uOttawa

Predicting Glass Sponge (Porifera, Hexactinellida) Distributions in the North Pacific Ocean and Spatially Quantifying Model Uncertainty

Fiona Davidson

Supervisor: Dr. Anders Knudby

Thesis Committee:
Dr. Michael Sawada
Dr. Jeremy Kerr

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for a Master of Science in Geography

Department of Geography, Environment and Geomatics
University of Ottawa

Abstract

Predictions of species' ranges from distribution modeling are often used to inform marine management and conservation efforts, but few studies justify the model selected or quantify the uncertainty of the model predictions in a spatial manner. This thesis employs a multi-model, multi-area SDM analysis to develop a higher certainty in the predictions where similarities exist across models and areas. Partial dependence plots and variable importance rankings were shown to be useful in producing further certainty in the results. The modeling indicated that glass sponges (*Hexactinellida*) are most likely to exist within the North Pacific Ocean where alkalinity is greater than 2.2 μmol l^{-1} and dissolved oxygen is lower than 2 ml l^{-1}. Silicate was also found to be an important environmental predictor. All areas, except Hecate Strait, indicated that high glass sponge probability of presence coincided with silicate values of 150 μmol l^{-1} and over, although lower values in Hecate Strait confirmed that sponges can exist in areas with silicate values of as low as 40 μmol l^{-1}. Three methods of showing spatial uncertainty of model predictions were presented: the standard error (SE) of a binomial GLM, the standard deviation of predictions made from 200 bootstrapped GLM models, and the standard deviation of eight commonly used SDM algorithms. Certain areas with few input data points or extreme ranges of predictor variables were highlighted by these methods as having high uncertainty. Such areas should be treated cautiously regardless of the overall accuracy of the model as indicated by accuracy metrics (AUC, TSS), and such areas could be targeted for future data collection. The uncertainty metrics produced by the multi-model SE varied from the GLM SE and the bootstrapped GLM. The uncertainty was lowest where models predicted low probability of presence and highest where the models predicted high probability of presence and these predictions differed slightly, indicating high confidence in where the models predicted the sponges would not exist.

Keywords: species distribution modeling, spatial ecology, glass sponges, spatial model uncertainty, GLM, GAM, BRT, MaxEnt, standard error, standard deviation.

Acknowledgements

I would first like to thank my thesis supervisor Dr. Anders Knudby for the kindness, continued support, and academic guidance he has extended to me throughout my undergraduate and master's degrees. Working with Dr. Knudby towards the end of my undergraduate degree was largely the reason I decided to pursue graduate school, so I am grateful to him for introducing me to the academic possibilities within this exciting field.

I would also like to thank my committee members Dr. Michael Sawada and Dr. Jeremy Kerr. I am grateful to have been able to take GIS courses with Dr. Sawada during my studies here at the University of Ottawa, and I am thankful to him for his time and valued feedback on my work over the last couple of years. I am also grateful to Dr. Jeremy Kerr for providing me with the opportunity to explore the ecological side of my otherwise geography-based work, and for the time he has taken to provide guidance on my work.

I would like to express my sincere gratitude to my friends and family who have supported me in countless ways. To my close friends I have gone through graduate school alongside, I am forever thankful for the shared support and encouragement you provided. Michelle, thank you for making my time here infinitely better, our friendship has kept me sane throughout our countless classes (and now two degrees) together. Edmar, thank you for always being generous with your help, especially when I was struggling with my maps. Miranda, thank you for your continuous encouragement, you have been a wonderful colleague.

Finally, I would like to thank Fisheries and Oceans Canada, the Province of Ontario, the University of Ottawa, and the North Pacific Marine Science Organization for providing funding which helped make this thesis possible and for introducing me to the world of travelling for academic conferences.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area under the receiving operating characteristic (ROC) curve
BRT	Boosted regression trees/generalized boosting model
CI	Confidence interval
EEZ	Exclusive economic zone
FDA	Flexible discrimination analysis
GAM	Generalized additive model
GLM	Generalized linear model
MARS	Multiple adaptive regression splines
MaxEnt	Maximum entropy distribution modeling
MPA	Marine protected area
PA	Presence absence data
RF	Random forest
ROC	Receiving operating characteristic curve
SDM	Species distribution model
SD	Standard deviation
SE	Standard error
TSS	True skill statistic

Chapter 1. Introduction

Biogeographical patterns for benthic marine taxa are poorly understood due to a lack of accessible geospatial information. Knowledge of the spatial distribution of species is a crucial prerequisite for the understanding of ecosystem functioning and processes as well as conservation management (Reiss et al., 2011). Fauna is more difficult to access and monitor in marine environments than in terrestrial environments. Due to the resulting limitation of data on marine taxa, predictive modeling methods are often employed in order to estimate their full distribution from the available data (Guisan et al., 2000; Guisan et al., 2005). Species distribution models (SDMs) used for this purpose, also referred to as habitat suitability models or ecological niche models, estimate the relationship between species' coordinate data and the environment within which they exist (Franklin, 2009; Elith et al., 2011). While terrestrial SDM work is a fairly robust field, marine applications of SDM have been more recent in their developments (Reiss et al., 2011; Robinson et al., 2011). The field of marine SDM has been stimulated by increasingly available large-scale environmental data on ocean biogeochemistry and the need for prediction methods to quantify and estimate changes in species distribution in response to climatic changes (Reiss et al., 2011). However, a systematic review of 236 published papers on marine-based SDMs (Robinson et al., 2017) noted some shortcomings typical in the field. Only 9% of the reviewed studies tested their model results against independent data, which is generally accepted as an unbiased method of assessing model performance, and 94% of the reviewed studies failed to report the amount of uncertainty derived from data deficiencies and model parameters. When model predictions are evaluated against independent data, there is usually no spatial component to the summary statistics or visualization of patterns of uncertainty such as spatial clustering or links with specific predictor variables (Elith et al., 2002b). A popular method of interpreting and calculating prediction uncertainty involves measuring similarities between distribution predictions, where multiple taxa, models, areas and/or spatial resolution are tested and the resultant predictions are compared (Monk et al., 2012; Pennino et al., 2016; Svensson et al., 2013). Besides similarity calculations, measuring and quantifying prediction uncertainty is an underdeveloped aspect of marine SDM work, and, along with testing several SDMs, will be explored in this thesis.

This thesis will focus on hexactinellid sponges in the North Pacific Ocean. The class Hexactinellida (kingdom Animalia, phylum Porifera), consists of between 400-500 species in two subclasses which contain five orders, 17 families, and 118 genera (Reiswig et al., 1983). They are often referred to as glass sponges because their skeletons are composed of spicules of silica. Hexactinellid sponges are sessile, relying on filter-feeding to obtain the macroscopic detritus material they subsist on (Atwater et al., 2001). After hatching, sponge larvae drift in the water column for a limited time before settling on the seafloor as juvenile sessile sponges (Maldonado, 2006). Observations of planktonic larval life in laboratories indicate that most larvae are anchiplanic, which means they generally remain in the water column for minutes to a few days, usually less than two weeks (Maldonado, 2006). Larvae are known to disperse under the influence of hydrodynamic processes that operate at a spatial scale of tens of meters to kilometers, and are not thought to be affected by active substratum selection, which operates at a smaller scale of centimeters to meters (Maldonado, 2006). Little is known about dispersion in specifically hexactinellid sponges, and limited information about species within the hexactinellid class is available in the dataset used for this thesis. Therefore it must be mentioned that this thesis operates under the assumption that glass sponges of different species react similarly to their environment, due to a lack of more specific data.

When the sponges die, their siliceous skeletons remain and future sponges grow directly on them, forming reefs. Their skeletons have left a fossil record as far back as the Cambrian/Pre-Cambrian, which would make them possible the earliest living metazoans on earth (Leys, 2003). While they are found in every ocean in the world, they remain a rare taxa with a seemingly specific set of environmental conditions required to thrive. Research on deep-sea reef-forming benthic taxa is crucial as they are important indicators of the health of benthic ecosystems and often increase biodiversity where they are found (Knudby et al., 2013; Beazley et al., 2013). Their high diversity, large biomass, complex physiology and chemistry, and long evolutionary history lend sponges (and their endosymbionts) to play a key role in diverse ecological processes, including but not limited to predation, habitat provision, nutrient cycling, food chains, and bioerosion (Rützler, 2004). Unfortunately, benthic marine taxa are vulnerable to climate change as well as human activities such as fisheries, specifically deep-sea trawling (Rooper et al., 2017). Glass sponge reefs in the waters off the coast of British Columbia have recently been permanently protected from trawling and other human activities through the establishment of marine protected areas (MPAs). Since little is known about the distribution of this taxa outside British Columbia coastal waters, applying SDMs to glass sponges throughout the North Pacific will shed light on this otherwise difficult-to-research taxa.

1.1 Thesis Objectives

The primary goals for this research are threefold;

1) To map hexactinellid (glass) sponge distribution for the entire North Pacific Ocean, as well as several smaller areas, by testing several commonly used species distribution modeling methods,

2) To assess the model outputs both in terms of the relative importance of different environmental variables in making predictions about glass sponge presence/absence, as well as the specific dependence of glass sponge presence probability on these environmental variables, and

3) To compare existing methods for mapping prediction uncertainty.

