4km, Masked, SMI, NASA GSFC OBPG, R2019.0, Global, 0.04166°,	NASA Earth Observing System	https://coastwatch.pfeg.noaa.gov/erddap/griddap/erdMH1sstdmdayR20190SQ.html	erdMH1sstdmdayR20190SQ	2003-present	Monthly Composite	4 km	Remotely sensed	NetCDF
Chlorophyll- <i>a</i> , Aqua MODIS, NPP, L3SMI, Global	NASA Earth Observing System	https://coastwatch.pfeg.noaa.gov/erddap/griddap/erdMH1chlamday.html	erdMH1chlamday	2003-present	Monthly Composite	4 km	Remotely sensed	NetCDF

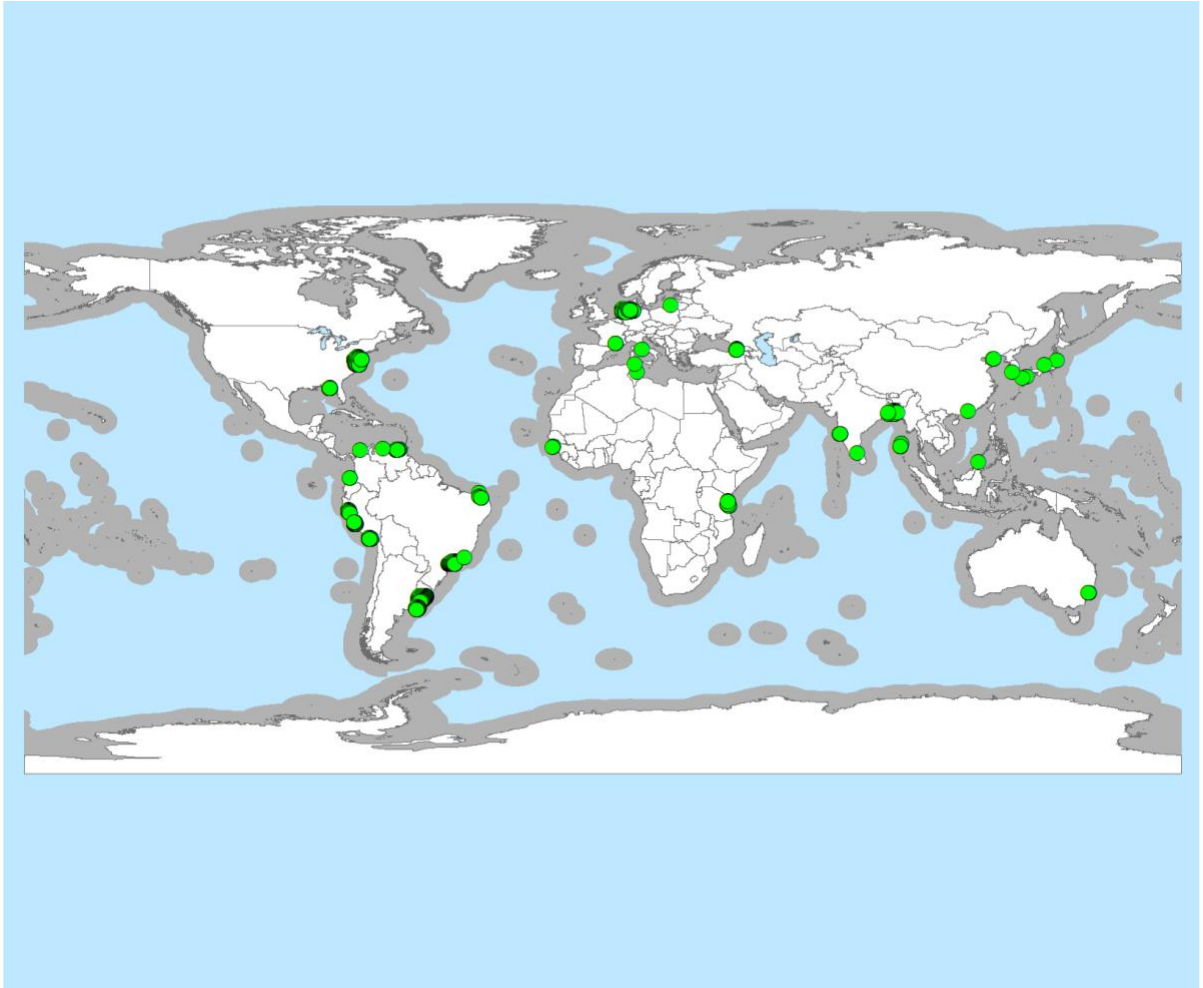


Figure 1. Study area and geolocation of *Vibrio cholerae* in coastal areas. Exclusive economic zone around the world (grey) was used to limit the satellite-derived data of seawater conditions in coastal areas. Literature occurrence records of *Vibrio cholerae* (green points).

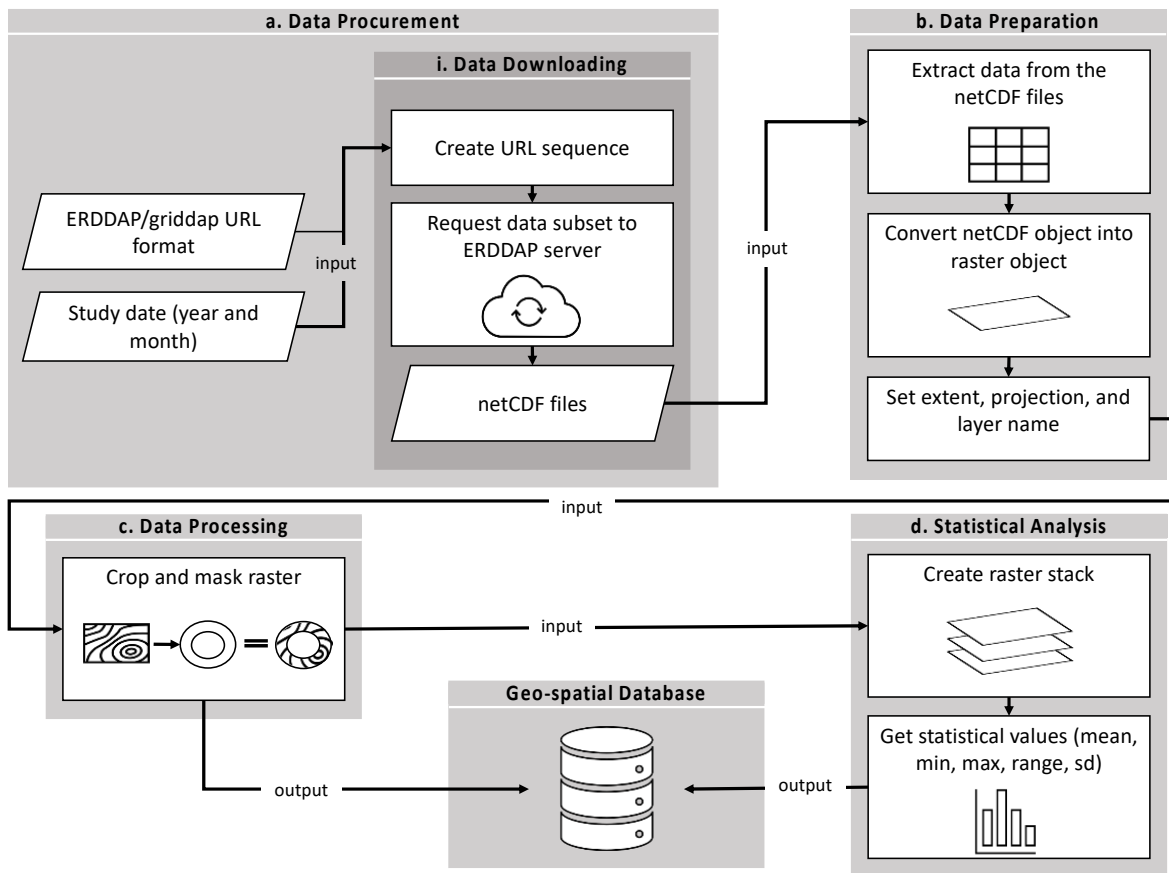


Figure 2. Database workflow diagram. (a) Remotely sensed data were downloaded from the NASA ERDDAP server in the form of NetCDF files. (b) Data were then transformed into a raster object. (c) Data were then cropped and masked to the exclusive economic zone and imported as GeoTIFF. (d) Data were analyzed to include statistical analyses and exported as raster files.

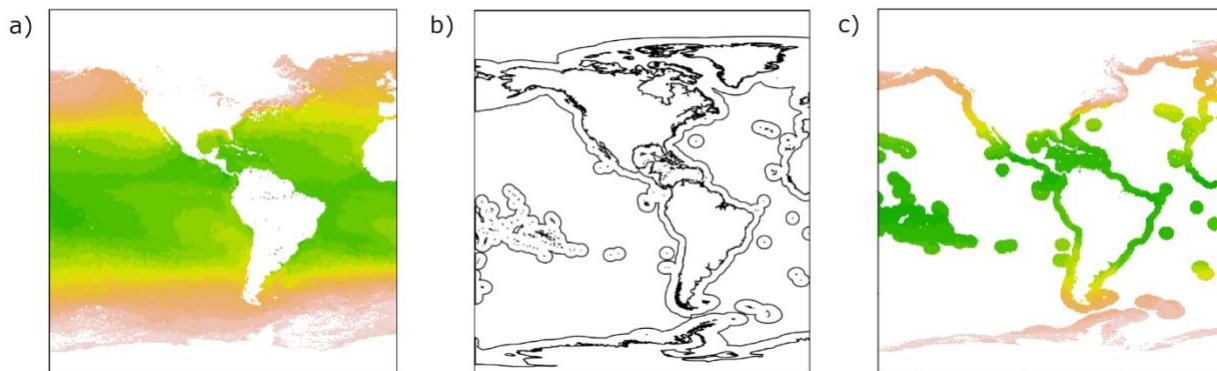


Figure 3. Data masking and cropping. Example of masking and cropping a raster. a) Raster from original NetCDF. b) Economic Exclusive Zone (solid lines). c) Raster after crop and mask.

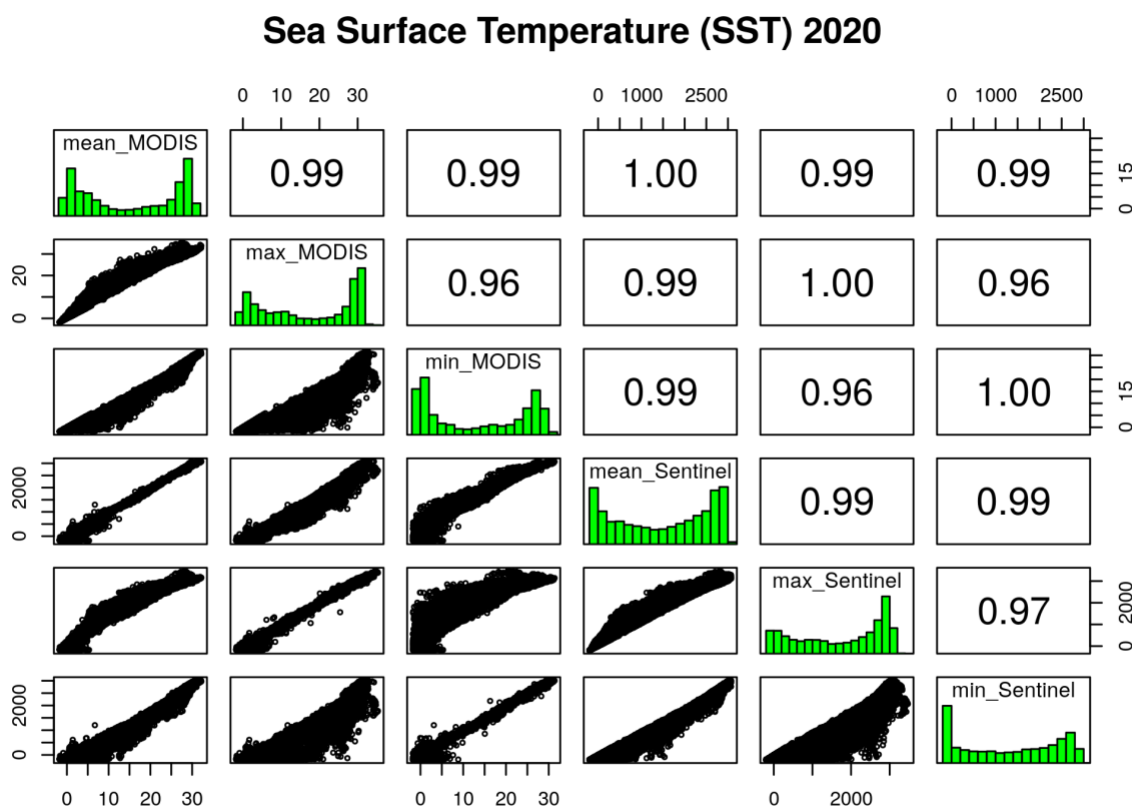


Figure 4. Sea surface temperature correlation between MODIS and Sentinel-3 data during the year 2020. Correlation for mean, minimum, and maximum, between both sensors results in a high positive correlation with a Pearson correlation coefficient of $r > 0.96$ for all scenarios.

Chlorophyll-a (Chlo-a) 2010

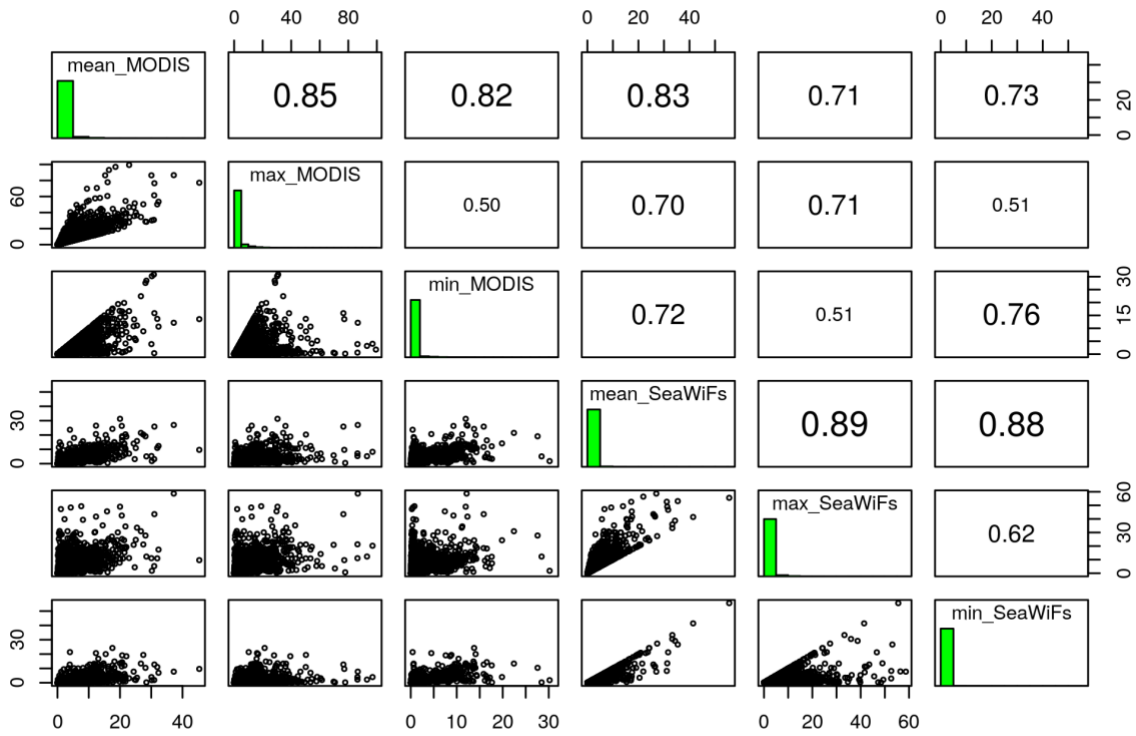


Figure 5. Chlorophyll-*a* correlation between MODIS and SeaWiFS data during the year 2010. Correlation for mean, minimum, and maximum, between both sensors results in a high positive correlation with a Pearson correlation coefficient of $r > 0.51$ for all scenarios.