1.2 Thesis Format

This thesis is presented in article format, including a manuscript intended for publication. Chapter 1 provides an introduction for the thesis, as well as covering the thesis objectives and the format. Chapter 2 presents a manuscript titled “Predicting Glass Sponge (Porifera, Hexactinellida) Distributions in the North Pacific Ocean and Spatially Quantifying Model Uncertainty”. This manuscript will be submitted to a journal in the field of applied marine ecology and ecological modeling for publication such as *Deep-Sea Research Part I: Oceanographic Research Papers*. Chapter 3 provides a general conclusion of the manuscript presented in Chapter 2. Chapter 4 contains the literature cited for this thesis.

Chapter 2: Predicting Glass Sponge (Porifera, Hexactinellida) Distributions in the North Pacific Ocean and Spatially Quantifying Model Uncertainty

2.1 Introduction

2.1.1 Species Distribution Modeling in Marine Environments

Species distribution modeling allows for the understanding of processes that create habitat distribution patterns, and has become increasingly important in the face of threats such as habitat destruction, species invasions, pollution and climate change (Robinson et al., 2011). SDM algorithms require high-quality species presence/absence records as well as high-quality environmental information to infer the macroecological preferences of species (Tyberghein et al., 2011). By transferring SDMs from terrestrial to marine environments, the validity of the model and its predictive performance will be affected by the unique physical properties of marine habitats (Robinson et al., 2011). This is largely due to the fact that marine ecosystems have significantly less permanence than terrestrial ecosystems; for example, a treeline or grasslands may remain stable during a timeline of decades, while ecological and physical conditions in the water are in continual flux (Longhurst, 2007).

2.1.1 Existing Guidance on Model Selection

Model complexity has increased greatly over time from environmental matching (e.g. BIOCLIM, DOMAIN) to more complex non-linear relationships between species and their environment (e.g. generalized additive models (GAMs), MaxEnt) (Elith et al., 2009). BIOCLIM is an early SDM package which relates the bioclimatic environment species exist within to a number of environmental predictor variables, such as temperature or elevation (Booth et al., 2013). Due to the now numerous SDM methods, there is some difficulty in selecting an appropriate algorithm. The advice that would assist making an informed choice of method is currently scattered throughout literature (Elith et al., 2009). It remains difficult to know which model is ‘best’ for the given data before comparing multiple models. This thesis will therefore focus on several commonly used SDMs.

Input data required for SDM work involves biological data: information about the species (single or multiple species) distribution, and environmental data: usually raster data describing the landscape the species is found within (Pearson, 2010). Biological data can be obtained in numerous ways: from surveys, museum collections, or personal collection and may be *presence-only* (PO, coordinates of where the species has been observed), or *presence/absence* (PA, coordinates of where the species has not been observed). Generally models are thought to have more ecological validity when fit with PA data as opposed to PO data, however the quality of absence data is often questioned due to possibility of 'false absences', which refers to instances when a species was present but not detected, or the environment was suitable but the species was absent (Pearson, 2010). Environmental data refers to predictor variables depicting climate, topography, land cover and vegetation, substrate, and other physical and chemical attributes of the area being modeled (Franklin et al., 2010). Spatial scale is often considered when collecting data and has two components: extent and resolution. Spatial extent refers to the size of the area being modeled and spatial resolution refers to the size of grid cells of the data. It is often common for datasets with large extents to have coarse resolution, and small extents to have high resolution (Pearson, 2010). As with other deep-sea species modeling efforts, due to the lack of information available concerning the niche environmental preferences of the relevant taxa it is difficult to ascertain the importance of individual environmental variables prior to modeling. When working with taxa for which there are limited data, environmental input layers are by necessity often selected primarily based on their availability and presumed relevance, and less important variables can be identified and removed during the modeling process.

Statistical Models

The Linear Model

Linear multiple regression models predict the response variable (Y) from a vector of multiple predictor variables, X = (X_1, X_2,…,X_p):

$$\widehat{Y} = \widehat{\beta}_o + \sum_{j=1}^{p} X_j \, \widehat{\beta}_j + \varepsilon \qquad \text{(Eq. 1)}$$

where $\hat{\beta}$ is the vector of estimated coefficients and $\hat{\beta}_o$ is an estimated constant known as the intercept (Franklin et al., 2010). The error term, ε, is normally distributed with zero mean and constant variance, and the variance of Y is constant across observations (Franklin et al., 2010).

Generalized Linear Models (GLMs)

While Franklin & Miller (2010) note that ecological data often violate the assumptions of the linear model, GLMs are often used in modeling and can be described as extensions of the linear model that can cope with non-normal distributions of the response variable (Venables et al., 1994). Distributions that are often used to characterize response variables in ecology include Gaussian, Poisson, binomial, negative binomial, and gamma (Franklin et al., 2010).

The linear model can be generalized using a link function that describes how the mean of Y depends on linear predictors, and a variance function that describes how the variance of Y depends on its mean (Chambers et al., 1992). The equation for the GLM can be seen in Equation 2:

$$\delta\big(\mathrm{E(Y)}\big) = \mathrm{LP} = \hat{\beta}_o + \sum_{j=1}^{p} X_j\, \hat{\beta}_j + \varepsilon \qquad \text{(Eq.2)}$$

where the predictor variables (far right side of the equation) are combined to produce a linear predictor, LP, and the expected value of Y, E(Y), is related to the LP through the link function, δ() (Franklin et al., 2010). Formulating a GLM for SDM involves selecting the response distribution and the link function (collectively known as the family of the GLM), the variance function, and the predictors (Franklin et al., 2010). The link function describes how the mean of Y depends on the linear predictor. For a binary response variable, a binomial distribution and logit link function are used.

Generalized Additive Models (GAMs)

Generalized additive models (GAMs) differ from GLMs in their ability to identify and describe a non-linear relationship between response and predictor variables; they are non-parametric extensions of GLMs (Franklin et al., 2010).

$$\delta\big(E(Y)\big) = LP = \hat{\beta}_o + \sum_{j=1}^{p} X_j\, f_j + \varepsilon \qquad \text{(Eq. 3)}$$

where the coefficients of the GLM are replaced by a smoothing function, f (Franklin et al., 2010). The fit of a GAM model is generally evaluated by testing the non-linearity of a predictor versus the non-parametric fit (Franklin et al., 2010). GAMs are used for characterizing non-linear response curves of species because they can suggest the shape of the parametric response curve and are thus more flexible than GLMs (Franklin et al., 2010). GAMs are popular in SDM work because they tend to have high prediction accuracy, they have been subjected to comparisons with other models and have proven to be useful (Franklin et al., 2010).

Machine Learning Models:

Maximum Entropy (MaxEnt) Distribution Modeling

The MaxEnt model was created in order to make predictions and inferences from incomplete data (Phillips et al., 2006), for example presence-only data. MaxEnt is one of the most common forms of SDM and "has been described as especially efficient to handle complex interactions between response and predictor variables" (Fourcade et al., 2014). MaxEnt is an acronym created for the concept of maximum entropy modeling (Guinotte et al., 2014), which extrapolates the likelihood a species has of existing in any specific geographic space. This can also be defined as a measure of dispersiveness. The underlying principle is that one should assume uniform distributions are preferred, given certain constraints (Nigam et al., 1999). Since becoming available in 2004, MaxEnt has been used to publish diverse projects including: finding correlates of species occurrences, mapping current distributions, and other related tasks in ecological, evolutionary, conservation and biosecurity applications (Elith et al., 2011).

MaxEnt has often been explained as estimating a distribution across geographic space (Phillips et al., 2006). Elith et al. (2011) give an alternative view: a characterization that focuses on comparing probability densities in covariate space. Their research examines how MaxEnt can be understood by looking at Bayes' rule:

$$\Pr(\gamma = 1|z) = f_1(z)\Pr(\gamma = 1) \;/\; f(z) \qquad \text{(Eq. 4)}$$

where $\gamma = 1$ indicates presence, $\gamma = 0$ indicates absence, z indicates a vector of environmental covariates. It must be assumed that all environmental variables z are available landscape-wide, and L is the extent of the landscape. $f(z)$ can be defined as the probability density of covariates across L, $f_1(z)$ can be defined as the probability density of covariates across locations within L where the species is present, and $f_0(z)$ can be defined as where the species is absent (Elith et al., 2011). The quantity to be estimated is the probability of presence of the species, conditioned on the environment: $\Pr(\gamma = 1|z)$.

Equation 4 can theoretically be explained by the following: that if the conditional density of the covariates at presence sites is known, $f_1(z)$, and if the unconditional density of covariates across the study area is known, $f(z)$, the prevalence $\Pr(\gamma = 1)$ is the only remaining value necessary to calculate the probability of occurrence (Ward, 2007; Elith et al., 2011). First, MaxEnt's core output involves estimating the ratio $f_1(z)/\, f(z)$. This gives insight about which features are important and how suitable one place is compared to another, which is the core of the MaxEnt model output. This explanation of MaxEnts' structure by ecologists rather than statisticians can be helpful in understanding the complicated processes that the data undergo.

Boosted Regression Trees (BRT)

Boosted regression trees (BRT) is an ensemble method for fitting statistical models that differs from conventional techniques to fit a single parsimonious model; BRTs combine the strength of two algorithms: regression trees and boosting (Elith et al., 2008). Regression trees are models that relate a response to their predictors by recursive binary splits, and boosting is an adaptive method which combines simple models to give improved prediction performance (Elith et al., 2008).