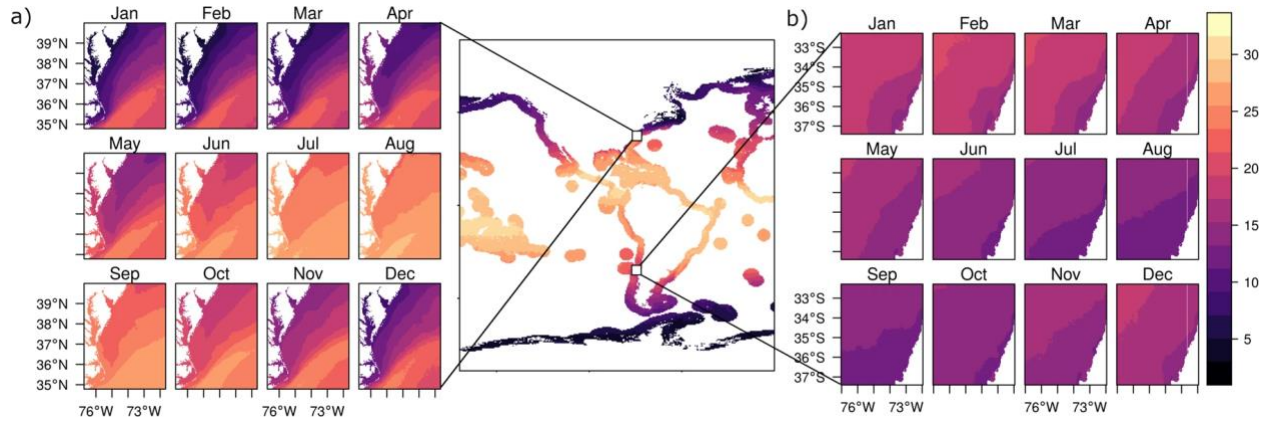


Figure 6. Sea surface temperature mean monthly values from 2003–2020. (a) Temperate zone monthly averages between the years 2003–2020 (east coast of the United States). **(b)** Subtropical zone monthly averages between the years 2003–2020 (coast of Chile).

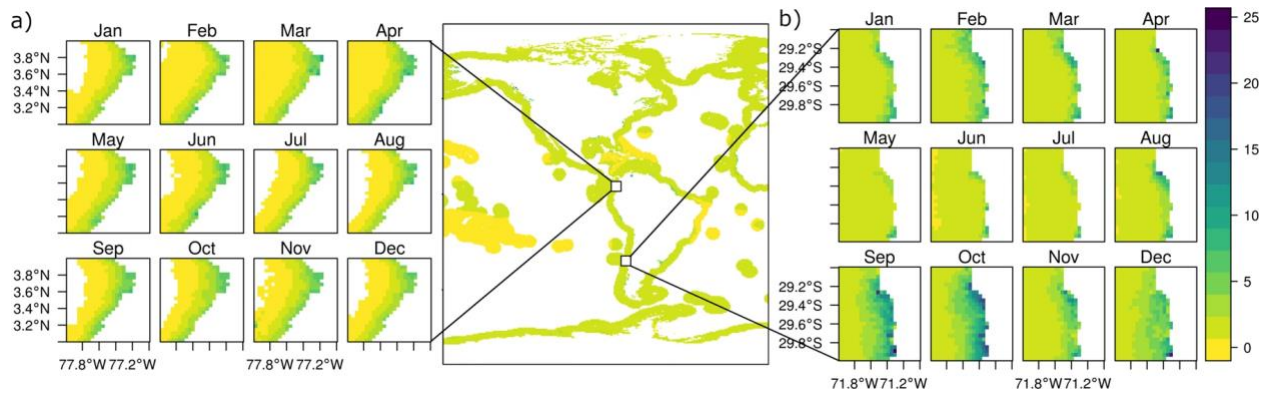


Figure 7. Chlorophyll-a mean monthly values from 2003–2020. (a) Tropical zone monthly averages between the years 2003–2020 (coast of Ecuador and Colombia). **(b)** Subtropical zone monthly averages between the years 2003–2020 (coast of Chile).

CHAPTER 3

EXPLORING CLUSTERING ALGORITHMS TO PREDICT THE REALIZED NICHE OF AQUATIC ORGANISMS

Abstract

Most infectious diseases in animals do not occur randomly. Instead, diseases in livestock and wildlife are predictable in terms of the geography, time and species affected. This, in turn, allows to quantify and trace specific environmental conditions associated with disease occurrence. Machine learning modeling is an analytical tool recently employed in spatial epidemiology to map vector-borne disease transmission, accounting for multivariate environmental conditions. Nevertheless, understanding the effects of biotic and abiotic environmental conditions on the distribution and abundance of environmentally transmitted diseases (i.e., non-vector-borne) remains in its infancy. Inadequate availability of analytical methods that analyze biotic and abiotic variables limits capacities to anticipate where directly-transmitted diseases can emerge. To solve this problem, I propose an adaptation of the density-based spatial clustering algorithm based on ecological theory to model relationships between biotic and abiotic environmental conditions as predictors of *Vibrio cholerae*, a water-borne/food-borne pathogen. This study focused on using classical ecological niche theory to implement a density-based clustering algorithm, i.e., density-based spatial clustering of applications with noise (DBSCAN), to predict the spatial distribution of *V. cholerae* in seawaters globally. This method may provide opportunities to use biotic and abiotic data to understand how global changes affect water-borne infectious diseases globally without the need for a priori assumptions of the shape of the response of organisms to biotic factors. More specifically, a revised version of the *Marble* algorithm is proposed, an adaptation of DBSCAN for realized ecological niche estimates. The assessment demonstrates that *Marble* can

accurately reconstruct and predict the global distribution of *V. cholerae*. As such, *Marble* can be used to estimate the distribution of emerging infectious diseases in aquatic ecosystems using presence-only data, accounting for abiotic and biotic variables, and without the need for delimitation of study areas.

Introduction

Vibrio cholerae is an aquatic bacterium with evolutionary origins in coastal sea waters and causes four million human infections annually (Ali et al., 2012). The transmission of *V. cholerae* depends on many environmental and social factors related to the risk of exposure and vulnerability of the population affected (R. R. Colwell, 1996; E. K. Lipp et al., 2002). Understanding the epidemiology of *V. cholerae* and the environmental drivers of transmission using ecological niche modeling may help improve cholera control and prevention. Ecological niche modeling (ENM) uses computer algorithms, including machine learning techniques like clustering algorithms, to predict the distribution of a species across geography, space, and time using environmental data (Escobar, 2020b; Escobar & Morand, 2020). Clustering analysis is a machine-learning technique that forms the association of a collection of objects based on their similarity. Clustering analysis classifies records such that records with similarities are grouped. These groups and associations are called clusters (Diday & Simon, 1976). Here, I investigate a cluster analysis, an unsupervised machine learning technique that aims to find patterns (e.g., sub-groups, size of each group, and common characteristics), to group similar records in environmental space as a proxy of a species' realized ecological niche (Diday & Simon, 1976; Lee, 1981).

A cluster can be defined as an area (two dimensions) or volume (three or more dimensions) of density in the environmental space (**E**-space) where records have similar properties and features (Figure 1) (Diday & Simon, 1976; Qiao, Lin, et al., 2015). Cluster analysis can be used as a method of unsupervised learning, given that clustering algorithms can define patterns from a dataset without reference to known or labeled outcomes. Unsupervised learning can be used to discover the underlying structure of the data without a *priori* knowledge of distribution (Barlow, 1989), making this approach appealing for studies with limited knowledge of the biology and ecology of organisms (e.g., pathogen species causing emerging infectious diseases, cryptic species, rare or endangered species). Clustering analysis is a common technique in statistical analysis, with applications that include pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics (Nizar et al., 2004) but has been rarely used in ecological niche modeling (Escobar 2020).

The notion of the degree of similarity (or dissimilarity) between the individual objects being clustered is central to cluster detection. There are four major categories for cluster analysis: partitioning, hierarchical, grid-based methods, and density-based methods (Mann & Navneet, 2013). This chapter focuses on density-based clustering methods to estimate the environments occupied by a species. Clustering algorithms using environmental data are used in ENM to predict the distribution of a species across space and time. Identifying clusters of occurrences in the **E**-space allows for characterizing the species' environmental tolerances. A popular clustering method in ENM include is one-class support vector machines (SVM; Blonder et al., 2014), which is used to estimate hypervolumes and is available in the hypervolume package in R.

Hutchinson first proposed the n -dimensional hypervolume to quantify species niches (Hutchinson, 1957). A set of n variables representing biologically important and independent axes

are identified, and the hypervolume is defined by a set of points within the n -dimensional space that reflects suitable values of each variable (e.g., climate, topography; R. K. Colwell & Rangel, 2009; Hutchinson, 1957; Peterson et al., 2011). Hutchinson's n -dimensional hypervolume concept is used for the interpretation of ecological niches as geometric shapes and has provided a foundation for researchers across different fields of ecology and evolution (Garnier et al., 2015; McNerny & Etienne, 2012; Peters, 1991). The n -dimensional hypervolume concept has been applied in a range of research areas, including niche-based ecology (e.g., Araújo & Guisan, 2006; Jackson & Overpeck, 2000; Soberón & Nakamura, 2009), functional ecology (e.g., Lamanna et al., 2014; Moles et al., 2007; van Kleunen et al., 2010), and community phylogenetics (e.g., Pavoine et al., 2013; Webb et al., 2002; Wiens & Graham, 2005).

A few methods have been developed to measure and estimate Hutchinson's n -dimensional hypervolume (Blonder et al., 2014; Peterson et al., 2011). The general mathematical problem is how to best estimate a hypervolume from a set of observations when disparate predictors are used, e.g., biotic and abiotic (Araújo & Rozenfeld, 2014; Simões & Peterson, 2018). Ideally, a hypervolume estimation procedure should: (1) directly delineate the boundaries of the hypervolume; (2) not assume a fixed distribution of observations; (3) account for disjunction or holes; (4) not be sensitive to outlier points; (5) not assume a priori species responses to the biotic and abiotic conditions and (6) produce a bounded result (i.e., not predict infinite values) (Blonder et al., 2014; Peterson et al., 2011; Qiao, Escobar, & Peterson, 2017; Qiao, Soberón, et al., 2015). Ideally, such methods should also be computationally efficient.

In 2014, Blonder et al. formalized the concept of hypervolume in ENM as an R package called *hypervolume*. They proposed the use of the multidimensional KDE procedure to directly estimate the n -dimensional hypervolume in environmental space from a set of geographic observations.

This method takes the environmental conditions enclosed by a contour of a kernel density estimate, which can account for non-normal, rotated, and holey data, including outliers, therefore addressing the most important procedures for generating an n -dimensional hypervolume (Blonder et al., 2014). A caveat, however, is that KDE defines the clusters based on user-specified parameters.

Here, I propose a revised version of the *Marble* algorithm, an adaptation of density-based spatial clustering with noise (DBSCAN) for realized ecological niche estimates first introduced by Qiao, Lin, et al. in 2015. I explored the use of DBSCAN to estimate the realized niche of a species using multiple abiotic and biotic variables (Hahsler et al., 2019a; Qiao, Lin, et al., 2015; Sander et al., 1998). DBSCAN is an alternative hypervolume method originally proposed in 1996 by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu and is one of the most popular density-based clustering algorithms (Sander et al., 1998). *Marble* inference of the number of clusters is data-driven and can reconstruct clusters of arbitrary shapes. *Marble* searches for clusters (i.e., a set of points concentrated in some dense space of the whole set) of species' occurrences in \mathbf{E} -space, which will result in a distribution map when projected to \mathbf{G} -space (Qiao, Lin, et al., 2015).

Methods

Environmental Data

I used the annual average raster data procured in *Chapter 2*, which contains fine-scale (~4 km) global processed remotely sensed data of monthly sea surface temperature and chlorophyll-a adjusted to the exclusive economic zone (country limits 200 miles off the coast). The seawater conditions dataset expands over 18 years (2003–2020). Sea surface temperature ($^{\circ}\text{C}$) was used as a proxy for seawater temperature (i.e., abiotic predictor), and chlorophyll-a ($\text{mg} * \text{m}^{-3}$) as a proxy

for phytoplankton density in global coastal waters (i.e., biotic predictor), recognizing that phytoplankton serves as a proxy for subsequent zooplankton blooms carrying the copepod vector of vibrios, including *V. cholerae* (R. R. Colwell & Huq, 2014). To reduce the dimensionality and multicollinearity of the environmental predictors, I ran a principal component analysis (PCA), choosing only the first three principal components that explained 99% of the overall variance (90%, 6%, 3%, respectively).

Occurrence Data

I used *V. cholerae* as case study to test the functionality of the *Marble* algorithm in a marine ecosystem at global scale. To recover some of the *V. cholerae* occurrence records, I developed a structured metanalysis review of the scientific literature reporting the species *V. cholerae* occurrence records, following Escobar et al. 2015 (Chapter 2). I searched and captured *V. cholerae* records addressing the following criteria: (i) records from seawaters in coastal areas, (ii) species corroborated in a laboratory, and (iii) records with metadata allowing detailed (locality level) geolocation (Figure 1 in Chapter 2). Before using the data for model calibration and testing, I filtered records to one per pixel to reduce spatial autocorrelation and divided them into subsets for calibration and evaluation following an 80/20 split.

DBSCAN Algorithm Implementation

Implementation of the *Marble* algorithm and subsequent analysis were executed using the statistical software R (R Core Team, 2022). The main *Marble* algorithm was coded in C++

language and implemented in R using the *Rcpp* package. *Rcpp* is a package that allows seamless integration of R and C++ languages (Eddelbuettel, 2013; Eddelbuettel & Balamuta, 2018; Eddelbuettel & François, 2011). The implementation of the algorithm was divided into three stages: 1) DBSCAN algorithm; 2) implementation of ecological niche theory, and 3) analysis of parameters sensitivity. The first two stages are associated with the development of the *Marble* algorithm to address n -dimensional hypervolume theory, while the last stage focuses on the evaluation of the parameters for niche estimation. *Marble* requires two parameters: the minimum number of points (*minPts*) and a distance measure (epsilon, ϵ). The goal of these parameters is to partition the presence data in environmental space into three classes of points: (i) core point, (ii) boundary point, and (iii) noise or outliers (Figure 2).

The parameter *minPts* specifies a minimum number of data points that can form a cluster. ϵ is the distance measure that will be used to locate the presence points in the **E**-space clustered together. When using Euclidean distance, ϵ can be thought of as the maximum radius from a target presence point in **E**-space that broadcasts out in all environmental directions, forming a perimeter (i.e., the neighborhood) around the target presence point.

The selection of the parameters, *minPts* and ϵ , can be data-driven or defined by the user based on the study aims and the level of knowledge of the organism. For instance, the literature recommends following these criteria for selecting an appropriate value for *minPts*: generally, (1) the larger the data set, the larger the value of *minPts* should be, and (2) if the dataset is noisier, a large value of *minPts* should be selected, (3) *minPts* should be greater than or equal to the dimensionality of the data set, and (4) when the data are analyzed in more than two dimensions, the most appropriate value for *minPts* is $minPts = 2 * dim$, where *dim* is the dimensions of a particular data set (i.e., number of predictor variables; Sander et al., 1998, Ester et al., 1996).

Furthermore, the *Marble* algorithm can either take a user-defined value for ϵ or automatically estimates the appropriate value for ϵ when an omission threshold is specified as a measure of the error in the presence-only dataset. That is, the omission threshold represents a level of distrust in the data being used to calibrate the model measured as the proportion of the calibration presence points to be excluded based on the reliability of the data and mirrors the level of alpha error defined by the users in probability statistics (e.g., $\alpha=0.05$). The omission threshold can also be thought of as one minus sensitivity threshold (i.e., the complement of the sensitivity threshold), which can be interpreted as one minus the confidence in the data being used in calibration (Peterson et al., 2008).