The decision trees in BRT are tree-based models which partition the predictor space into rectangles, doing this using a series of rules to identify regions having homogeneous responses to predictors (Elith et al., 2008). Then, a constant is fitted to each region, with regression trees fitting the mean response for observations in that region. Fitting a single decision tree is often done by growing a large tree and afterwards pruning it by collapsing the weakest links (identified through cross-validation) (Elith et al., 2008). Decision trees are popular because they allow for information to be represented in an intuitive manner that is easy to visualize. Trees are insensitive to outliers and are able to accommodate missing data in predictor variables by using surrogates (Breiman et al., 1984).

2.2 Study Area and Data

2.2.1 Study Area

The Pacific Ocean, the largest and deepest of the earth's oceans, is about 15 times the size of the United States, and is almost equal in area to the total land area of the world. The ocean can be divided by the equator into two separate areas: the north and south. The study area for this project is contained by the boundaries of the North Pacific Ocean: bordered by the Arctic Ocean in the north, Asia in the west, the Americas in the east, and the equator in the south. It provides habitat for thousands of species, including cold-water sponges and corals. Due to the size of this study area, five sub-areas were delineated within the North Pacific to account for the likely varying physical and chemical environments across an area as large as the North Pacific. These sub-regions within the North Pacific include a) The US Exclusive Economic Zone (EEZ) around Alaska, b) The Canadian EEZ around British Columbia, and c) The US EEZ around the Washington-Oregon-California coast, as well as d) two smaller areas within the Canadian EEZ which were manually delineated but roughly correspond to i) Hecate Strait and ii) the shelf waters west of Vancouver Island (Figs. 1-4). It is likely that there are varying environments within an area as large as the North Pacific, as well as varying groups of sponges. Using a multi-area analysis ensures a more comprehensive attempt at capturing these potentially different species-environment relations.

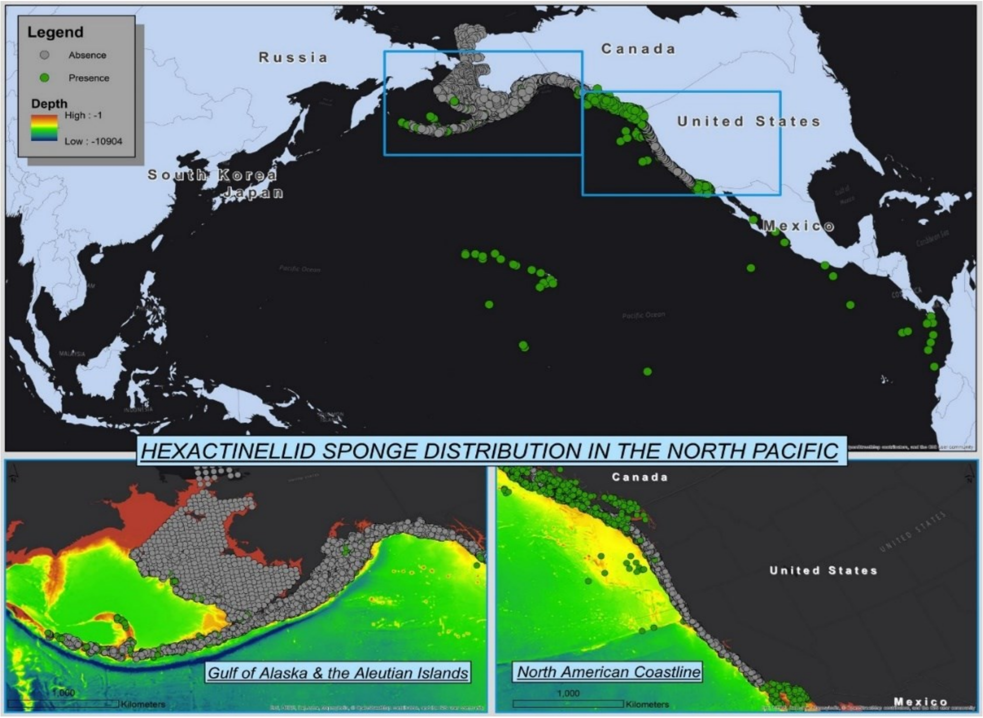


Figure 1. Hexactinellid Sponge Distribution in the North Pacific with insets of the Gulf of Alaska and British Columbia coastline.

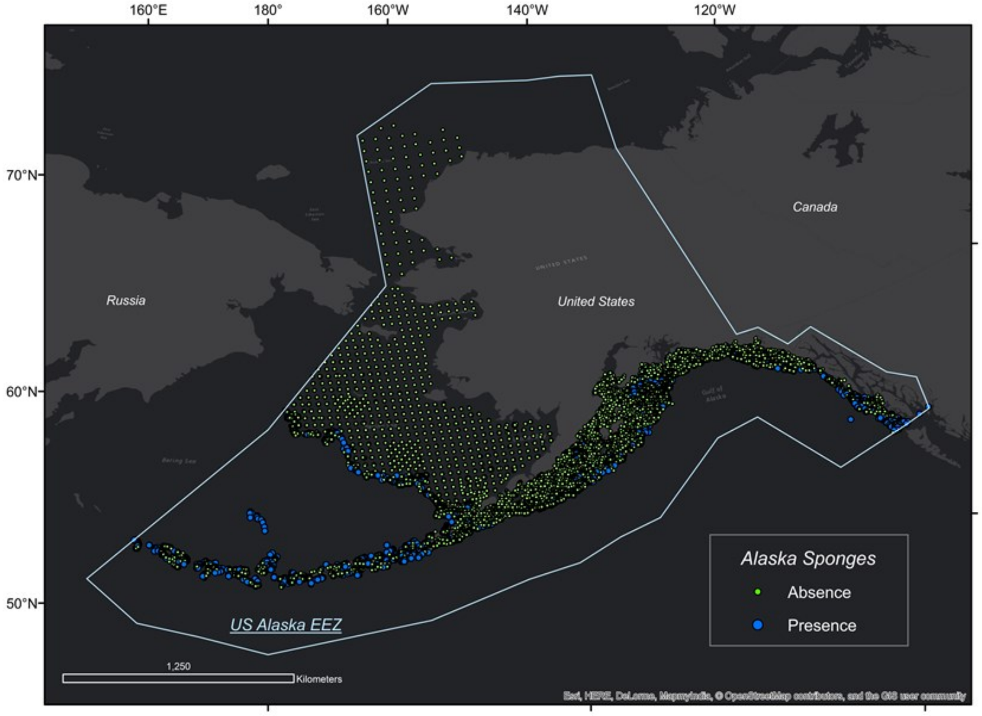


Figure 2. Alaska sub-area with contained sponge presence-absence (PA) data.

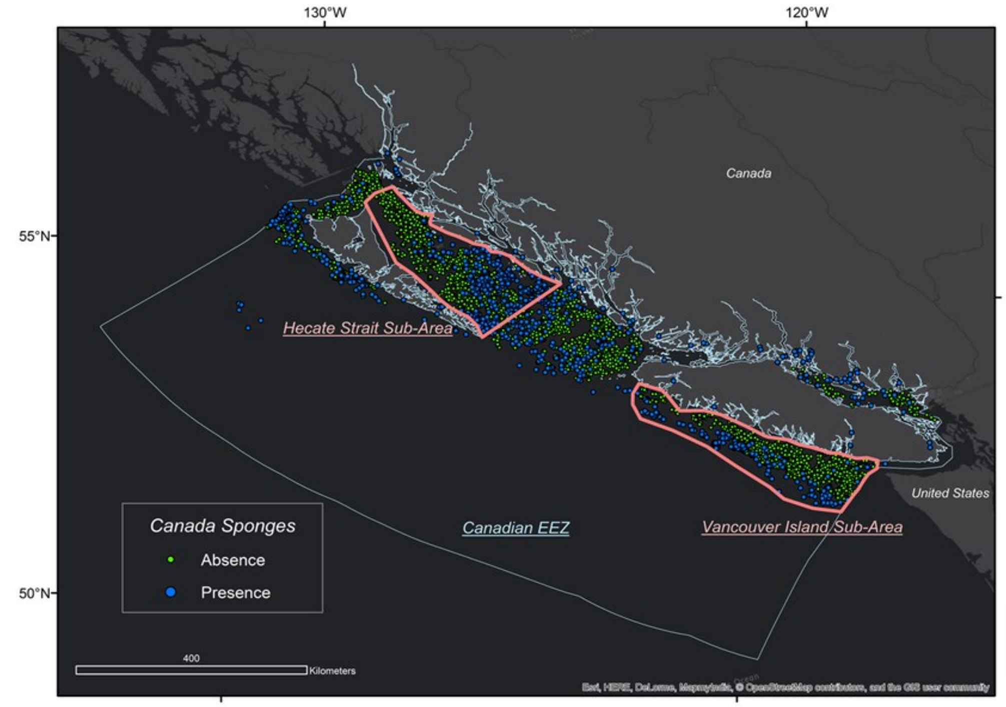


Figure 3. British Columbia, Hecate Strait and Vancouver Island sub-areas with contained sponge PA data.