Moreover, if *minPts* and omission threshold have been specified by the user, the *Marble* algorithm estimates the appropriate value for ϵ utilizing the kth-nearest-neighbor algorithm (kNN). kNN is a machine learning classification algorithm that helps determine for each presence point in **E**-space the closest neighboring presence point and records the distance between these points. Once the distance to the nearest neighbor for each presence point in **E**-space is determined by kNN, the ϵ is selected by sorting these resulting distances in increasing order. From the kNN calculated distances, I selected a subsample of the distances that correspond to one minus omission threshold (i.e., sensitivity threshold). For example, in a sample size of 100, I calculate 100 unique distances. If an omission threshold of 5% error is selected, only the top 95% (i.e., the first 95) of the distances will be selected, and the maximum value of those distances will represent ϵ .

During model calibration using *Marble*, presence points in **E**-space are partitioned into one of three classes of points: core, border, or noise points. A point, p , will be labeled as core point, q , if the number of points within distance ϵ (i.e., neighboring points) is greater or equal to the *minPts* threshold (Figure 3a). Contrarily, a point, p , which is within ϵ distance from a core point, q , and has neighboring points less than *minPts* within distance ϵ , then p is defined as a border point.

Border points can join a cluster but cannot extend it further. If a point in \mathbf{E} -space is neither a core nor a border point, it is defined as a noise point (i.e., outlier). Each of the points will belong to a unique cluster, and once a point is assigned to a cluster, it can no longer belong to another cluster. All presence points in \mathbf{E} -space belonging to that specific cluster form a neighborhood (N_ϵ). In other words, a neighborhood is defined as,

$$N_\epsilon(p) : \{q \mid d(p, q) \leq \epsilon\} \quad (1)$$

Two other key concepts in the building of the *Marble* algorithm are the concept of density reachability and connectivity. There are two distinctions for reachability, for directly and not directly reachable density points (Figure 3b). A point p is called directly-density reachable if it has a core point in its neighborhood (N_ϵ). In other words, a core point can be found within a particular distance (ϵ) from another p . Satisfying the following conditions (Qiao, Lin, et al., 2015):

$$p \in N_\epsilon(q) \text{ and } N_\epsilon(q) \geq \text{minPts} \quad (2)$$

where a presence points in \mathbf{E} -space, p , is density reachable from point q (i.e., a neighboring presence point labeled as a core point) if it is connected through a series of core points (Figure 3c). For instance, if there is a chain of points p_1, \dots, p_n and $p_1 = q$ and $p_n = p$, such that p_{i+1} is directly density-reachable from p_i , the point p is defined as ‘density-reachable from q .’ In terms of connectivity, a point p is said to be density-connected to a point q if there is a core point, o , that chains them together, and both p and q are density reachable from o (Figure 3d).

Although the statistical software R as a package called *dbscan* with defined parameterizations to analyze data frames (Hahsler et al., 2019b), I developed my own functions coded in C++ software to have more control over the model performance, input requirements, and outputs analyzed in \mathbf{E} -space and projecting the model outputs to \mathbf{G} -space. This version of DBSCAN

implemented in *Marble* followed the protocol illustrated in Figure 4. Because *Marble* intends to mitigate environmental interpolations, *Marble* was designed to not assume particular species response curves for specific predictors, which is a desirable feature when modeling the effects of abiotic or biotic variables with unknown influence over species distributions and fitness.

Based on the above definitions, the realized ecological niche is feasible with *Marble*, i.e., abiotic and biotic conditions where organisms occur are estimated as hypervolumes in **E**-space (Qiao, Lin, et al., 2015). Hypervolumes can then be defined as multivariate clusters of abiotic and biotic conditions where species actually occur and consistently survive. Presence points in **E**-space classified in the point classes of core and border points by *Marble* are considered within the realized niche of the given species; otherwise, presence points are just considered noise. The final set of clusters created in **E**-space is then projected to **G**-space (Figure 5c). If more than one cluster is found, *Marble* carefully identifies and labels clusters as independent units, which may be a desirable feature for studies to assess niche conservatism vs. differentiation, niche expansion vs. niche filling, and metapopulation and lineage-level assessments. Here, to simplify the output, the resulting raster is transformed into binary outputs, but continuous values can be generated via permutations to generate an ensemble model. In binary models, the labeling of independent clusters becomes irrelevant, and the only important information is whether that geographic pixel belongs to the realized niche of the species. Therefore, the resulting raster will only have binary values (belongs to the realized niche or not, based on the observed data; Figure 6 e and f). Binary outputs facilitates to compare ENM among *Marble* parameters, environmental variables, periods, or species. Nevertheless, future research could include assessing the role of clustering position, size, and shape to address demographic, biogeographic, and evolutionary questions.

Sensitivity Analysis

I performed a sensitivity analysis to evaluate the effects of parameter values on the final realized niche estimates. For the sensitivity analysis, I created an extended matrix of all possible parameter combinations based on previously agreed-on sets of values for each parameter. That is, using values for *minPts* with a lower bound of two and upper bound of four times the dimension of the data (e.g., for a three-dimensional data set, upper bound of *minPts*=12; Eq. 3). For the omission threshold, I determined the best range of values to be between 5% to 10% confidence error (Eq. 4). In this case,

$$\text{minPts} = \{x, x_{i-1} + 2, \dots \mid 2 \leq x \leq 12\} \quad (3)$$

$$\text{omission threshold} = \{x, x + 0.01, \dots \mid 0.01 \leq x \leq 1\} \quad (4)$$

Furthermore, the presence points in **E**-space were split using 5-fold cross-validation. K-fold cross-validation is a resampling method that uses different portions (i.e., *k* number of folds) of the data to evaluate and calibrate a model on different iterations (Refaeilzadeh et al., 2009). Therefore, a 5-fold cross-validation will split the data into five different portions, and a different individual portion will be used for model evaluation in each iteration, resulting in five different evaluation and calibration sets of presence points in **E**-space. The cross-validation sets combined with the parameter combination matrix resulted in a total of 330 model replications (i.e., 66 replications per fold).

Using the calibrated model in **E**-space, I evaluated the relationship and effect between each of the parameters (i.e., *minPts*, ϵ , and omission threshold) and the model behavior and response in terms of the number of independent identified clusters as a proxy of model fit. Calibrated models were then projected into **G**-space and evaluated using the evaluation presence points in **E**-space

(i.e., presence points not used for calibration) for each fold in terms of the proportion of predicted suitable area, omission rate, and cumulative binomial probability (CBP; test based on the omission threshold and the proportion of predicted suitable area) (Anderson et al., 2003; Escobar et al., 2015b, 2018; Peterson, 2012). Omission threshold and CBP allowed us to measure Type I error (i.e., false presence) for every model replication, and determine whether evaluation presence points in **G**-space are predicted better than by chance, given the proportion of area predicted as suitable (Anderson et al., 2003; Escobar et al., 2015b, 2018; Peterson, 2012). I also used a modification of the conventional receiver operating characteristic (ROC) approach more suited for ENM (Peterson et al., 2008). A conventional ROC analysis involves plotting sensitivity (i.e., true positive rate) against $1 - \text{specificity}$ (i.e., false positive rate). The plot in ROC space of sensitivity versus $1 - \text{specificity}$ displays how well an algorithm classifies instances as the threshold changes. In ENM, threshold changes mean that the area predicted as suitable also changes. Regions in the ROC space near sensitivity=1 and proportion of area predicted as suitable=0, will represent the better performing models (Peterson et al., 2008).

Results

Here, I implemented *Marble*, a machine learning ENM method that combines DBSCAN with ecological niche theory to estimate the realized niche of organisms by linking environmental and geographic dimensions in R (R Core Team, 2022), which I implemented from a modified DBSCAN algorithm in C++. I used *V. cholerae* as a case study to understand how this novel method performs to estimate the distribution of an aquatic, non-free-living organism (i.e., parasite) on a global scale. This implementation of clustering analysis and ecological niche theory was termed *Marble* to retain the original idea proposed by Qiao, Lin, et al. (2015), who implemented

an early version of DBSCAN to estimate species' potential distribution using abiotic variables in a web-based platform that has been discontinued. In this study, I also evaluated the sensitivity of *Marble* parameters on ecological niche model outputs.

Parametrization General Results

Each of the three *Marble* parameters (i.e., ϵ , *minPts*, omission threshold) impact ENM estimates. As expected, given the way ϵ is calculated (i.e., using *minPts* and omission threshold), model results show that ϵ and sensitivity threshold (i.e., one minus omission threshold) are positively correlated (Figure 7). In other words, as the omission threshold increases, so will ϵ . Similarly, *minPts* and ϵ are positively correlated. Although as *minPts* is increased, the values of ϵ do not increase in a linear fashion; they instead resemble a positive logarithmic curve. That is, there is a point where an increase in *minPts* would correspond to a minimal change in ϵ values (i.e., ϵ reaches an equilibrium after which there is no considerable change in the model) (Figure 7A). Furthermore, both *minPts* and ϵ (and sensitivity threshold) have an impact on the number of clusters to be detected. The higher *minPts* and ϵ , the lower the number of resulting clusters. Lastly, based on ROC, sensitivity versus commission rate (i.e., false positives rate) curve (grouped by *minPts*), I identify that the model can predict better than random in all model replications to predict an aquatic organism (i.e., *V. cholerae*) (Figure 8).

Parameter Selection E-space

A higher number of clusters in *Marble* represents a tight fit of the realized niche model to the available occurrence data and may represent a lack of transferability (i.e., the limited ability of the

model to predict new occurrences beyond the calibration area). A lower number of clusters may encompass more of the occurrence and environmental data, which could result in model overfitting. For ϵ , results show that, in general, small values should be preferred, and only a fraction of the points should be within ϵ distance of each other to avoid overfitting. A larger ϵ will produce broader clusters and a broader realized niche closer to a species fundamental niche, and a smaller ϵ will build smaller clusters and a narrower realized niche closer to the species presence records. The range of values for ϵ is dependent on the specific presence of data in **E**-space being used. That is, the lower bound for ϵ values correspond to the minimum Euclidean distance between any two points in the data set. Conversely, the upper bound of the ϵ values would correspond to the maximum Euclidean distance. If ϵ is selected to be less than the lower bound (i.e., underfitting), no clusters will form in **E**-space; if larger than the upper bound, the resulting model will be a single cluster encompassing all data presence points in **E**-space (i.e., overfitting). Determining the range of values of ϵ in **E**-space, however, proves to be complex and not intuitive. Thus, incorporating the omission threshold as an optional parameter for *Marble* would allow the users to bypassing the need to specify a specific ϵ value, allowing for a more intuitive parameter selection for the models at the cost of introducing potential bias by the user.

In general, a value of omission threshold is preferred (i.e., omission threshold > 0) to include only a fraction of the presence points in **E**-space within ϵ distance from each other, which will reduce extrapolation and interpolation of models. Results show that, for the *V. cholera* data, selecting a value of omission threshold greater than 2.5% and less than 5% would result in a balance between over and underfitting the data (i.e., smaller ϵ values; Figure 7).

Similarly, to ϵ and omission threshold, a higher *minPts* will produce more robust clusters and a broader realized niche, and a smaller *minPts* will build smaller clusters and a narrower realized

niche, with more presence points in **E**-space classified as noise. The preference for *minPts* should lean toward selecting values twice or at a maximum of three times the number of predictor variables (i.e., the dimensionality of the data). Selecting *minPts* within this range (i.e., $minPts = 2 * dim$ to $3 * dim$) also results in low computational time (Figure. 11c). Although, in the case of a larger dataset with thousands of occurrences in **E**-space, for datasets with abundant nosier occurrences, and occurrence data that contains multiple duplicated records in **E**-space a selection of a larger *minPts* value may be necessary.

Geographic Space

In terms of the projection of potential distribution of *V. cholerae*, i.e., **G**-space, which each omission threshold, *Marble* identified different clusters in **E**-space that were only defined as suitable (i.e., belongs to the realized niche, as shown in Figure 6c and d). For omission thresholds greater than 3%, *Marble* consistently identified the tropics and subtropics areas as suitable. For omission thresholds, less than 3% of the tropic and subtropic area shows suitability with the addition of the temperate zones. Lastly, the poles are only shown as suitable when the omission threshold is less than 1%. Based on CBP, all models, except for those with an omission threshold equal to 0%, result in a p -value < 0.001 , indicating all models are able to predict better than by random. The statement that all models predict better than random, is also supported by the ROC curve (Figure 8). Models in the ROC evaluation fell above the 1:1 sensitivity versus proportion of area predicted as suitable line in all parameter combinations, indicating better prediction than random; with $minPts=1$ having the overall better performance in terms of proportion of predicted area versus sensitivity.

Discussion

In this study, I performed a detailed parametrization of the *Marble* algorithm to identify the adequate parameters for the algorithm and determine the capabilities of *Marble* to reconstruct the realized niche of an aquatic organism with limited surveillance in marine waters, *V. cholerae*. I found that *Marble* was most sensitive to the user-defined omission threshold (i.e., level of error in the presence-only data), followed by ϵ , and lastly, *minPts*. For the *V. cholerae* presence data, *minPts* had little to no effect on the output of the models in **G**-space and only differed in time complexity (i.e., lower *minPts* higher time complexity; Figure 9). Overall, time complexity was low, averaging 40 min for long-performing models (Figure 9). This study introduces a new low-complexity ENM tool for the reconstruction of species' realized niche that allows for the integration of abiotic and biotic factors, and it is not sensitive to geographic extents.

Many studies have modeled the niches of species. Despite that, most fail to describe the specific niche being modeled (i.e., fundamental or realized niche) (Escobar & Craft, 2016). Since the conception of the idea of *Marble*, the focus was to reconstruct the realized niche of species. The realized niche of species is defined to be the set of all environmental states that would permit a species to exist in the presence of restrictive factors (i.e., abiotic and biotic factors; Peterson et al., 2011).

A key component of *Marble* is that it is being conceived for the practical and theoretically sound inclusion of both biotic and abiotic variables in the model. The addition of biotic factors to ENM has shown contrasting results in the literature, affecting predictive power and model complexity positively, negatively, or neutrally (but usually positively; Araújo & Luoto, 2007). It is known that species distributions can be summarized by three limiting factors: biotic interactions

(**B**), abiotic conditions (**A**), and parts of the world that have been accessible to the species via dispersal over relevant periods of time (i.e., movement capacities; **M**) (**BAM** framework; Soberón & Peterson, 2005). Nevertheless, historically, climatic variables (**A**) have been the focus in ENM, whereas biotic (**B**) interactions are usually not considered for model calibration (Pearson & Dawson, 2003; Soberón, 2007; Soberón and Nakamura, 2009). Another major challenge in ecological niche modeling is determining the geographic extent (**M**) to use during the model calibration, as it has been proven to influence the outcome of traditional correlative models (Barve et al., 2011) in terms of model parametrization (VanDerWal et al., 2009), validation (Lobo et al., 2008), and comparisons (Warren et al., 2008). *Marble*, however, due to its algorithmic framework, focuses on the presence point in **E**-space alone, and it does not require background or absence data to find environmental clusters.

Recent evidence in the literature suggests that the abiotic components of the species' fundamental niche are more similar among terrestrial organisms than typically expected (Araujo et al., 2013), leading to the conclusion that species distributions might often result from biotic factors (Araujo & Rozenfeld, 2014). It has been argued that biotic interactions determine whether a species thrives or withers in each environment but that the spatial effects associated with these interactions are lost on a broad scale (Whittaker et al., 2001, Pearson & Dawson, 2003; McGill, 2010; Araujo & Rozenfeld, 2014). The idea that distributions of species on the geographic extent and coarse resolution are rarely affected significantly by biotic factors has been termed the 'Eltonian Noise Hypothesis' (ENH; Soberón & Nakamura, 2009). On the other hand, biotic interactions are known to affect species' spatial distribution via mechanisms such as predation, competition, and mutualism (Anderson, 2017; Simoes & Peterson, 2018). In the case of parasites, biotic interactions are imperative, e.g., a mosquito needs to feed on another organism to survive.

Therefore, biotic interactions can also be positive. For example, the presence of prey (i.e., food resources), hosts, and vectors, that facilitate the occurrence of the species in the determined areas.

Modeling exercises using abiotic features consider species distribution inside the Grinnellian niche, which comprises the environmental variables (Grinnell, 1917). Even so, some studies suggest the important role of biotic interactions as a predictor of spatial distribution even on a large scale (Meier et al., 2011; Araújo et al., 2014). By including variables representing biotic interactions, it assumes a Hutchinsonian view of the niche (i.e., composing the realized niche) (Hutchinson 1957). Mapping the realized niche by incorporating the knowledge of a species' biology and ecology, comparatively with the fundamental niche, may lead to more accurate and refined predictions of species distribution models mainly by avoiding commission errors (false positive). This chapter found that biotic and abiotic predictors could be used to model the potential distribution of *V. cholerae*.

Ecological niche models of aquatic ecosystems are more complex and less explored in the literature than terrestrial ecosystems (Melo-Merino et al., 2020). Traditional models aim to reconstruct the distribution of disease-causing bacteria and generally focus on the reconstruction of the geographic distribution of the host or vector species (i.e., the species that carries the pathogen) rather than directly modeling the geographic distribution of the pathogen (Melo-Merino et al., 2020). Here, I found that *Marble* can directly model pathogens providing an opportunity to further the knowledge of directly transmitted diseases. Aquatic organisms, like *V. cholerae*, are influenced by biotic (e.g., plankton) and abiotic (e.g., temperature) factors surrounding their aquatic environment (R. R. Colwell, 1996; R. R. Colwell & Huq, 2014; E. K. Lipp et al., 2002). Assuming a Hutchinsonian niche, *Marble* was able to estimate the realized niche of *V. cholerae* under different parameters' scenarios. That is, *Marble* found, even in high omission threshold

scenarios (i.e., high error rate), that coastal countries like Ecuador, Peru, Uruguay, Northern Argentina, Eastern United States, and countries in the North Sea (i.e., Denmark, Norway, Sweden, Finland), Western Africa, Madagascar, Mozambique, and countries by the Coral Sea (i.e., Easter Australia, Fiji) have suitable conditions for *V. cholerae* to survive.

Currently, there are limited tools to make estimates of potential distribution that allow the use of abiotic and biotic predictors. The use of both abiotic and biotic predictors is the future of ENM (Ashraf et al., 2021; Simões & Peterson, 2018). *Marble* could help differentiate between the fundamental niche and the realized niche, avoiding generalizations of ecological niche models without a definition of the niche being modeled. Future work may include the analysis of virtual species or other species, ideally, one from which the environmental and geographic distributions are known. Furthermore, in the way *Marble* is currently implemented, predictor variables must be transformed and normalized using PCA to enable the use the measurement of the Euclidean distance. The addition of different distance measures (e.g., Mahalanobis distance) in the future may allow the calculation of distances between uncorrelated predictors and predictors with different units without the need for transformation. Although, advantages of *Marble* are that it does not require background or presence data, and it does not require any delimitation of **M** for model calibration.

Reference

- Ali, M., Lopez, A. L., Ae You, Y., Eun Kim, Y., Sah, B., Maskery, B., & Clemens, J. (2012). The global burden of cholera. *Bulletin of the World Health Organization*, *90*(3), 209–218. <https://doi.org/10.2471/BLT.11.093427>
- Anderson, R. P., Lew, D., & Peterson, A. T. (2003). Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling*, *162*(3), 211–232. [https://doi.org/10.1016/S0304-3800\(02\)00349-6](https://doi.org/10.1016/S0304-3800(02)00349-6)
- Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. <https://doi.org/10.1111/j.1365-2699.2006.01584.x>

- Araújo, M. B., & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, 37(5), 406–415. <https://doi.org/10.1111/j.1600-0587.2013.00643.x>
- Ashraf, U., Chaudhry, M. N., & Peterson, A. T. (2021). Ecological niche models of biotic interactions predict increasing pest risk to olive cultivars with changing climate. *Ecosphere*, 12(8). <https://doi.org/10.1002/ECS2.3714>
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1(3), 295–311. <https://doi.org/10.1162/neco.1989.1.3.295>
- Blonder, B., Lamanna, C., Violle, C., & Enquist, B. J. (2014). The n -dimensional hypervolume. *Global Ecology and Biogeography*, 23(5), 595–609. <https://doi.org/10.1111/geb.12146>
- Colwell, R. K., & Rangel, T. F. (2009). Hutchinson’s duality: The once and future niche. *Proceedings of the National Academy of Sciences*, 106(Supplement_2), 19651–19658. <https://doi.org/10.1073/pnas.0901650106>
- Colwell, R. R. (1996). Global climate and infectious disease: The cholera paradigm. *Science*, 274(5295), 2025–2031. <https://doi.org/10.1126/science.274.5295.2025>
- Colwell, R. R., & Huq, A. (2014). Vibrios in the environment: Viable but nonculturable *Vibrio cholerae*. In *Vibrio cholerae and Cholera* (pp. 117–133). ASM Press. <https://doi.org/10.1128/9781555818364.ch9>
- Diday, E., & Simon, J. C. (1976). *Clustering analysis* (pp. 47–94). https://doi.org/10.1007/978-3-642-96303-2_3
- Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer New York. <https://doi.org/10.1007/978-1-4614-6868-4>
- Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, 72(1), 28–36. <https://doi.org/10.1080/00031305.2017.1375990>
- Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8). <https://doi.org/10.18637/jss.v040.i08>
- Escobar, L. E. (2020). Ecological niche modeling: An introduction for veterinarians and epidemiologists. *Frontiers in Veterinary Science*, 7. <https://doi.org/10.3389/fvets.2020.519059>
- Escobar, L. E., & Craft, M. E. (2016). Advances and limitations of disease biogeography using ecological niche modeling. *Frontiers in Microbiology*, 07. <https://doi.org/10.3389/fmicb.2016.01174>
- Escobar, L. E., & Morand, S. (2020). Disease ecology and biogeography. *Frontiers in Veterinary Science, Special Issue*. <https://www.frontiersin.org/research-topics/12035/disease-ecology-and-biogeography>
- Escobar, L. E., Qiao, H., Cabello, J., & Peterson, A. T. (2018). Ecological niche modeling re-examined: A case study with the Darwin’s fox. *Ecology and Evolution*, 8(10), 4757–4770. <https://doi.org/10.1002/ece3.4014>
- Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015). A global map of suitability for coastal *Vibrio cholerae* under current

- and future climate conditions. *Acta Tropica*, 149, 202–211. <https://doi.org/10.1016/j.actatropica.2015.05.028>
- Garnier, E., Navas, M.-L., & Grigulis, K. (2015). *Plant functional diversity*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198757368.001.0001>
- Hahsler, M., Piekenbrock, M., & Doran, D. (2019a). dbSCAN: Fast density-based clustering with R. *Journal of Statistical Software*, 91. <https://doi.org/10.18637/jss.v091.i01>
- Hahsler, M., Piekenbrock, M., & Doran, D. (2019b). dbSCAN: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1). <https://doi.org/10.18637/jss.v091.i01>
- Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0), 415–427. <https://doi.org/10.1101/SQB.1957.022.01.039>
- Jackson, S. T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, 26(S4), 194–220. <https://doi.org/10.1017/S0094837300026932>
- Lamanna, C., Blonder, B., Violle, C., Kraft, N. J. B., Sandel, B., imova, I., Donoghue, J. C., Svenning, J.-C., McGill, B. J., Boyle, B., Buzzard, V., Dolins, S., Jorgensen, P. M., Marcuse-Kubitza, A., Morueta-Holme, N., Peet, R. K., Piel, W. H., Regetz, J., Schildhauer, M., ... Enquist, B. J. (2014). Functional trait space and the latitudinal diversity gradient. *Proceedings of the National Academy of Sciences*, 111(38), 13745–13750. <https://doi.org/10.1073/pnas.1317722111>
- Lee, R. C. T. (1981). Clustering analysis and its applications. In *Advances in Information Systems Science* (pp. 169–292). Springer US. https://doi.org/10.1007/978-1-4613-9883-7_4
- Lipp, E. K., Huq, A., & Colwell, R. R. (2002). Effects of global climate on infectious disease: the cholera model. *Clinical Microbiology Reviews*, 15(4), 757–770. <https://doi.org/10.1128/CMR.15.4.757-770.2002>
- Mann, A. K., & Navneet, K. (2013). Review paper on clustering techniques. *Global Journal of Computer Science and Technology*.
- McInerney, G. J., & Etienne, R. S. (2012). Ditch the niche - is the niche a useful concept in ecology or species distribution modelling? *Journal of Biogeography*, 39(12), 2096–2102. <https://doi.org/10.1111/jbi.12033>
- Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, 415, 108837. <https://doi.org/10.1016/j.ecolmodel.2019.108837>
- Moles, A. T., Gruber, M. A. M., & Bonser, S. P. (2007). A new framework for predicting invasive plant species. *Journal of Ecology*, 0(0), 071119203335006-???. <https://doi.org/10.1111/j.1365-2745.2007.01332.x>
- Nizar, G., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: A brief survey. *A Review of Machine Learning Techniques for Processing Multimedia Content*, 1, 9–16.

- Pavoine, S., Gasc, A., Bonsall, M. B., & Mason, N. W. H. (2013). Correlations between phylogenetic and functional diversity: Mathematical artifacts or true ecological and evolutionary processes? *Journal of Vegetation Science*, 24(5), 781–793. <https://doi.org/10.1111/jvs.12051>
- Peters, R. H. (1991). *A critique for ecology*. Cambridge University Press.
- Peterson, A. T. (2012). Niche modeling - Model evaluation. *Biodiversity Informatics*, 8(1). <https://doi.org/10.17161/bi.v8i1.4300>
- Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213(1), 63–72. <https://doi.org/10.1016/j.ecolmodel.2007.11.008>
- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press. <https://doi.org/10.1515/9781400840670>
- Qiao, H., Escobar, L. E., & Peterson, A. T. (2017). Accessible areas in ecological niche comparisons of invasive species: Recognized but still overlooked. *Scientific Reports*, 7(1), 1213. <https://doi.org/10.1038/s41598-017-01313-2>
- Qiao, H., Lin, C., Jiang, Z., & Ji, L. (2015). Marble algorithm: A solution to estimating ecological niches from presence-only records. *Scientific Reports*, 5(1), 14232. <https://doi.org/10.1038/srep14232>
- Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, 6(10), 1126–1136. <https://doi.org/10.1111/2041-210X.12397>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (4.0.3). <https://www.r-project.org/>
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of Database Systems* (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565
- Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2), 169–194. <https://doi.org/10.1023/A:1009745219419>
- Simões, M. V. P., & Peterson, A. T. (2018). Importance of biotic predictors in estimation of potential invasive areas: The example of the tortoise beetle *Eurypedus nigrosignatus*, in Hispaniola. *PeerJ*, 2018(12). <https://doi.org/10.7717/peerj.6052>
- Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 106(Supplement_2), 19644–19650. <https://doi.org/10.1073/pnas.0901637106>
- Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2(0). <https://doi.org/10.17161/bi.v2i0.4>

- van Kleunen, M., Weber, E., & Fischer, M. (2010). A meta-analysis of trait differences between invasive and non-invasive plant species. *Ecology Letters*, 13(2), 235–245. <https://doi.org/10.1111/j.1461-0248.2009.01418.x>
- Webb, C. O., Ackerly, D. D., McPeck, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33(1), 475–505. <https://doi.org/10.1146/annurev.ecolsys.33.010802.150448>
- Wiens, J. J., & Graham, C. H. (2005). Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics*, 36(1), 519–539. <https://doi.org/10.1146/annurev.ecolsys.36.102803.095431>

FIGURE AND TABLES

CHAPTER 3

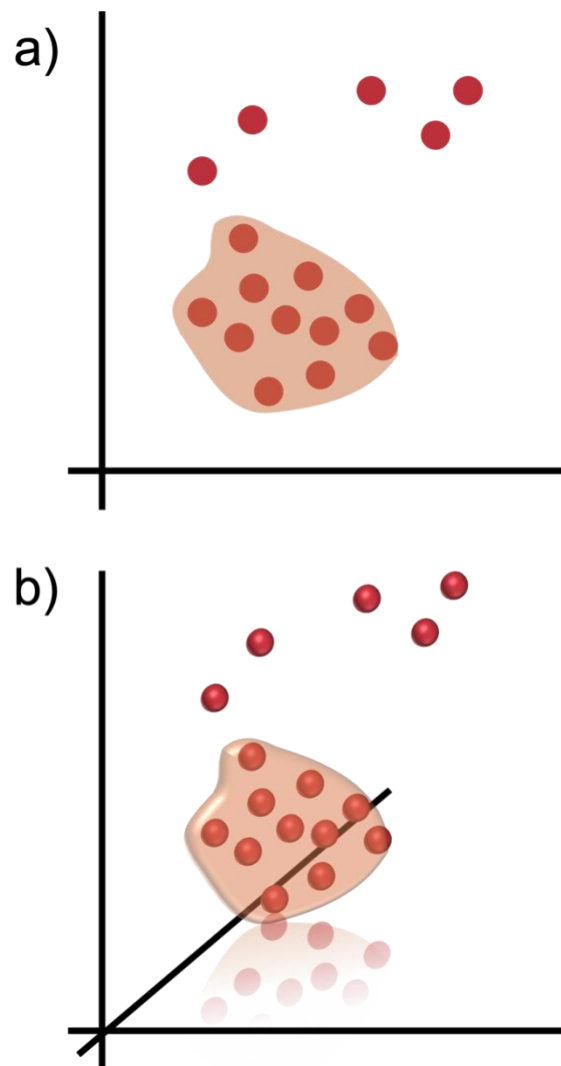


Figure 1. Clustering example. a) cluster in 2-D, called an area; b) cluster in 3-D, called a volume. Data points that fall closer to each other most of the time will be part of a cluster. Note that the clusters may vary in shape. When plotting environmental values in an XY or XYZ plot, this pertains to **E**-space (i.e., environmental space)

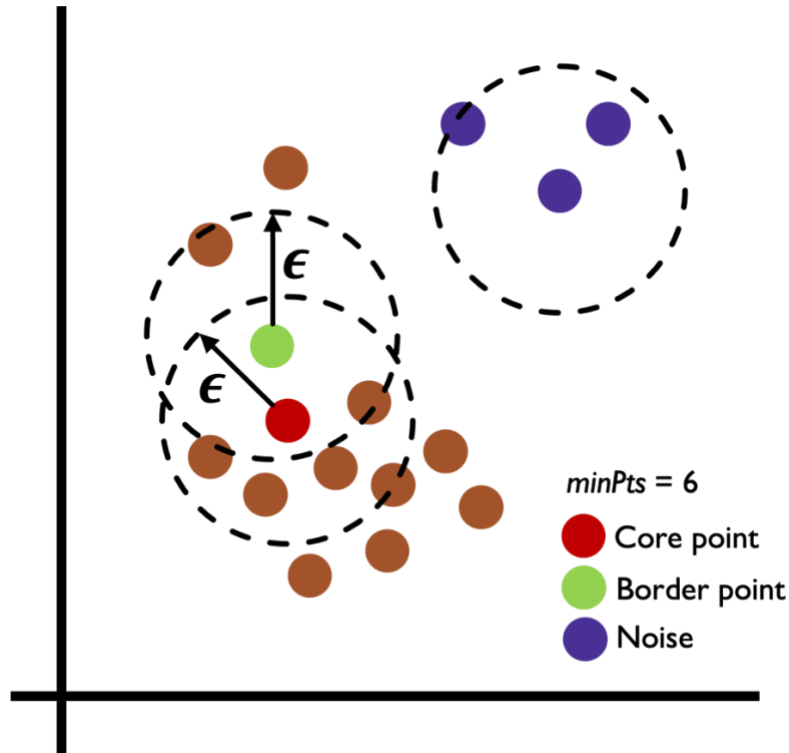


Figure 2. DBSCAN point labeling example. This figure shows one example of the labeling of points in DBSCAN. In this figure, only one core point (red) and one border point (green) are labeled as illustrative examples. Note brown points may also be core or border points. The radius of the neighborhood ϵ is denoted by the dotted lines. Outliers or noise points are shown in purple.