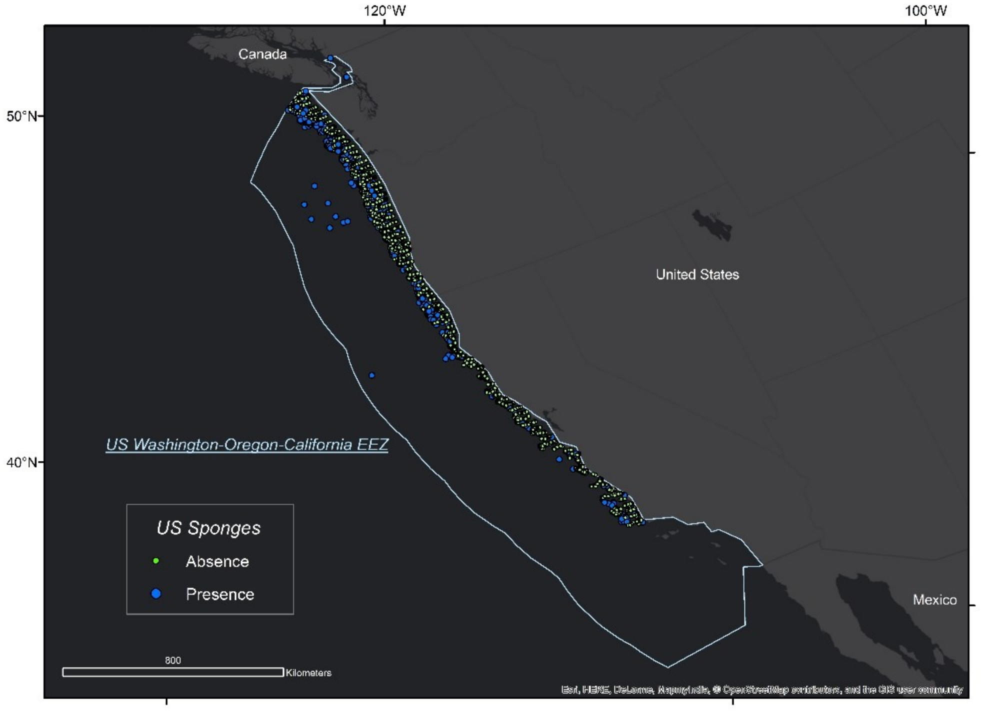

120°W
100°W
50°N
40°N
Canada
United States
Mexico
US Washington-Oregon-California EEZ
US Sponges
● Absence
● Presence
800
Kilometers
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

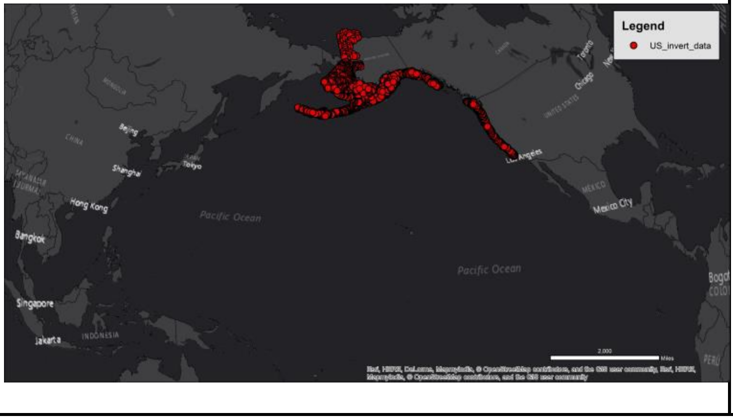
Figure 4. United States Washington-Oregon-California sub-area with contained sponge PA data.

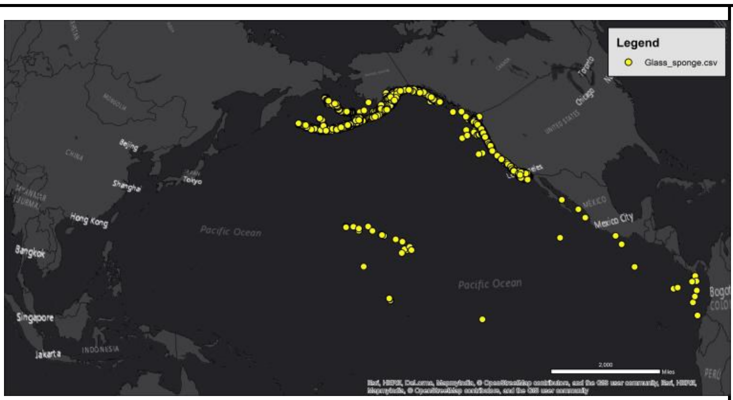
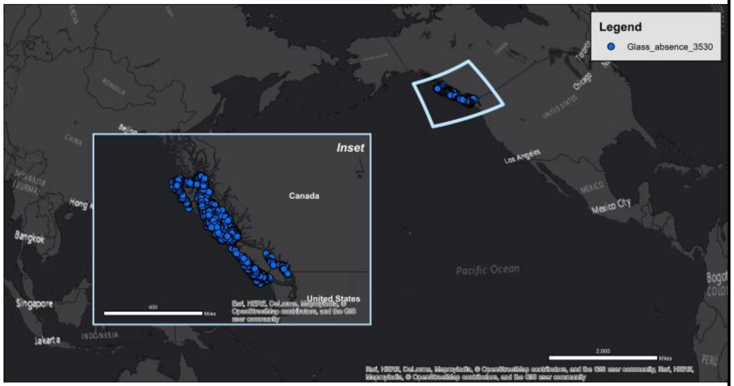
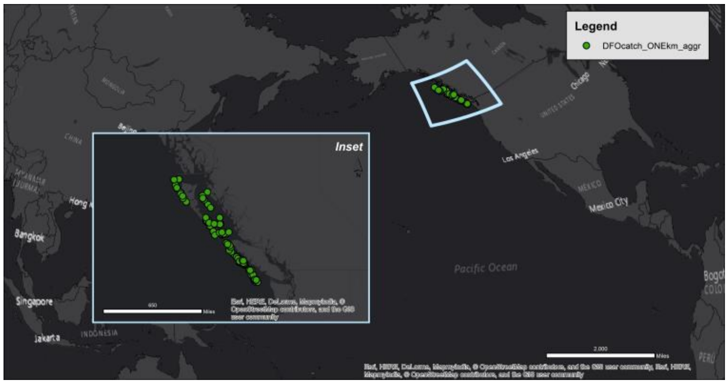
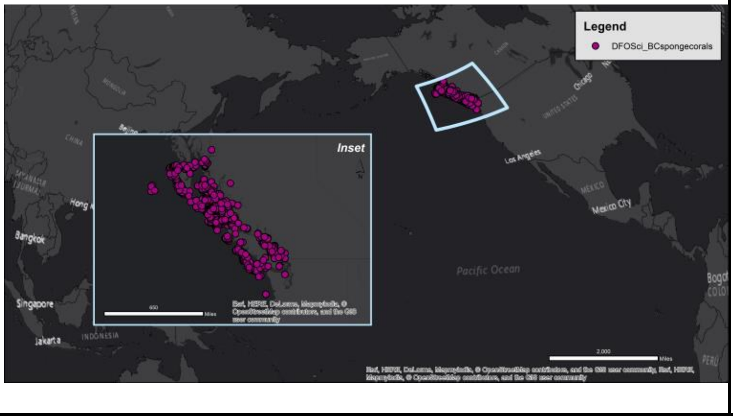
2.2.2 Biological Data

Presence and absence glass sponge data were obtained from trawl surveys conducted by Fisheries and Oceans Canada (DFO) and the US Government. Data from several surveys were collected and merged to create a dataset containing both presence and absence information for 42,113 coordinate locations sampled between 1996 and 2016. The dataset contains 16,148 presence points and 25,684 absence points. As can be seen in Figures 1–4, and Table 1, the species presence/absence points are located largely in coastal waters along the coast of North America, and out along the Aleutian Islands, with a few data points from Hawaii. No data are available from the western North Pacific.

In an attempt to decrease sample bias, the original dataset of 42,113 coordinate points was thinned based on environmental variation (See Methods section for more detail). After the data was thinned based on local environmental variation of the predictor variables, the resultant dataset had 12,467 sponge presence absence points.

Table 1. Hexactinellid sponge data: location of data points and source.

Data	Geographic Extent	No. of Presences	No. of Absences
US bottom trawl surveys from Alaska (1996-2016) and US West Coast (1996-2004)	 Legend ● US_invert_data Beijing Shanghai Tokyo Hong Kong Bangkok Singapore Jakarta INDONESIA CHINA Pacific Ocean Pacific Ocean Los Angeles Chicago Toronto UNITED STATES MEXICO Mexico City PERU 2,000 Miles	1008	22,322

Presence data from North Pacific Marine Science Organization (PICES) Working Group 32	 Legend Glass_sponge.csv Pacific Ocean Pacific Ocean 2,000 Miles	14,134	0
DFO commercial bycatch logs	 Legend Glass_absence_3530 Inset Canada United States 400 Miles Pacific Ocean 2,000 Miles	0	3530
DFO commercial catch records aggregated to 1km grid	 Legend DFOcatch_ONEkm_aggr Inset 400 Miles Pacific Ocean 2,000 Miles	251	0
DFO research databases and museum records	 Legend DFOSci_BCspongecorals Inset 400 Miles Pacific Ocean 2,000 Miles	868	0

Total		42,113

2.2.3 Environmental Data

Environmental variables were selected based on availability and presumed likelihood of being relevant to the distribution of glass sponges. Potential environmental variables relevant to the distribution of glass sponges have been compiled from various sources through the North Pacific Marine Science Organization (PICES) Working Group 32 (Table 2). The environmental data layers are in a raster format, with a cell size of 1000m by 1000m, using an azimuthal equidistant projection with a central meridian of -180. The values reflect the near-sea floor part of the water column.

Table 2. Environmental variables, units and reference.

Variable Name	Units	Reference
Alkalinity	μmol l^{-1}	Steinacher et al. (2009)
Aragonite saturation state	Ω_{ARAG}	Steinacher et al. (2009)
Aspect	degrees	Becker et al. (2009)
Calcite saturation state	Ω_{ARAG}	Steinacher et al. (2009)
Depth	m	Becker et al. (2009)
Dissolved inorganic carbon	μmol l^{-1}	Garcia et al. 2014a
Eastness	degrees	Wilson et al. (2007)
Nitrate	μmol l^{-1}	Garcia et al. 2014b
Northness	degrees	Wilson et al. (2007)
Oxygen	ml l^{-1}	Garcia et al. 2014a
Phosphate	μmol l^{-1}	Garcia et al. 2014b
Roughness	unitless	Wilson et al. (2007)
Rugosity	unitless	Becker et al. (2009)
Salinity	pss	Zweng et al. 2013
Silicate	μmol l^{-1}	Garcia et al. 2014b
Slope	degrees	Becker et al. (2009)
Temperature	°C	Locarnini et al. 2013
TPI (Topographic Position Index)	unitless	Wilson et al. (2007)
TRI (Terrain Ruggedness Index)	unitless	Wilson et al. (2007)

2.3 Methods

Many of the choices in the following methodology section were made in an effort to maximize reproducibility of this study, however had other methods been selected, the results could have differed. Within the SDM field, reproducibility is a common problem. Advanced modeling techniques, data selection and processing require many choices to be made which decreases the replicability, yet is nonetheless common and relatively unavoidable in SDM.

2.3.1 Data Pre-Processing

Spatial sampling bias, a common problem in marine and terrestrial SDM, decreases the accuracy and interpretability of SDM outputs. Spatial filtering is a common method of removing spatial bias as a data pre-processing step. For example, Boria et al. (2014) filtered clustered data to discard any data point within 10 miles of another point, and Varela et al. (2014) applied an environmental filter which discards presence points that are too clustered in environmental space. The following steps were taken to spatially thin the data in a manner which takes into account the differences in areas with high environmental variation and areas with low environmental variation. The principle at the basis of this method is that areas with low environmental variation across space require less geographically dense data to cover environmental variability than do areas with high environmental variation, and they can therefore be thinned more than areas with high variation in an effort to reduce bias in the dataset.

1. The local standard deviation of each predictor variable was calculated for a 9x9km window centered on each cell, and then normalized to a 0-1 scale. The 19 normalized values were then added together to produce a single raster with a theoretical value range of 0-19, quantifying local environmental variation across the study area.
2. A histogram was plotted to view the frequency distribution of this local environmental variation (Fig. 5). If this histogram had been multimodal, spatial areas corresponding to each local maximum, i.e. clusters of low or high local environmental variation, could have been identified. However, the histogram was unimodal, so instead quintiles were calculated to separate the study area into 5 regions ranging from lowest to highest local environmental variation. The maximum value for each quintile can be seen in Table 3.

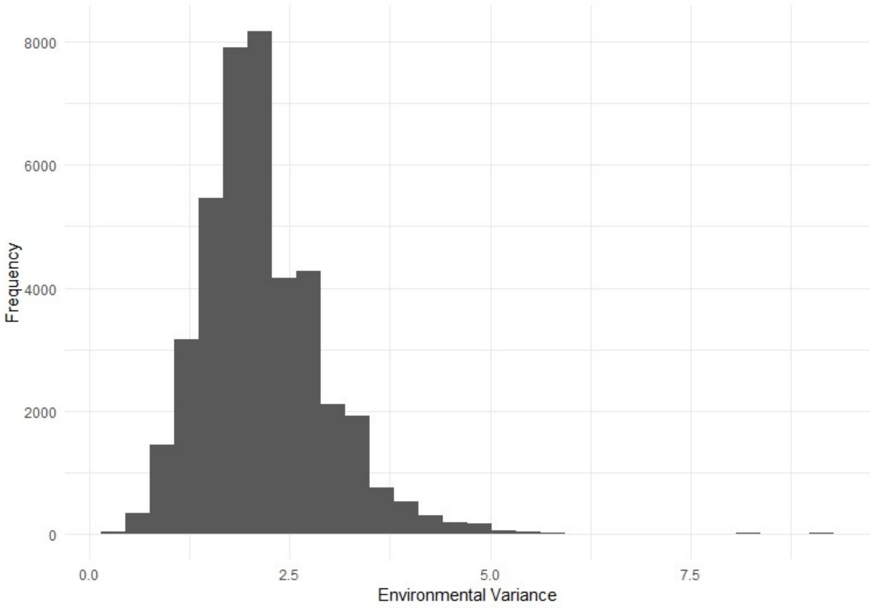


Figure 5. Histogram showing environmental variance data distribution.

Table 3. Quintile breaks for the variance data (environmental variance values from Figure 5).

20%	40%	60%	80%	100%
1.576018	1.893480	2.200223	2.730262	9.180244

3. Five subsets of the presence/absence data were then generated, one for each quintile, and semi-variograms were generated based on the bathymetric values from each subset. Depth was chosen to be the predictor for which to produce semi-variograms for several reasons: 1) depth often is one of the most important variables in any SDM for hexactinellid sponges in this study area, and 2) it can be used as a proxy for many other variables in this study.
4. The semi-variograms all used 1000m bins and a cut-off of 25,000m to ensure standardization (Fig. 6). For each plot, the distance at which semi-variance increased to more than 500 was noted. For the first quintile (representing the area with lowest local environmental variation) this distance was ~15,000m, for the second quintile it was ~7500m, and for the third, fourth and fifth quintile it was ~2500m. The semi-variance value of 500 was selected visually to provide a range of reasonable distances that were considered suitable to inform the scale of spatial thinning.

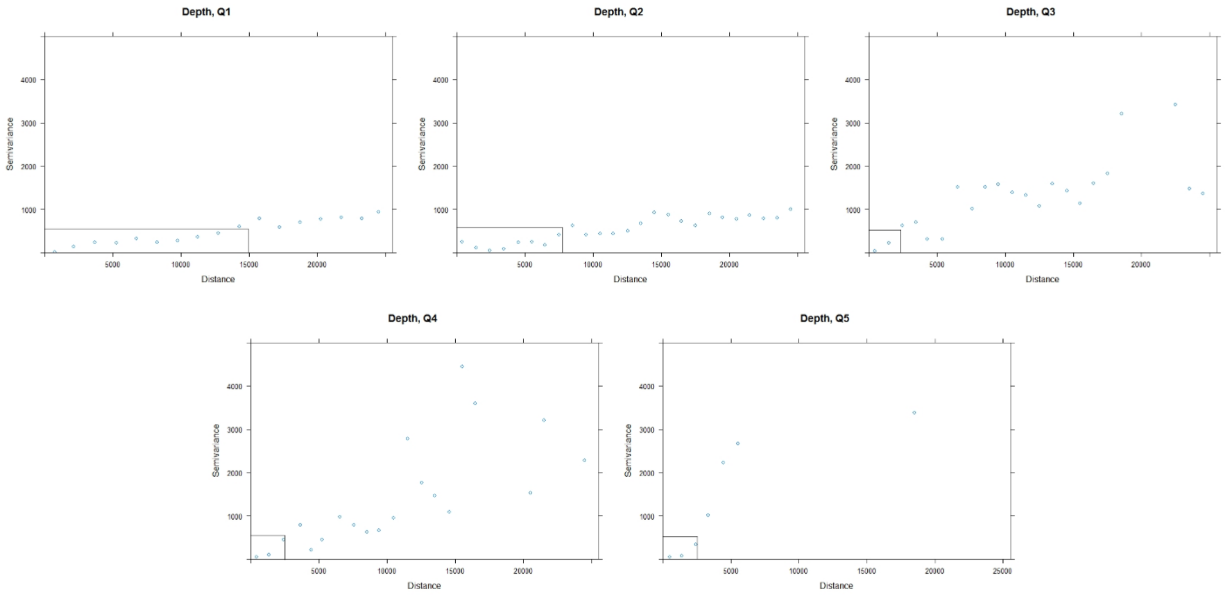


Figure 6. Semi-variogram plots showing five quintiles, with semi-variance value of 500 indicated to show approximate calculation of range value.

5. Based on these semi-variograms, block-based thinning was applied to the presence/absence data:
   i. A grid with 15,000m cell sizes was overlaid on the study area, and the mean value of the local environmental variation raster was calculated for each cell. For those cells falling in the first quintile, i.e. with mean local environmental variation <1.576 (see Table 3), the presence/absence data were aggregated according to the following rules:
      a) If no presence/absence observations were found in the cell, the output would be empty.
      b) If there were at least as many presence as absence observations in the cell, the output would be a presence point, located in the center of the cell.
      c) If there were more absence than presence observations in the cell, the output would be an absence point, located in the center of the cell.
   ii. This process was repeated with 7500m cells applied to points in the second quintile, and with 2500m cells applied to points in the third, fourth and fifth quintiles.

iii. The outputs were combined to produce a single set of spatially thinned presence/absence observations. Out of the original 42,113 presence and absence points, 12,467 remained after thinning. A sample section from the Aleutian Arc can be seen in Figure 7, showing the difference between the original points and the thinned points.