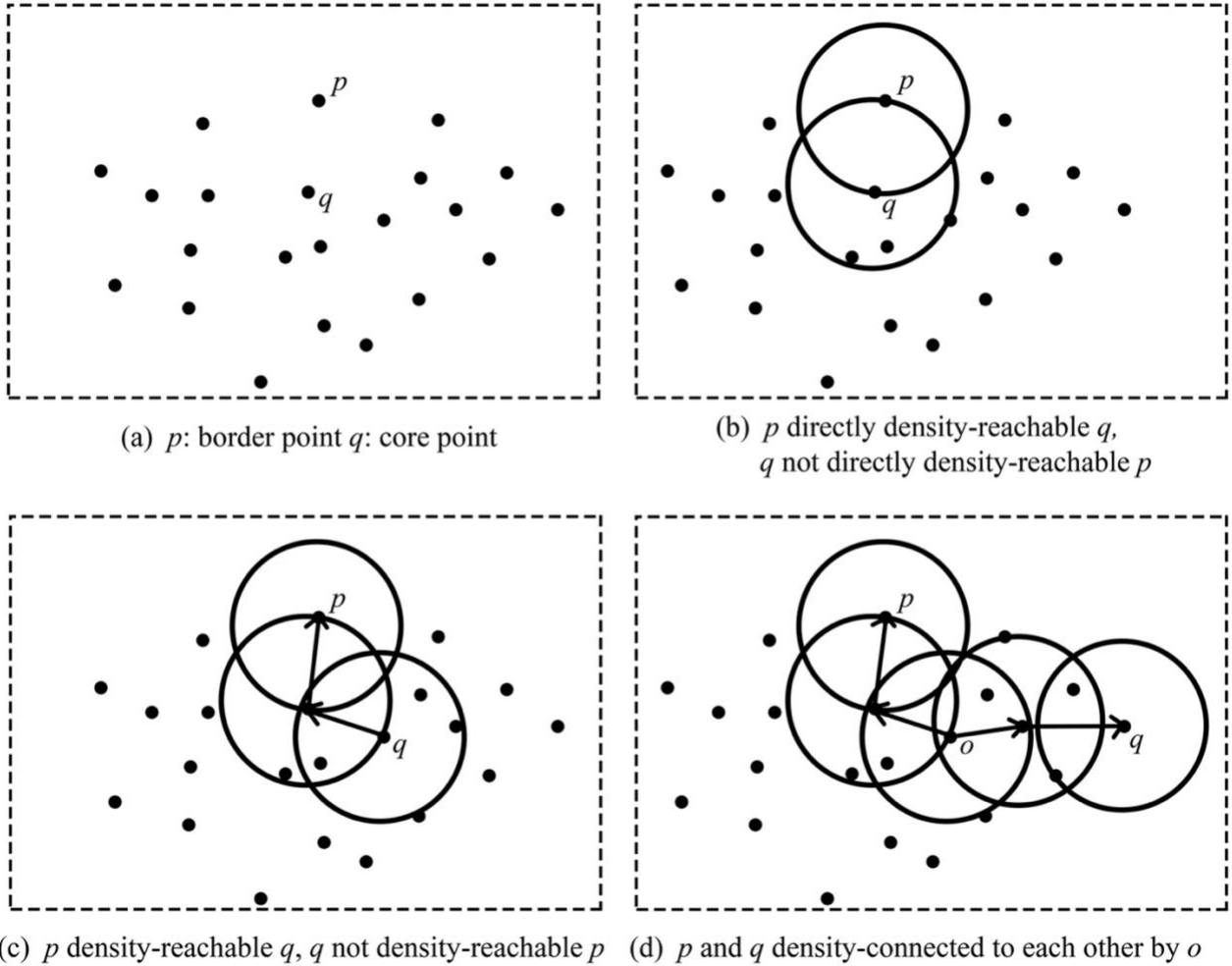


Figure 3. Basic concepts of Marble algorithm. a) shows examples of two classes of points: core (q) and border (p) points; b) illustrates the concept of direct density-reachability; (c) illustrates the concept of density-reachability; d) illustrates the concept of density-connectivity. Original figure from Qiao et al. (2015).

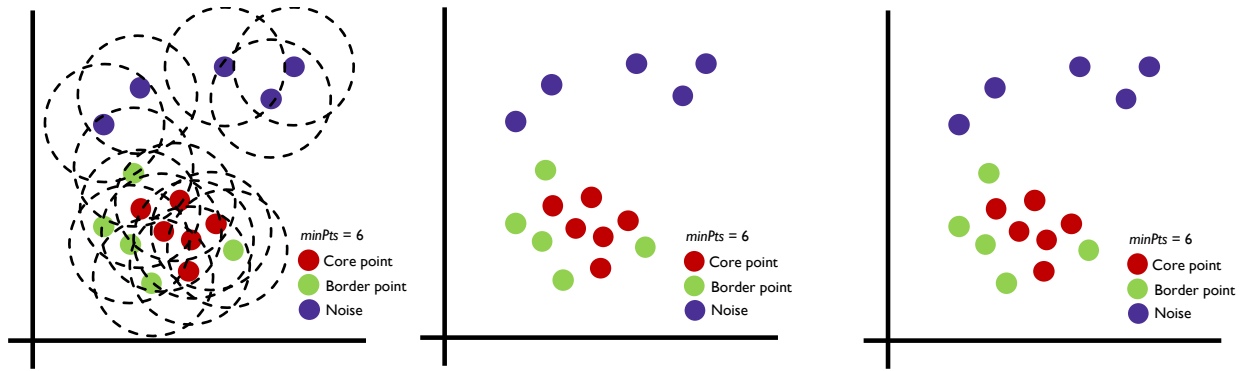


Figure 4. DBSCAN algorithmic steps. 1) arbitrary select a point p ; 2) retrieve all points density-reachable from p based on ϵ and $minPts$; 3) If p is a core point, a cluster is started; 4) each point p in the cluster will broadcast out their own perimeter looking to find new members to join the cluster. 5) If p is a border point, no points are density-reachable from p , and DBSCAN visits the next point of the database, and 6) continue the process until all the points have been processed.

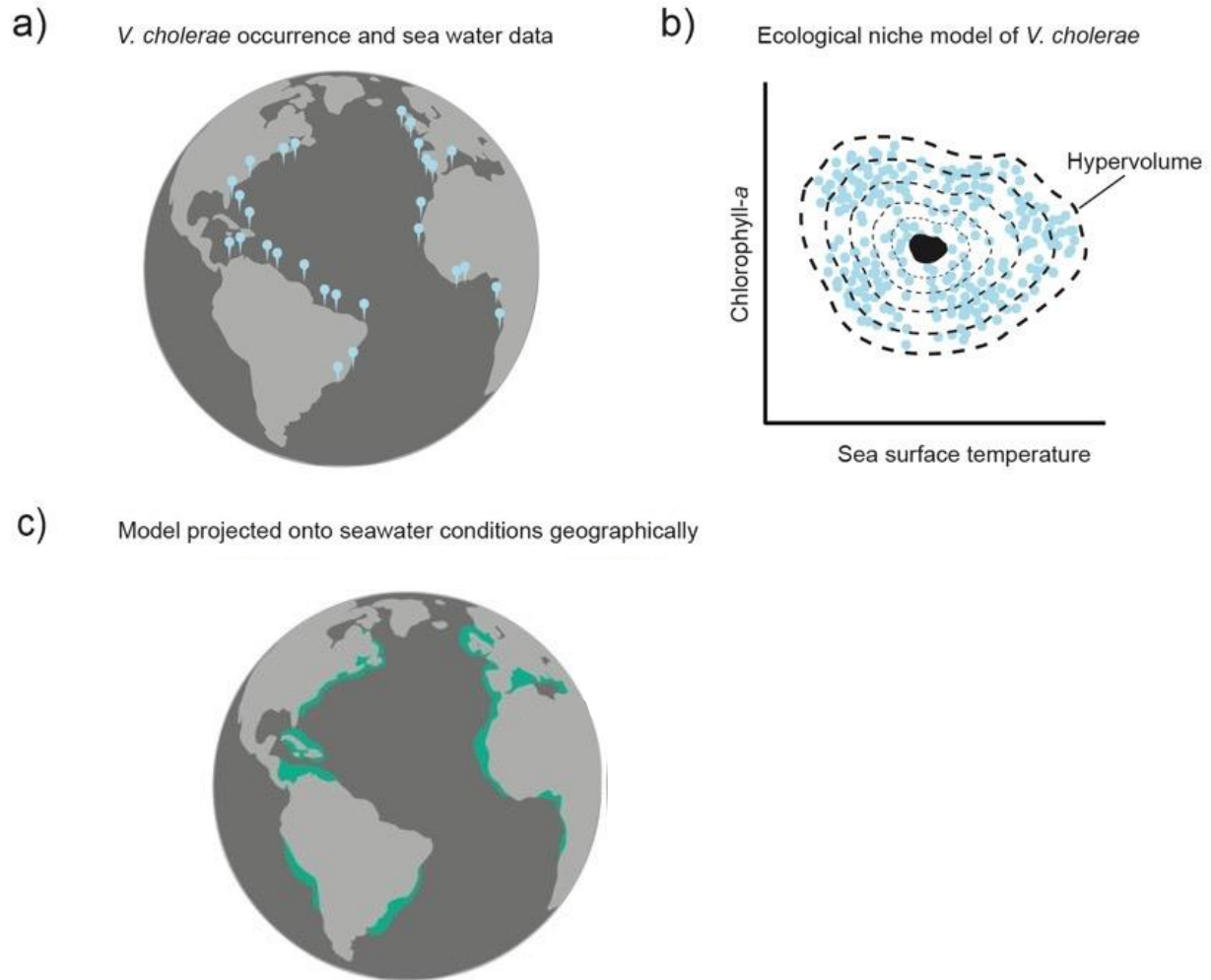


Figure 5. Analytical framework of *Vibrio cholerae* bacteria ENM. a) *Vibrio cholerae* records from coastal seawaters during the last two decades were collected, curated, standardized, and coupled with satellite-derived coastal water conditions at 4 km resolution for the years from 2003 to 2020. b) Data integration accounting by locality and date of the *V. cholerae* record was developed using machine-learning clustering algorithms to build hypervolume models of *V. cholerae* environmental suitability for growth and reproduction.

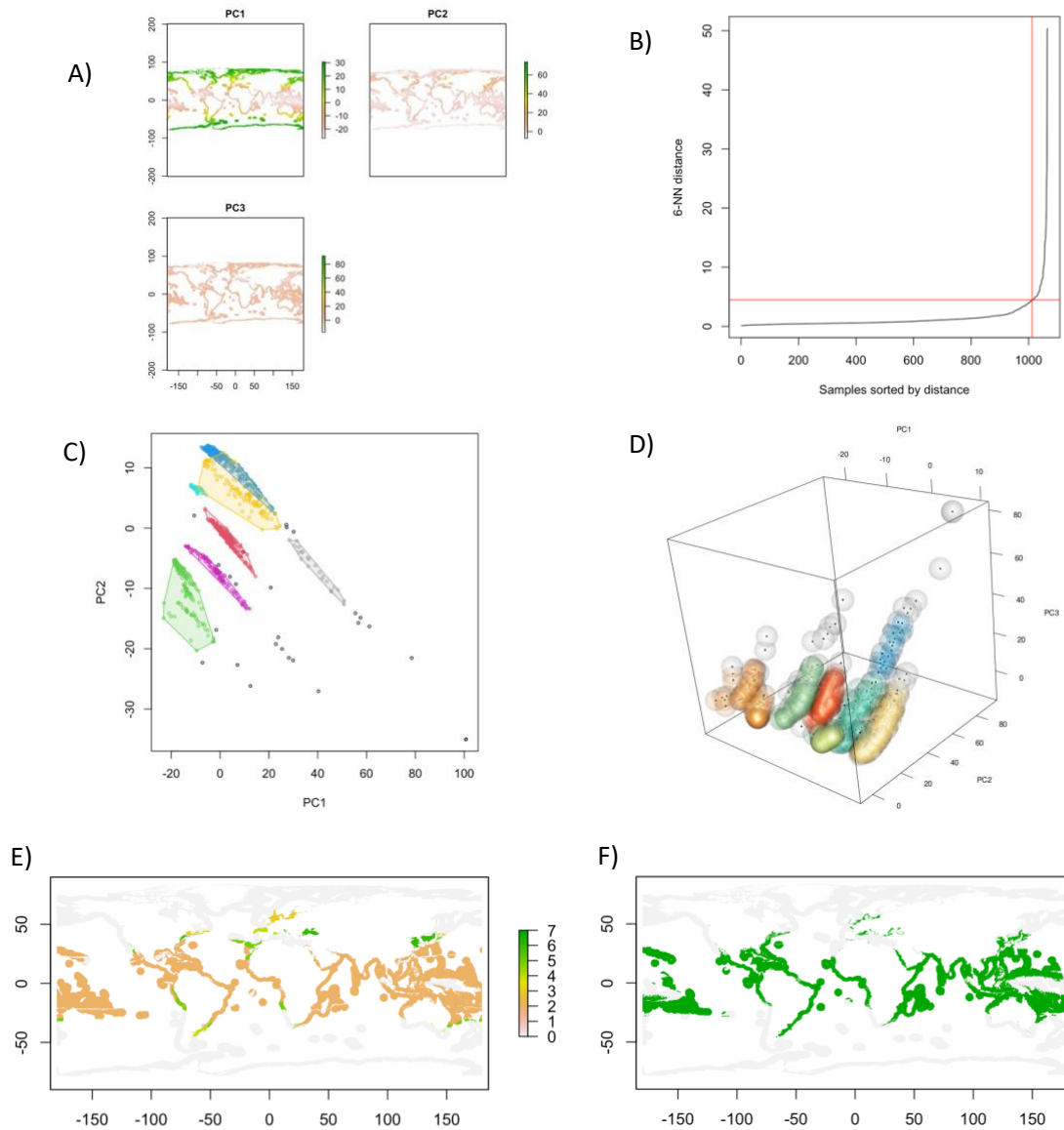


Figure 6. Methods of the implementation *Marble* to reconstruct the realized niche of *Vibrio cholerae*. A) environmental and biological predictors were normalized and transformed using PCA. B) The selection of parameters to use in *Marble* was selected following literature recommendations for $minPts = 2 * dim$ (Sander et al., 1998, Ester et al., 1996), and ϵ (y-intercept of the red horizontal line) was calculated using the “elbow method” by calculating distances with KNN using an omission threshold of 5% (red vertical line). C and D illustrate *Marble* clusters in **E**-space c) in 2-d and d) in 3-d. E) show the resulting projection of *Marble* from **E** to **G**-space, retaining the clusters. F) resulting rasters after binarization (i.e., labeling as suitable and unsuitable, ignoring cluster number).