2.3.2 Species Distribution Modeling

For each of the predefined areas, the following SDM types were tested using the 'Biomod2' (Thuiller et al., 2016) package in R (R Core Team, 2013): GAM, BRT and MaxEnt (resulting in 18 model-area combinations). The following parameters for modeling were used:

- 'Number of Evaluation Runs: 3'; Running three evaluations means the calibration and evaluation is run 3 separate times independently, which allows for a more robust test of the models when independent data is not available.
- 'Data Split: 80%'; this sets 80% of the data aside for calibration of models, with the remaining 20% used for validation.
- Model accuracy measures: KAPPA, TSS, AUC; KAPPA refers to Cohen's Kappa Coefficient, and TSS to True Skill Statistic (Zhang et al., 2015). Both Kappa and TSS are threshold-dependent measures of model accuracy. They range from −1 to +1, where +1 indicates perfect agreement between predictions and observations and values of 0 or less indicate agreement no better than random classification (Landis et al., 1977). The Area Under the receiver operator characteristic Curve (AUC) is an effective, threshold-independent model evaluation indicator and is also independent of prevalence (i.e. the frequency of occurrence) of the target species (Zhang et al., 2015). Ranges used to interpret accuracy metrics from these statistics can be found in Table 4.

Table 4. Model accuracy ranges for AUC, Kappa and TSS measures (Zhang et al., 2015).

	Poor	Moderate	Good/Excellent
AUC	< 0.7	0.7 - 0.9	> 0.9
Kappa and TSS	< 0.4	0.4 – 0.8	> 0.8

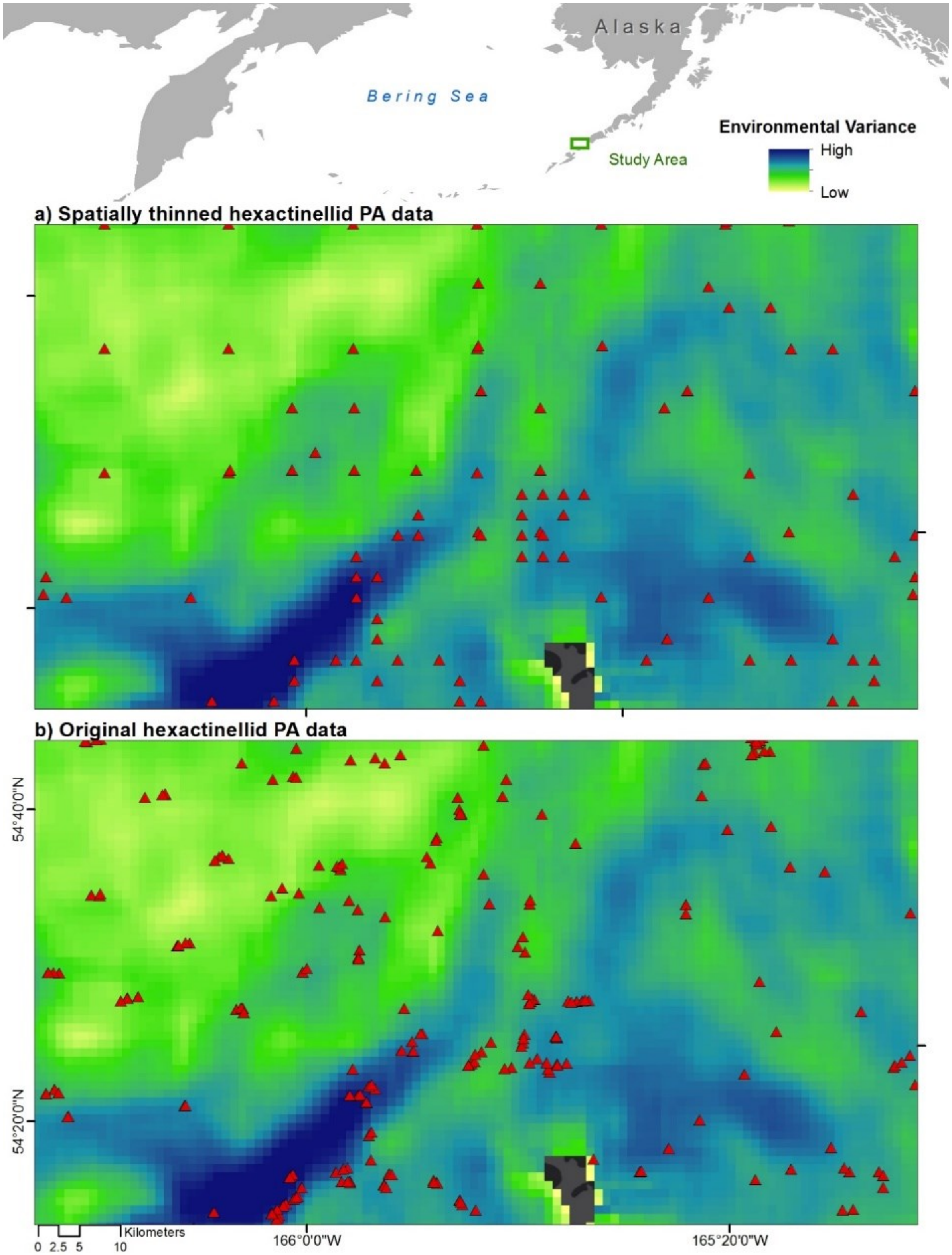


Figure 7. Sample area showing spatially thinned data and original data on the Aleutian Arc.

2.3.3 Interpretation of Model Results

Model output can be classified into two types: aspatial and spatial. Aspatial outputs from models consist of variable importances and partial dependence plots depicting fitted functions relating probability of occurrence to each selected predictor, while spatial outputs consist of GIS layers depicting the probability of presence across the study area (Ferrier et al., 2002).

Variable importance values were calculated for every model-area combination. These values quantify to what extent each predictor variable contributes to the predictions made by the model. The variable importance value is calculated as the result of one minus the correlation between the original prediction and the prediction with only the individual variable of interest. While the individual values are dependent on the algorithm used, they can still be used to provide relative information on predictor importances within the model.

Partial dependence plots provide a graphical representation of how likely the species is to be present, given a gradient of the specific environmental predictor. As with variable importance, these plots are calculated when the model is built, by averaging every other predictor variable except the one chosen predictor, and the change in model response is measured in relation to changes in the one variable. Partial dependence plots showing results from multiple algorithms can be used to visually compare species responses to environmental variable values. Partial dependence plots were generated for predictor variables that ranked in the top 25% in variable importance in several models and areas. The predictor variables chosen for closer analysis were alkalinity, oxygen, silicate, and phosphate. Additionally, frequency distribution plots were produced for alkalinity, oxygen, silicate, and phosphate in each area, showing the percentage of presence (as opposed to absence) data points across the range of environmental variable values. This information can shed light on which value ranges the species most commonly exists within. Partial dependence plots can be compared for similarities, variable importance was also considered in an effort to find strong trends in how predictors contribute to different models/areas. It can be posited that if the response curves of a particular predictor variable are similar across multiple models/areas, the variable importance is likely to be high as well. When the response curves vary significantly, it is more likely those variables ranked toward the lower range of variable importance for the model.

2.3.4 Mapping Prediction Uncertainty

For the purpose of testing and comparing uncertainty metrics spatially, a binomial GLM was fit to the Hecate Strait boundary (Fig. 3), a subset of the original, North-Pacific-wide dataset. 1,255 presence/absence points were included within the area and three environmental predictors with high variable importance were selected and clipped to the same extent; alkalinity, oxygen and silicate. A GLM was selected because it can provide a model-based uncertainty measure that can be mapped in addition to the actual model predictions. The logit-link function was selected for the binomial GLM because it is appropriate for binary data and ensures the predicted values will be between 0 and 1 (Kindt et al., 2005). Three spatially explicit uncertainty metrics were compared using this GLM model:

2.3.4.1 GLM Prediction Standard Error

The first uncertainty metric involved producing partial dependence plots from the GLM outputs and adding confidence intervals to the partial dependence plots. In order to obtain the predicted values from the estimates of the coefficients, the inverse link function needs to be calculated. Using the inverse link function, the confidence interval was calculated as the fitted value plus/minus two times the standard error on the link scale. CIs were calculated for alkalinity, oxygen and silicate. Adding confidence intervals can provide information on why certain areas would have predictions with high or low confidence. Next (and separate from the confidence intervals), the standard error of the prediction was calculated. Standard error provides the absolute measure of the typical distance between the data points and the regression line, in the units of the dependent variable. The standard error of the prediction was then written to a raster and thus the uncertainty of the model can be seen spatially.

2.3.4.2 Bootstrapped GLM Standard Deviation

The second uncertainty metric was obtained by bootstrapping the GLM. Bootstrapping is an approach to statistical inference based on building a sampling distribution for a statistic by resampling repeatedly from the data. 200 bootstrap samples were created from the data. GLMs were then calibrated on the bootstrap samples, still using alkalinity, oxygen and silicate as predictors for the models. The calibrated models were then used to make predictions, and the standard deviation was calculated for the predictions.

2.3.4.3 Standard Deviation of Multiple Model Predictions

The third uncertainty metric aims to test if standard deviations are geographically comparable among a variety of models. By running eight models available in the Biomod2 package (GLM, BRT, GAM, FDA, MARS, RF, MAXENT.Phillips and MAXENT.Tsuruoka) on the Hecate Strait subset, the standard deviation of all the predictions can be calculated, as for the GLM bootstrapping above. This provided a spatial view of where the models produced similar results and where they differed.