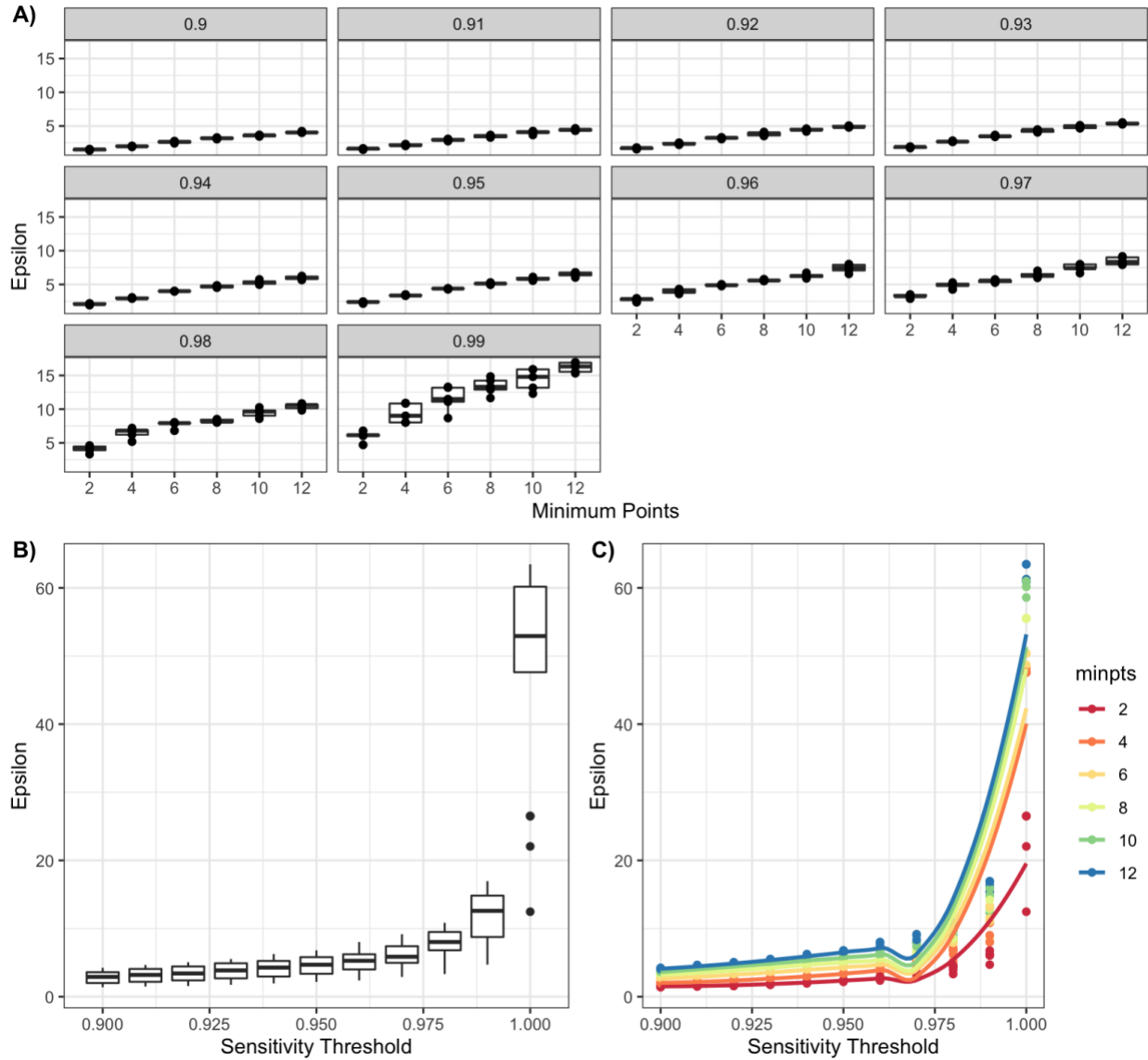


Figure 7. Epsilon (ϵ) distribution based on omission threshold and $minPts$. (A) The panel grid shows the different distribution of ϵ separated by sensitivity threshold. The x-axis represents the number of $minPts$. Variation in resulting ϵ increases as the sensitivity threshold increases (i.e., level of confidence in the input data; one minus omission threshold). Noting, however, how the values follow a logistic curve, approaching some equilibrium value for ϵ as $minPts$ are increased. (C) Illustrates the distribution on the resulting ϵ using KNN as $minPts$ varies across a range of sensitivity thresholds (i.e., one minus omission threshold). Colored lines show the likely distribution of values grouped by $minPts$. Similarly, (B) Boxplot illustrates the distribution of the resulting ϵ when all $minPts$ values are grouped by a specific sensitivity threshold. Both figures indicate that as confidence in the input data grows, values for ϵ will increase, as well as if $minPts$ increase.

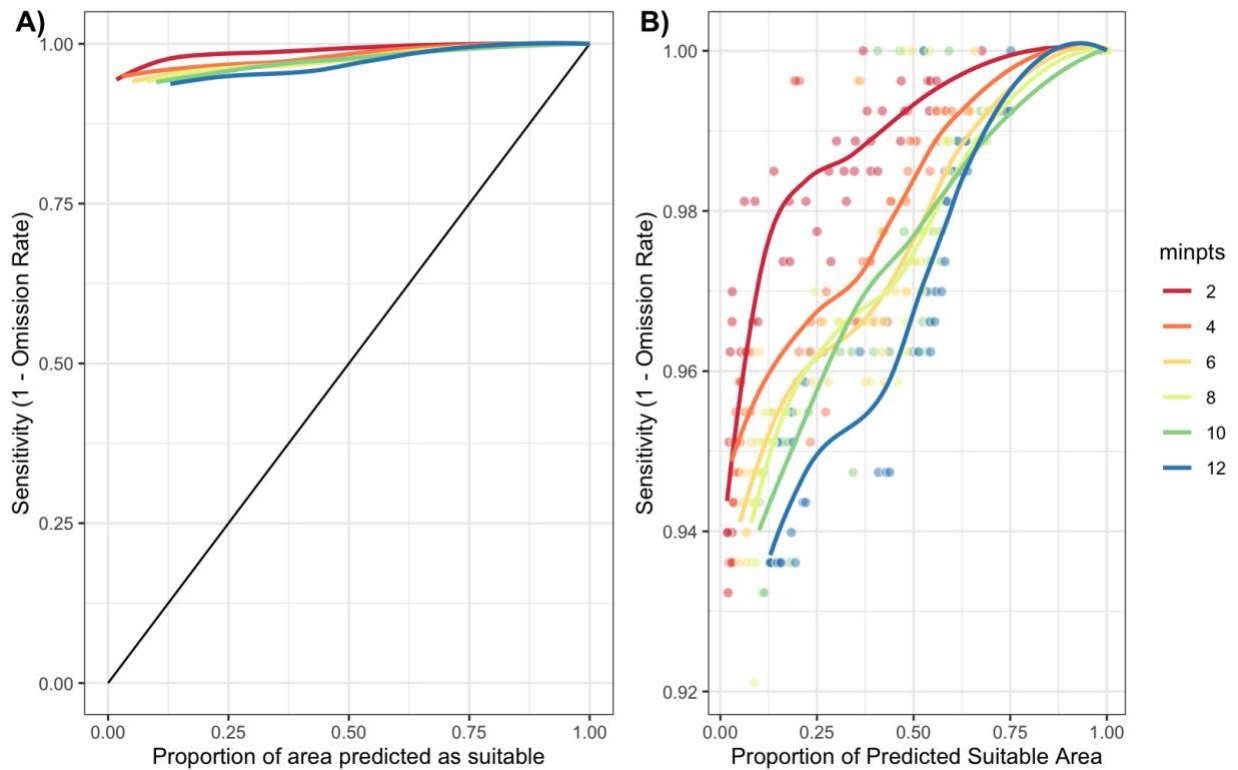


Figure 8. Sensitivity versus commission rate output from cross-validation using evaluation data. This figure illustrates a variation of ROC evaluation, where the x-axis corresponds to the proportion of area predicted as suitable instead of the commission rate (i.e., false positive rate). Although proportion of suitable predicted area has a 1:1 ratio when compared to commission rate (Peterson et al., 2008). The independent evaluation data for each cross-validation fold (i.e., five folds) grouped by *minPts* was used to develop the curves. (B) is the zoom-in of the figure in panel (A). In panel (B) each color represents a different *minPts* value. Lines show the distribution grouped by *minPts* of the original data points (colored dots). Each colored dot represents one of the 330 model replications from the parametrization matrix, colored by the *minPts* value. Overall, model outputs from all *minPts* show that they predict better than by random choice.

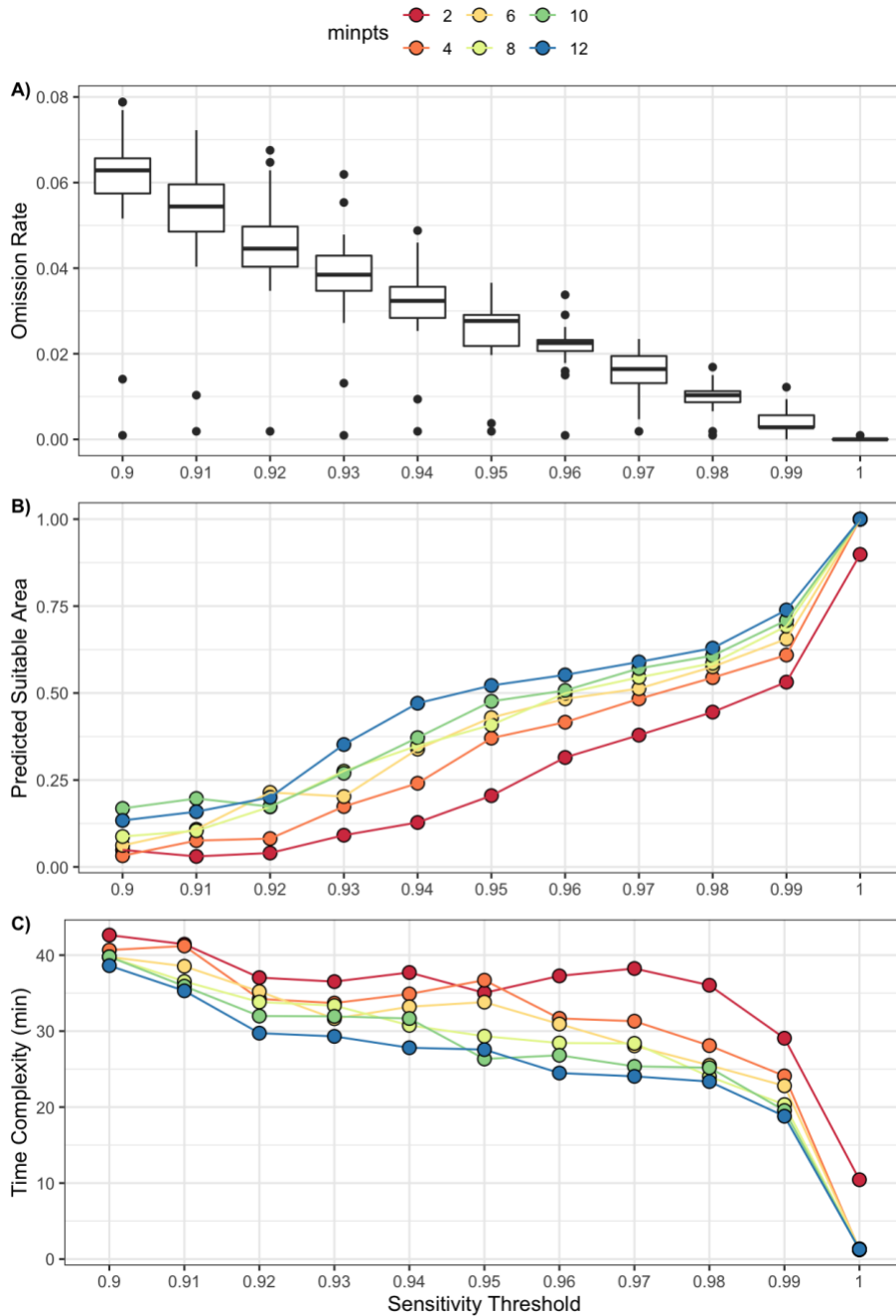


Figure 9. Evaluation of *Marble* in G-space based on selected user-defined sensitivity threshold. The panel figure illustrates the performance of *Marble* through different value parameters of *minPts* and sensitivity threshold (i.e., 1 – omission threshold). Panel (A) shows the variation of omission rate through different thresholds of sensitivity. Boxplots show a negative correlation between variable omission rate and sensitivity threshold. Panel (B) illustrates the change in the predicted area through different sensitivity thresholds grouped by values of *minPts*. Panel (C) illustrates the differences in time complexity grouped by *minPts*. Legend for both Panel B and C is at the top of the figure.

CHAPTER 4

CLUSTER-BASED ECOLOGICAL NICHE MODELING AGAINST TRADITIONAL CORRELATIVE ECOLOGICAL NICHE MODELING

Abstract

A wide variety of modeling techniques have been developed in spatial epidemiology to estimate the ecological niche of species. Although, there is still a need for more data-driven methods that require little knowledge of the natural history of the organism, especially for systems with limited data (e.g., emerging disease of which there is limited information about its ecology). Here I focused on comparing the performance between a new presence-only ecological niche modeling technique using density-based clustering (*Marble*) against traditional correlative algorithms, including Generalized Linear Models (GLM), and Maximum Entropy (Maxent), Boosted Regression Trees (BRT), Support Vector Machines (SVM), and Generalized Additive Model (GAM). Although it is advised that no single algorithm should be used for disease mapping, instead, different algorithms should be employed for a more informed and complete understanding. Density-based clustering algorithms avoid the need to generate pseudoabsences or background data to address requirements of correlative models and remove the need to determine study areas a priori. Results show that *Marble* had comparable performance with traditional correlative models. *Marble* constantly resulted in a desirable ratio of omission rate (i.e., low degree of interpolation) and area predicted as suitable (i.e., lower omission error and lower degree of extrapolation). Identifying the advantages and disadvantages of each of the ecological niche models explored here can help inform study designs for many taxa under diverse data-availability circumstances.

Introduction

Ecological niche models (ENMs) have been employed as a predictive tool in diverse research applications, including biological conservation, invasion ecology, and disease mapping (Escobar et al., 2018). There is not a “one-size-fits-all” model (Qiao, Soberón, et al., 2015), model selection should be based on the data available, assumptions necessary, and research question (Escobar et al., 2018; Peterson et al., 2015; Soberón & Peterson, 2005).

Correlative ENM estimate the potential distribution of a species as a function of geographically referenced climatic predictor variables using multiple regression approaches (Escobar, 2020; Escobar et al., 2017, 2018; Peterson & Soberón, 2011). Given a set of geographically referred observed presences of a species and a set of environmental conditions, an algorithm finds the most likely environmental ranges within which a species is observed (Peterson & Soberón, 2011; Peterson & Soberón, 2012; Soberón & Nakamura, 2009). In other words, correlative ENMs model the observed distribution of a species as a function of occupied and unoccupied environmental conditions. Correlative ENMs assume that species are at equilibrium with their environment and that the relevant environmental variables have been adequately sampled (Soberón & Nakamura, 2009).

Model selection and validation is a critical step in the model procedure; it is the measure of how useful and trustworthy models are (Qiao, Escobar, & Peterson, 2017; Wenger & Olden, 2012). Previous studies have assessed ecological niche modeling performance based on their spatial fit with the calibration data and on correct prediction of independent evaluation occurrence data. This study compares the predictive ability of seven ENM methods based on presence-background data and correlations, including Generalized Linear Models (GLM), and Maximum Entropy (Maxent), Boosted Regression Trees (BRT), Support Vector Machines (SVM), and Generalized Additive Model (GAM) and presence-only data and cluster analysis, *Marble*.

Methods

The predictive power of five different ENMs that use presence-background data (BRT, GLM, SVM, GAM, and Maxent) and a new presence-only protocol, *Marble*, were explored. Correlative presence-background models were executed using the package *sdm* in the statistical software R (Naimi & Araújo, 2016; R Core Team, 2022), and the *Marble* models were executed with a new generated code linking C++ and R languages (Chapter 3) for. For experimental purposes, model calibration was done using default parameters for each method to allow easy model replication (Elith et al., 2006). Standardized data of the tree white oak (*Quercus alba*) was implemented, commonly used in ecological niche modeling comparisons, which contains presence data and is available in the *hypervolume* package in R. This dataset selection allowed to compare *Marble* performance against other methods in a terrestrial ecosystem for a carefully curated, standardized, and broadly accepted dataset of plants. Based on the known observation of *Q. alba*, the study area encompasses Canada, The United States, and Mexico (Figure 1). Further, the *Q. alba* data were filter to one observation per pixel to reduce auto correlation in the models. For the environmental predictor, I procured bioclimatic variables form WorldClim and ran a PCA to reduce dimensionality and selected principal components that explained >90% of the variance. Models were calibrated using only a subset of the presence data (i.e., 80% of the data), and the other set of presence points (i.e., 20% of the data) was held out for model evaluation. Finally, to supplement the presence-background models a random set of 10,000 background point was generated from the study area.

Using the calibration subsets and random-generated background point, I ran the *sdm* function in R for each of the methods (BRT, GLM, SVM, GAM, and Maxent), separately. After obtaining the calibrated models (i.e., fitted models), models were projected into the study area using the

function *predict* within the *sdm* package (Naimi & Araújo, 2016). Then, using the algorithm defined in Chapter 3 I fit and project all *Marble* models. For all the *Marble* models, the parameters *minPts*=3 and a range of values for omission threshold=0% (i.e., final model will include all the calibration presence points) to 10% (i.e., final model will only include 90% of the calibration presence points) were used, resulting in 10 different omission scenarios binary models.

All other algorithms (BRT, GLM, SVM, GAM, and Maxent) generate continuous outputs. To facilitate comparison with *Marble*'s binary output, all other algorithms were transformed into binary outputs. To transform continuous outputs into binary outputs and evaluate binary-output models, a range of omission threshold (i.e., proportion of evaluation presence point to exclude for calibration of the models) between 0% and 10%, comparable to *Marble* thresholds, were used. Similar evaluation metrics as in Chapter 3 were implemented, i.e., the evaluation of the proportion of area predicted suitable as a proxy of extrapolation and interpolation, omission rate (i.e., how well the model can predict independent evaluation presence points correctly), and the cumulative binomial probability (CBP), to test if models were able to predict independent occurrence records better than by chance.

Result

The environmental space defined by the first three principal components explained 95.8% of the overall variance: 53.1%, 31.3%, and 11.4%, respectively. After filtering the presence point to one per pixel, resulting in with 347 unique observations of *Q. alba* (Figure 1). Figure 2a, shows the resulted calibrated models for all continuous output models (i.e., BRT, GLM, SVM, GAM, and Maxent).

Using independent evaluation data, the models' calculated omission rate and proportion of area deemed suitable provided detailed quantile information on model performance. Overall, all models were able to predict the unseen instance better than by chance alone (Figure 3b). GAM, Maxent, and *Marble*, had the lowest omission rate to predicted suitable area ratio (Figure 3a). Between, GAM, Maxent, and *Marble*, GAM resulted in the highest mean omission error rate, while *Marble* had the lowest omission error rate overall. In fact, *Marble* had the lowest omission error out of all the models (BRT, GLM, SVM, GAM, and Maxent). BRT had low overall omission error rates with not much variability, although it predicted on average 70% of the geographic extent to be suitable with more than 10% variability in predicted area (i.e., high degree of extrapolation; Figure 3a, solid red circle). Moreover, with omission threshold >5% BRT resulted in a CBP of 1, that is, BRT did not predict better than random when omission threshold >5% (i.e., all area is predicted as suitable, high degree of extrapolation; Figure 2b BRT).

Maxent showed the lowest degree of extrapolation, followed by GAM and then *Marble*. For all other models, BRT, GLM, SVM, GAM, and Maxent, the CBP tests determined that in all omission threshold scenarios models can predict better than by random chance (i.e., $CBP < 0.001$). SVM had the highest average omission rate, with an average value of 9%. Overall, SVM had an omission error of >10% for omission threshold values >5%, with a very low degree of extrapolation (Figure 3a and Table 1). GLM resulted in overall balanced ratio of omission rate and area predicted suitable (Figure 3a). GLM models, on an average, had higher value of omission rate and area predicted as suitable when compared to than Maxent, *Marble*, and GAM, but lower omission error and area predicted as suitable than SVM and BRT, respectively.

Discussion

Models were evaluated based on their spatial fit with the calibration data (i.e., degree of extrapolation and interpolation) and on correct prediction of independent evaluation data (i.e., omission error rate). Since continuous-output and binary-output models both are evaluated with different metric continuous outputs from Maxent, GLM, KDE, SVM, and BRT were converted to binary based on a thresholding to be comparable to *Marble's* output. I found that *Marble* had comparable performance to traditional correlative models, and for the case of modeling *Q. alba's* realized niche, *Marble* performed better than commonly used correlative models, including GLM, SVM, and BRT (Figure 3). These may have implications in the use of presence-background versus presence-only models and selection of evaluation metrics to validate models on their ability to reconstruction a species' realized niche.

This study used default parameters to facilitate comparisons among algorithms (Elith et al., 2006). Nevertheless, detailed parameterizations instead of default configurations may improve model performance at the cost of exhaustive experimentation to reach the optimal parameter values (Oliveira et al., 2017). Detailed model parameterization, however, is not the common practice in macroecology. Thus, default parameter were used to mirror common practices in modern ecological niche modeling applications (Carlson et al., 2022; Petitpierre et al., 2012).

A common practice for transforming continuous outputs to binary outputs is the use of minimum training presence thresholding. Minimum training presence thresholding consist in identifying the lowest predicted suitability value for the calibration presence points, so that the final model includes all the calibration presence points (0% omission threshold), which is suggested when comprehensive data curation protocols are employed (Phillips, Anderson & Schapire, 2006). Here 0% omission threshold was included under the assumption that *Q. alba* data contained high accuracy in the species identification and geolocation of records.

I suggest that the evaluation metric should be selected based on the model feature desired and the use intended (Soberón & Peterson, 2005). For realized niches, models should have a high-performance regarding model fit (i.e., degree of extrapolation and interpolation), with capacities to predict independent data not used during model calibration (i.e., low omission error). High model extrapolation (i.e., high proportions of area predicted as suitable) may be undesirable, as it assumes that the species can survive under conditions outside of their historic observed conditions to maintain populations, sometimes far outside known conditions. BRT models predicted broad areas suitable for the species (i.e., high degree of extrapolation), thus, this model might not be the most adequate to reconstruct *Q. alba*'s realized niche. *Marble*, however, constantly resulted in a desirable ratio of omission rate (i.e., low degree of intrapooltion) and area predicted as suitable (i.e, lower omission error and lower degree of extrapolation). With Maxent having the lowest degree of extrapoaltion and the second to last lowest degree of intrapoolation (first is *Marble*). Although noting that low interpolation and extrapolation, expressed in environmental space metrics as low interpolation and extrapolation, is also likely to be biologically unrealistic (Escobar et al., 2018). Basic physiology suggests that species will be able to survive under intermediate conditions, as physiological responses tend to be bell-shaped in terms of response of suitability to environmental conditions, rather than bimodal, intolerance to intermediate environmental conditions (Austin, Cunningham, & Fleming, 1984; Birch, 1953; Maguire, 1973; Qiao, Escobar, et al., 2016). Nevertheless, attempts to reconstruct species' realized niches may prefer reduced interpolation to identify environment used by the species instead of potentially used by the species, a more desirable feature when reconstructing fundamental niches.

In summary, the analyses conducted here support the idea that there is often not a single 'best' algorithm or method that fits with all ecological applications and data configurations for estimating

species niches (Guillera-Arroita et al., 2015; Qiao, Escobar, Saupe, et al., 2017; Qiao, Soberón, et al., 2015). However, results showed that *Marble* can perform similar to, or in some cases outperforms, extant correlative ecological niche models. *Marble* can reduce considerably the degree of extrapolation, resulting in projection of narrower niches staying closer to the realized niche of a species without overfitting the data (i.e., high degree of interpolation). Future research should consider modeling comparisons using virtual species data for which there is control over the uncertainty, biases, data amounts, and ecology of the species (e.g., niche breadth, endemism, rarity, abundance).

References

- Carlson, C. J., Albery, G. F., Merow, C., Trisos, C. H., Zipfel, C. M., Eskew, E. A., Olival, K. J., Ross, N., & Bansal, S. (2022). Climate change increases cross-species viral transmission risk. *Nature*, *607*(7919), 555–562. <https://doi.org/10.1038/s41586-022-04788-w>
- Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., ... E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Escobar, L. E. (2020). Ecological niche modeling: An introduction for veterinarians and epidemiologists. *Frontiers in Veterinary Science*, *7*. <https://doi.org/10.3389/fvets.2020.519059>
- Escobar, L. E., Qiao, H., Cabello, J., & Peterson, A. T. (2018). Ecological niche modeling re-examined: A case study with the Darwin's fox. *Ecology and Evolution*, *8*(10), 4757–4770. <https://doi.org/10.1002/ece3.4014>
- Escobar, L. E., Qiao, H., Lee, C., & Phelps, N. B. D. (2017). Novel methods in disease biogeography: A case study with heterosporosis. *Frontiers in Veterinary Science*, *4*. <https://doi.org/10.3389/fvets.2017.00105>
- Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*(3), 276–292. <https://doi.org/10.1111/geb.12268>

- Naimi, B., & Araújo, M. B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, *39*(4), 368–375. <https://doi.org/10.1111/ecog.01881>
- Oliveira, S. V. de, Romero-Alvarez, D., Martins, T. F., Santos, J. P. dos, Labruna, M. B., Gazeta, G. S., Escobar, L. E., & Gurgel-Gonçalves, R. (2017). Amblyomma ticks and future climate: Range contraction due to climate warming. *Acta Tropica*, *176*, 340–348. <https://doi.org/10.1016/j.actatropica.2017.07.033>
- Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. <https://doi.org/10.1016/j.ecolmodel.2007.11.008>
- Peterson, A. T., Papeş, M., & Soberón, J. (2015). Mechanistic and correlative models of ecological niches. *European Journal of Ecology*, *1*(2), 28–38. <https://doi.org/10.1515/eje-2015-0014>
- Peterson, A. T., & Soberón, J. (2012). Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza a Conservacao*, *10*(2), 102–107. <https://doi.org/10.4322/natcon.2012.019>
- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press. <https://doi.org/10.1515/9781400840670>
- Petitpierre, B., Kueffer, C., Broennimann, O., Randin, C., Daehler, C., & Guisan, A. (2012). Climatic niche shifts are rare among terrestrial plant invaders. *Science*, *335*(6074), 1344–1348. <https://doi.org/10.1126/science.1215933>
- Qiao, H., Escobar, L. E., & Peterson, A. T. (2017). Accessible areas in ecological niche comparisons of invasive species: Recognized but still overlooked. *Scientific Reports*, *7*(1), 1213. <https://doi.org/10.1038/s41598-017-01313-2>
- Qiao, H., Escobar, L. E., Saupe, E. E., Ji, L., & Soberón, J. (2017). A cautionary note on the use of hypervolume kernel density estimators in ecological niche modelling. *Global Ecology and Biogeography*, *26*(9), 1066–1070. <https://doi.org/10.1111/geb.12492>
- Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. <https://doi.org/10.1111/2041-210X.12397>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (4.0.3). <https://www.r-project.org/>
- Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, *106*(Supplement_2), 19644–19650. <https://doi.org/10.1073/pnas.0901637106>
- Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, *2*(0). <https://doi.org/10.17161/bi.v2i0.4>

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2), 260–267. <https://doi.org/10.1111/j.2041-210X.2011.00170.x>

FIGURE AND TABLES

CHAPTER 4

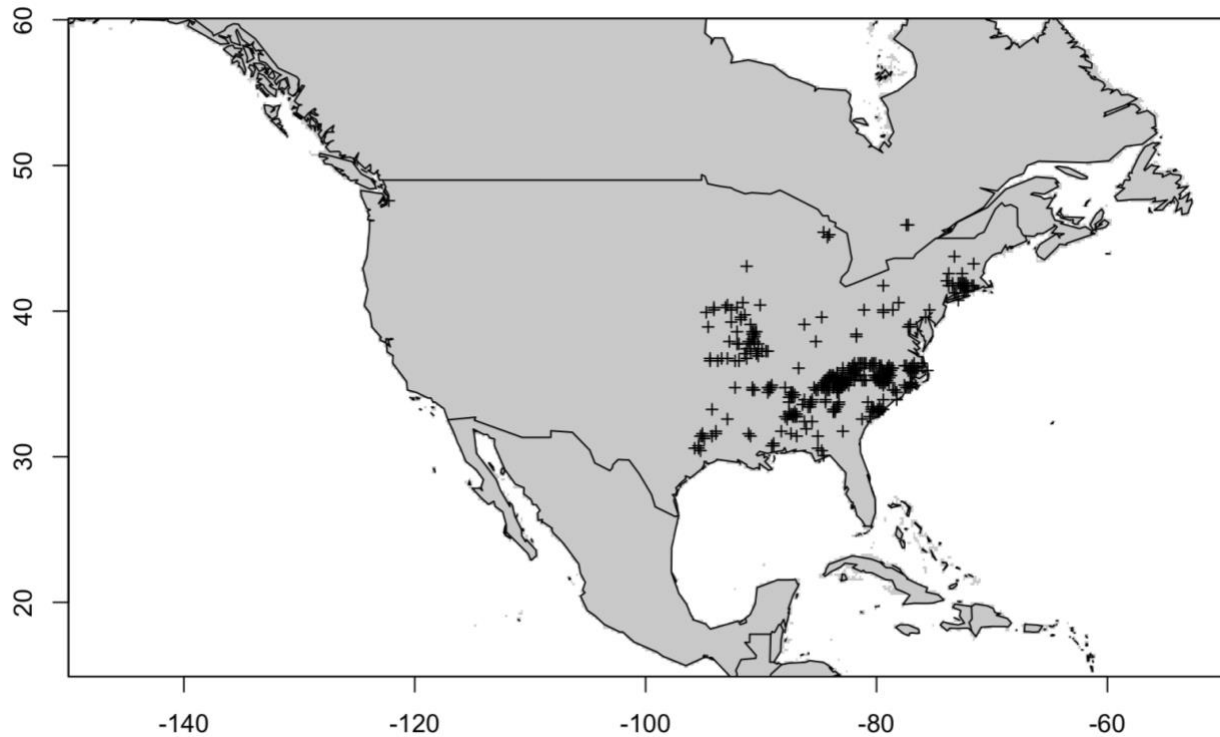


Figure 1. Presence points for *Quercus alba*. Figure illustrate the geolocation of the available presence records for *Quercus alba* (black crosses) commonly used for research and teaching purposes in ecological niche modeling and available in the *hypervolume* package in R. *Quercus alba* is mainly found the Southeast United States and a few observations have been recorded in the Canada territory in the western border with the United States. This is a good study model of a species in a terrestrial ecosystem and suitable to compare ecological niche modeling methods.