2.4 Results

2.4.1 Model Performance

Model fit statistics and variable importance values from the GAM, BRT, and MaxEnt models run on the North Pacific basin-wide data, as well as the five sub-areas are presented in Table 5. Figures 8-11 present the outputs of these models in the form of partial dependence plots and show the data distribution for alkalinity, oxygen, phosphate and silicate for each model/area. Only these four variables were selected because they had consistently high variable importance values.

As can be seen in Table 5, AUC values for the majority of the models were between 0.7-0.9. These values are interpreted to indicate these models performed moderately well (See Table 4 for value ranges associated with model accuracy) (Zhang et al., 2015). Two MaxEnt models for the BC and Alaska areas performed poorly, with AUC values of 0.655 and 0.428 respectively, and the GAM and BRT models for the Vancouver sub-area performed especially well, with AUC values of 0.946 and 0.978, the highest of all the models and areas. The Kappa and TSS values reported similar results in terms of models in the North Pacific, BC, and Alaska generally performing poorly, and models in the Vancouver and Hecate Strait sub-areas performing well (more detailed results for all models and areas can be found in Table 5).

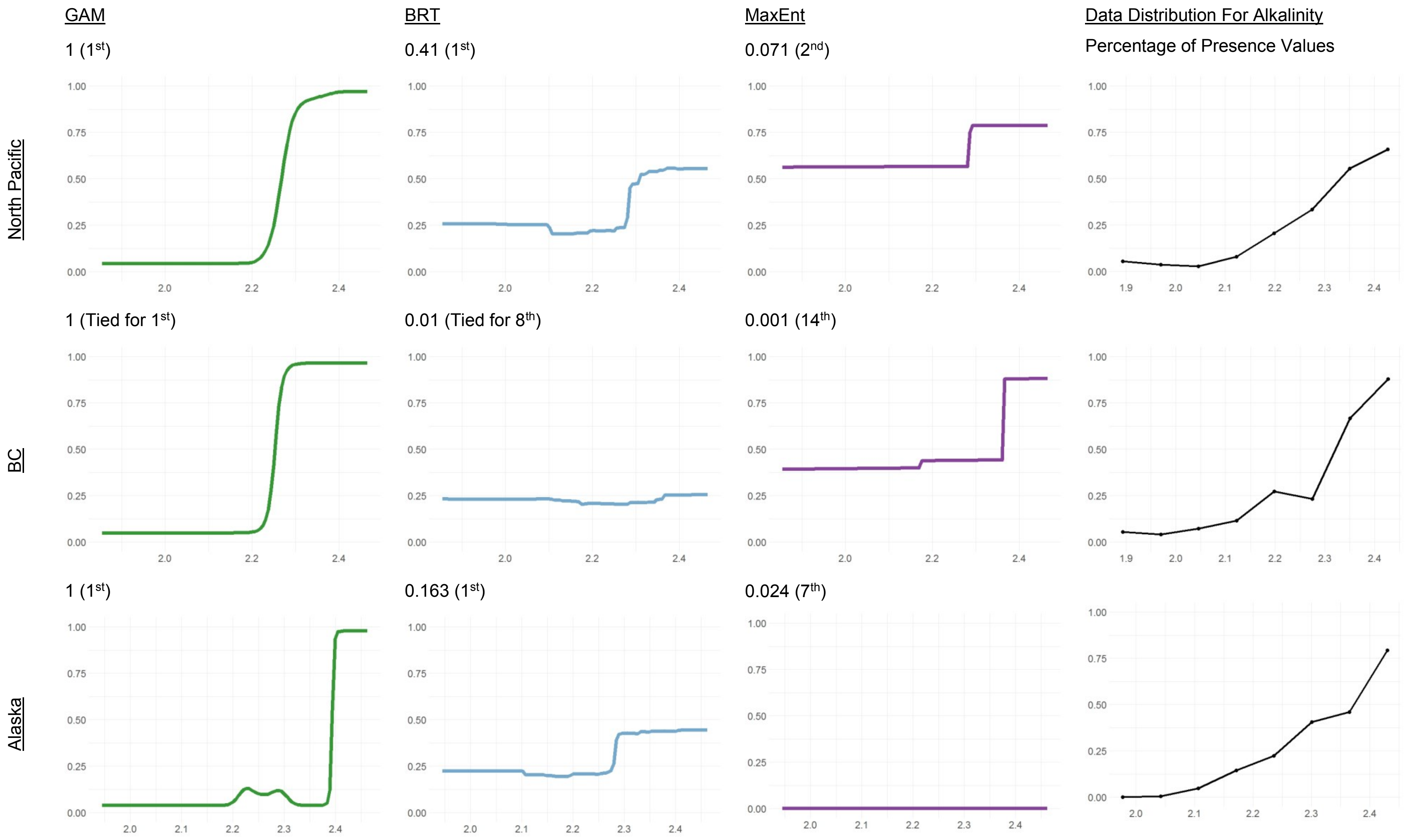
2.4.2 Variable Importance

For each area and model type, the model assigns a variable importance value to each of the 19 environmental predictors which were used as input to the model. The variables are arranged by importance to the model on a scale of 0-1. Individual variables ranked among the top 25% within each specific model are highlighted in Table 5. Alkalinity is ranked within the top 25% of variables in 13 out of 18 models. Oxygen is the next variable of highest importance being ranked within the top 25% of variables in 9 out of 18 models. Variables which are ranked within the top 25% in at least 4 of the 18 models include phosphate, silicate, temperature, nitrate, depth, omega aragonite, and omega calcite. The remaining variables are ranked within the top 25% for less than four models.

Table 5. Model Results, Fit Statistics and Variable Importance Values (top 25% of variables in each model are bolded in blue).

		North Pacific			***BC EEZ***			***ALASKA EEZ***			***US WOC EEZ***			***BC Hecate Strait***			***BC Vancouver Island***		
Model		GAM	BRT	MaxEnt	GAM	BRT	MaxEnt	GAM	BRT	MaxEnt	GAM	BRT	MaxEnt	GAM	BRT	MaxEnt	GAM	BRT	MaxEnt
Kappa		0.371	0.368	0.343	0.321	0.398	0.250	0.330	0.359	0.033	0.620	0.595	0.516	0.449	0.514	0.453	0.818	0.824	0.787
TSS		0.447	0.449	0.409	0.378	0.440	0.254	0.491	0.506	0.022	0.618	0.598	0.516	0.425	0.526	0.486	0.825	0.864	0.7
AUC Value		0.799	0.880	0.776	0.750	0.797	0.655	0.804	0.816	0.428	0.890	0.878	0.836	0.777	0.850	0.811	0.946	0.978	0.85
Variable Importance (Scale of 0-1)	Alkalinity	***1***	***0.410***	***0.071***	***1***	0.01	0.001	***1***	***0.163***	0.024	***1***	***0.22***	0.01	***1***	***0.136***	***0.137***	***1***	***0.013***	0.002
	Omega Aragonite	***0.973***	0.003	0.005	***0.703***	0.008	0	***0.837***	0.009	***0.05***	***0.679***	0.008	0.376	0.832	***0.048***	0.069	0.68	0.003	0
	Aspect	0.011	0.001	0	0.016	0.005	0	0.001	0.002	0.002	0	0	0.118	0.425	0.102	0.075	0.045	0.002	0
	Omega Calcite	***0.885***	0.009	0.015	***1***	0.021	0	***0.953***	0.014	0.023	***0.976***	0.011	0.219	***1***	0.005	0.046	***1***	0.001	0
	Depth	0.303	***0.025***	0.027	0.509	0.021	0.017	0.256	0.026	0.007	0.043	0.016	***0.7***	0.622	0.009	0.001	0.718	***0.101***	***0.166***
	Dissolved Inorganic Carbon	0.389	0.009	0	0.638	0.002	0.069	0.641	0.008	0.01	0.658	0.005	***0.629***	0.642	0.013	0.059	0.656	0.003	0.005
	Eastness	0.029	0.005	0.011	0.01	0.003	0.066	0.008	0.001	0	0.017	0.003	0.035	0.166	0.006	0	0.059	0	0
	Nitrate	0.367	***0.012***	***0.028***	0.559	***0.148***	***0.249***	0.35	0.021	***0.117***	0.2	0.027	0.322	0.099	***0.05***	***0.076***	0.609	0.003	***0.314***
	Northness	0.006	0.001	0	0.014	0.003	0	0.001	0	0	0.001	0	0	0.002	0.021	0.012	0.156	***0.016***	0.003
	Oxygen	***0.393***	***0.020***	***0.167***	0.48	0.007	0.111	0.113	***0.108***	***0.134***	0.405	***0.127***	0.239	***1***	0.011	0.074	***1***	0	***0.013***
	Phosphate	0.156	0.004	0.014	***1***	***0.121***	***0.13***	0.333	0.004	0.003	0.266	0.011	0.018	0.456	***0.048***	0.011	0.557	0	0
	Roughness	0.02	0.001	0.012	0.056	0.002	0.113	0.032	0.001	0	0.016	0.001	0.029	0.009	0.003	0	0.641	0.001	0.001
	Rugosity	0.003	0	0.004	0.05	***0.03***	***0.114***	0.025	0.001	***0.94***	0.015	0.003	0	0.029	0.006	0	0.443	0.001	0.001
	Salinity	0.021	0.008	0.014	0.633	0.018	***0.263***	0.116	***0.056***	0.021	0.017	***0.028***	***0.877***	0.589	0.011	0	0.562	0.002	0
	Silicate	0.241	0.002	***0.028***	0.482	0.002	0.071	***0.643***	***0.052***	0.003	0.504	***0.038***	0.001	***1***	0.008	0.028	0.821	0.001	0.009
	Slope	0.002	0.001	0.002	0.051	0.012	0.049	0.041	0.003	0.026	0.036	0.008	0	0.03	0.012	0	0.382	***0.053***	0
	Temp	0.257	0.002	0.001	0.211	***0.034***	0.059	0.144	0.011	0.041	***0.818***	0.019	***0.411***	0.599	0.002	0.034	***0.897***	0.001	0
	TPI	0.024	0.008	0.001	0.012	0.01	0.046	0.028	0.027	0.002	0.031	0.007	0.064	0.003	0.006	0.066	0.31	0.006	***0.031***
	TRI	0.003	0	0	0.004	0.001	0.037	0.032	0.001	0	0.257	0.001	0.021	0.009	0.025	***0.087***	0.449	0.004	0.002

Figure 8. Partial dependence plots for alkalinity in all areas and models with ranked variable importance.