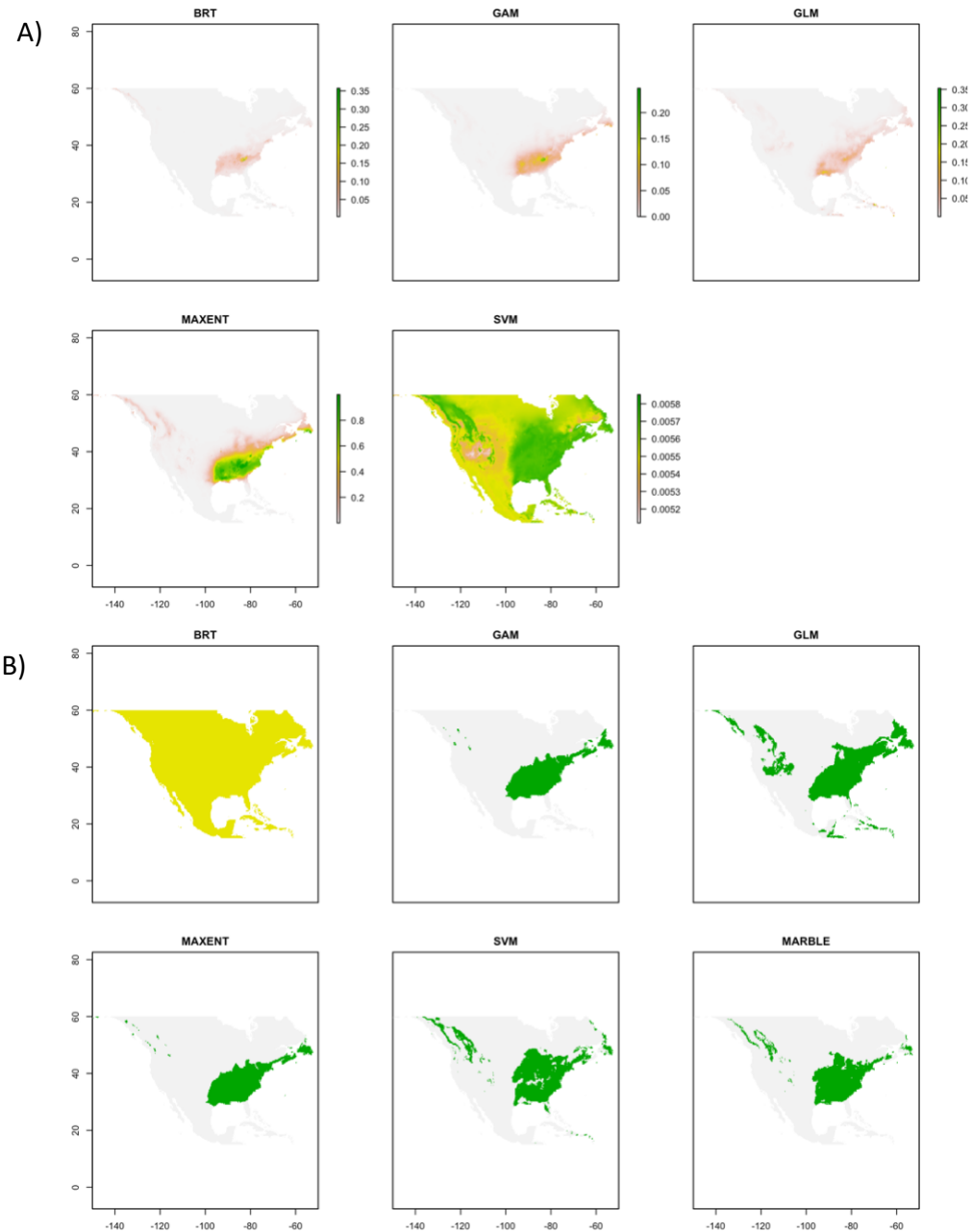


Figure 2. Continuous and binary models of *Quercus alba* in the study area (omission threshold=5%). (A) illustrates the continuous model projections for *Q. alba* potential distribution. Binary models (B) were generated based on a 5% omission threshold from calibration occurrences. BRT, boosted regression trees; GAM, generalized additive model; GLM, generalized linear models; MAXENT, and maximum entropy; and SVM, support vector machines.

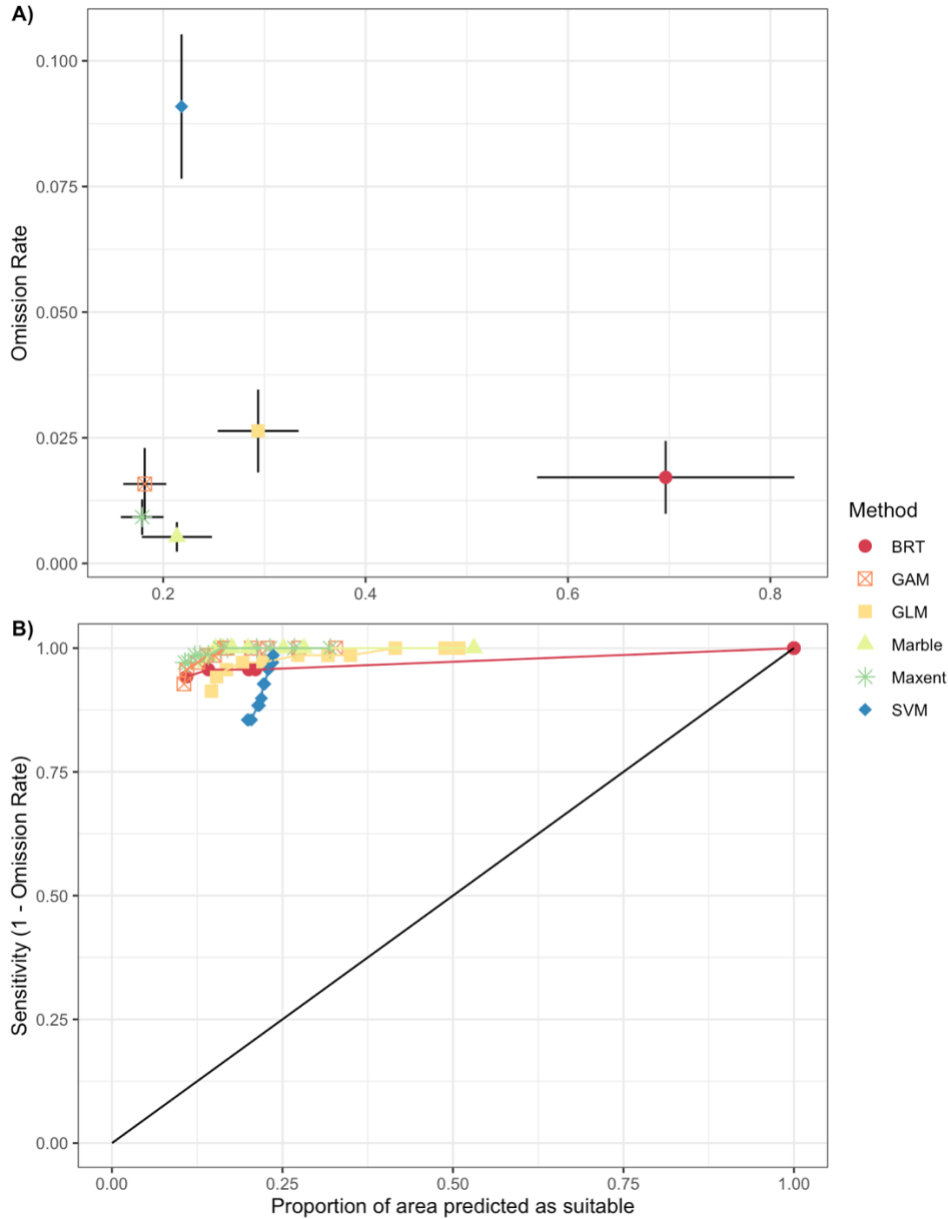


Figure 3. Model performance evaluation in G-space. Boosted regression trees (BRT), generalized additive model (GAM), generalized linear models (GLM), and maximum entropy (MAXENT), support vector machines (SVM), and clustering DBSCAN (MARBLE) models plotted in terms of their omission error rate (i.e., how well it predicted the evaluation presence points) and sensitivity (i.e., 1-omission rate). Figure (A) colored shapes distinguish between methods and represent the mean values among the different omission threshold for binarization (i.e., 0%-10% omission threshold). Vertical black lines represent the standard error of commission error (y-axis) and standard error in proportion of area predicted (x-axis). Figure (B) is a modification of the ROC curve (Peterson et al., 2008), anything above the 1:1 black diagonal line is considered to predict better than by random expectations. All methods fall above the 1:1 line, with BRT and SVM being the models with the lowest predictive performance and Maxent and *Marble* being the best performing models balancing accurate prediction and reduced extrapolation.

Table 1. Model evaluation results based on omission thresholds. Table shows the results obtained from the transformed binary maps based on the corresponding sensitivity threshold or 1-omission threshold (first column). CBP, Cumulative binomial probability; BRT, boosted regression trees; GAM, generalized additive model; GLM, generalized linear models; MAXENT, and maximum entropy; and SVM, support vector machines.

Omission Threshold (%)	Method	Area predicted suitable (%)	Omission Rate (%)	CBP
10%	BRT	10.9%	5.80%	<0.001
	GAM	10.6%	7.25%	<0.001
	GLM	14.6%	8.70%	<0.001
	Maxent	10.7%	2.90%	<0.001
	SVM	19.9%	14.49%	<0.001
	Marble	13.7%	2.90%	<0.001
9%	BRT	14.1%	4.35%	<0.001
	GAM	11.0%	4.35%	<0.001
	GLM	15.3%	5.80%	<0.001
	Maxent	11.2%	2.90%	<0.001
	SVM	20.0%	14.49%	<0.001
	Marble	14.1%	1.45%	<0.001
8%	BRT	20.1%	4.35%	<0.001
	GAM	12.6%	2.90%	<0.001
	GLM	16.8%	4.35%	<0.001
	Maxent	12.2%	1.45%	<0.001
	SVM	20.4%	14.49%	<0.001
	Marble	14.2%	1.45%	<0.001
7%	BRT	21.0%	4.35%	<0.001
	GAM	13.9%	1.45%	<0.001
	GLM	19.2%	2.90%	<0.001
	Maxent	12.9%	1.45%	<0.001
	SVM	21.4%	11.59%	<0.001
	Marble	15.2%	0.00%	<0.001
6%	BRT	100.0%	0.00%	1
	GAM	15.0%	1.45%	<0.001

	GLM	22.1%	2.90%	<0.001
	Maxent	14.1%	1.45%	<0.001
	SVM	21.6%	11.59%	<0.001
	Marble	16.3%	0.00%	<0.001
5%	BRT	100.0%	0.00%	1
	GAM	16.5%	0.00%	<0.001
	GLM	27.3%	1.45%	<0.001
	Maxent	15.9%	0.00%	<0.001
	SVM	21.9%	10.14%	<0.001
	Marble	17.5%	0.00%	<0.001
4%	BRT	100.0%	0.00%	1
	GAM	16.9%	0.00%	<0.001
	GLM	31.7%	1.45%	<0.001
	Maxent	16.9%	0.00%	<0.001
	SVM	22.2%	7.25%	<0.001
	Marble	17.6%	0.00%	<0.001
3%	BRT	100.0%	0.00%	1
	GAM	20.4%	0.00%	<0.001
	GLM	35.0%	1.45%	<0.001
	Maxent	21.2%	0.00%	<0.001
	SVM	22.4%	7.25%	<0.001
	Marble	19.9%	0.00%	<0.001
3%	BRT	100.0%	0.00%	1
	GAM	22.8%	0.00%	<0.001
	GLM	41.5%	0.00%	<0.001
	Maxent	23.2%	0.00%	<0.001
	SVM	22.9%	4.35%	<0.001
	Marble	25.1%	0.00%	<0.001
1%	BRT	100.0%	0.00%	1
	GAM	27.4%	0.00%	<0.001
	GLM	48.8%	0.00%	<0.001
	Maxent	26.8%	0.00%	<0.001
	SVM	23.6%	2.90%	<0.001
	Marble	28.1%	0.00%	<0.001

0%	BRT	100.0%	0.00%	1
	GAM	32.8%	0.00%	<0.001
	GLM	50.9%	0.00%	<0.001
	Maxent	32.0%	0.00%	<0.001
	SVM	23.7%	1.45%	<0.001
	Marble	53.1%	0.00%	<0.001

CHAPTER 5

CONCLUSION

This project studied a new data-driven ecological niche modeling method referred to as the *Marble* algorithm. This approach was originally proposed by Qiao et al. (2015) and allowed modelers to differentiate ecological niche models between the fundamental niche and the realized niche. This simple distinction improves previous generalizations of ecological niche models without a definition of the niche being modeled, which neglects a large body of ecological niche theory and limits the quality of the model interpretation (Escobar & Craft, 2016).

In order to model the realized niche, there is a need to account for the interaction between abiotic and biotic factors and a clear expectation of the environment actually occupied by the species instead of potentially occupied by the species, which is a desirable feature for models of the fundamental niche. Realized niches are defined as the set of abiotic and biotic factors and other restrictive factors that allows a species to exist (Peterson et al., 2011). An important feature of realized niche reconstructions is the consideration of biotic variables, which are difficult to obtain, interpret, and analyze in ENM, especially for large-scale studies (Araújo & Luoto, 2007; Guillera-Aroita et al., 2015; Zimmermann et al., 2010). This gap was addressed in Chapter 2 of this thesis, with the development of a database of global coastal conditions with both abiotic (i.e., temperature) and biotic (i.e., Chlorophyll-a, a proxy of phytoplankton biomass). Remotely sensed imagery can provide an opportunity to analyze large study areas during extended periods.

Based on the results, some of the advantages of *Marble* are that (1) clusters are based on distances, (2) the algorithm does not assume a fixed distribution, (3) it performs well with arbitrarily shaped clusters, (4) it is robust to outliers, and (5) clusters are bounded to \mathbf{G} -space. And an added advantage of this implementation of DBSCAN is the reduction in complexity of

parameter selection because of the addition of the more intuitive parameter, omission threshold. However, a disadvantage of *Marble* is that it cannot handle varying densities, i.e., clusters that may have different ϵ s in the same data set. *Marble* could help differentiate between the fundamental niche and the realized niche, avoiding generalizations of ecological niche models without a definition of the niche being modeled. *Marble*, therefore, could be a new low-complexity ENM tool for the reconstruction of species' realized niche that allows for the integration of abiotic and biotic factors, and it is not sensitive to geographic extents.

The use of density-based clustering algorithms, like *Marble*, may prevent extrapolation for a closer approximation of realized niches (Qiao, Lin, et al., 2015). In turn, I anticipate that density-based clustering models predict the spatial distribution of infectious diseases accurately, reducing Type I error (i.e., the model incorrectly predicts presence). Models such as generalized additive models (GAM), maximum entropy (Maxent), and *Marble* may be more suited for the reconstruction of narrower niches and, therefore, better at approximating the realized niche. *Marble* has comparable performance to extant correlative ecological niche models, and it can considerably reduce the degree of extrapolation (i.e., low Type I error) and remain close to the realized niche of a species without overfitting the data (i.e., high degree of interpolation).

Future research should include a rigorous comparison of other species' presence data to test for performance and stability, that is, test how consistent are the predictions are when calibrated with other species' data. As well as modeling comparison with virtual species data for which there is control over the uncertainty, biases, data amounts, and ecology of the species. Including comparisons of other presence-only methods, including BioClim, convex hulls, minimum volume ellipsoid (MVE), and Blonder et al. one-class support vector machine (SVM) and kernel density estimation (KDE) may provide a more comprehensive assessment of how *Marble* compares

against extant ecological niche models. Future analysis should also evaluate the transferability of *Marble*, i.e., evaluate how well *Marble* can predict over different temporal and spatial ranges. And last, implement the *Marble* algorithm as an open access R package available for the scientist community to use.

In summary, (1) there is a need to develop data-driven ecological niche models that perform accurately even when only limited amounts of data are available or when there is little to no knowledge of natural history of the species, (2) that the use of both abiotic and biotic predictors is the future of ENM (Ashraf et al., 2021; Simões & Peterson, 2018), and (3) for the accurate reconstruction of realized niches, models should have a low degree of extrapolation and interpolation, with capacities to predict independent data not used during model calibration (i.e., low omission error). *Marble*, a density-base clustering algorithm couple with ecological niche theory and the use of abiotic and biotic factors has the potential to help differentiate between the fundamental niche and the realized niche. Therefore being able to define from the start which ecological niche is being modeled; with the use of only limited amount of presence-only data.

Reference

- Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, 16(6), 743–753. <https://doi.org/10.1111/j.1466-8238.2007.00359.x>
- Escobar, L. E., & Craft, M. E. (2016). Advances and limitations of disease biogeography using ecological niche modeling. *Frontiers in Microbiology*, 07. <https://doi.org/10.3389/fmicb.2016.01174>
- Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24(3), 276–292. <https://doi.org/10.1111/geb.12268>

- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press. <https://doi.org/10.1515/9781400840670>
- Qiao, H., Lin, C., Jiang, Z., & Ji, L. (2015). Marble algorithm: A solution to estimating ecological niches from presence-only records. *Scientific Reports*, 5(1), 14232. <https://doi.org/10.1038/srep14232>
- Zimmermann, N. E., Edwards, T. C., Graham, C. H., Pearman, P. B., & Svenning, J.-C. (2010). New trends in species distribution modelling. *Ecography*, 33(6), 985–989. <https://doi.org/10.1111/j.1600-0587.2010.06953.x>