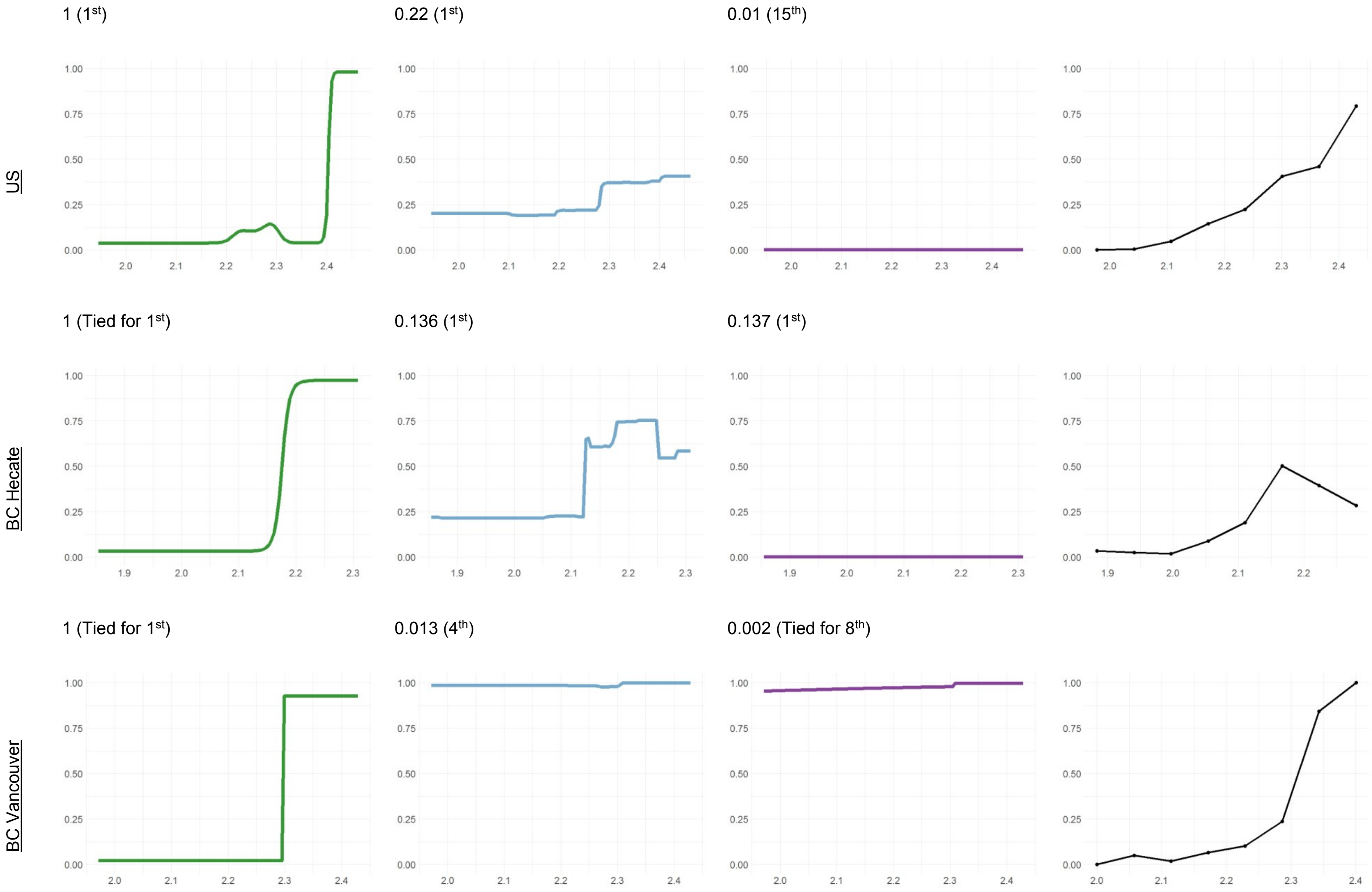
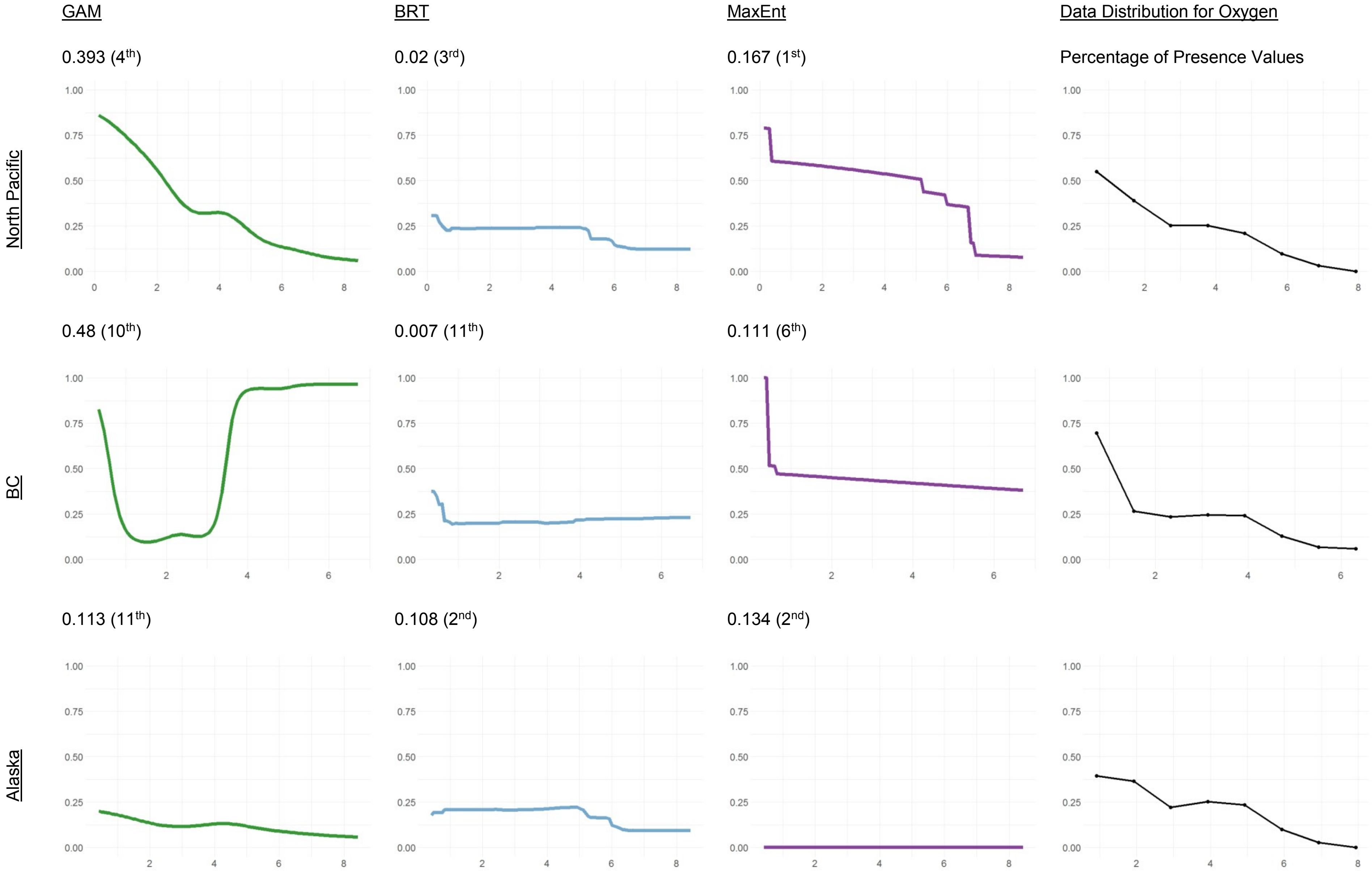

1 (1st)
0.22 (1st)
0.01 (15th)
US
1 (Tied for 1st)
0.136 (1st)
0.137 (1st)
BC Hecate
1 (Tied for 1st)
0.013 (4th)
0.002 (Tied for 8th)
BC Vancouver
1.00
0.75
0.50
0.25
0.00
1.9
2.0
2.1
2.2
2.3
2.4

Figure 9. Partial dependence plots for oxygen in all areas and models with ranked variable importance.



GAM

BRT

MaxEnt

Data Distribution for Oxygen

North Pacific

0.393 (4th)

0.02 (3rd)

0.167 (1st)

Percentage of Presence Values

BC

0.48 (10th)

0.007 (11th)

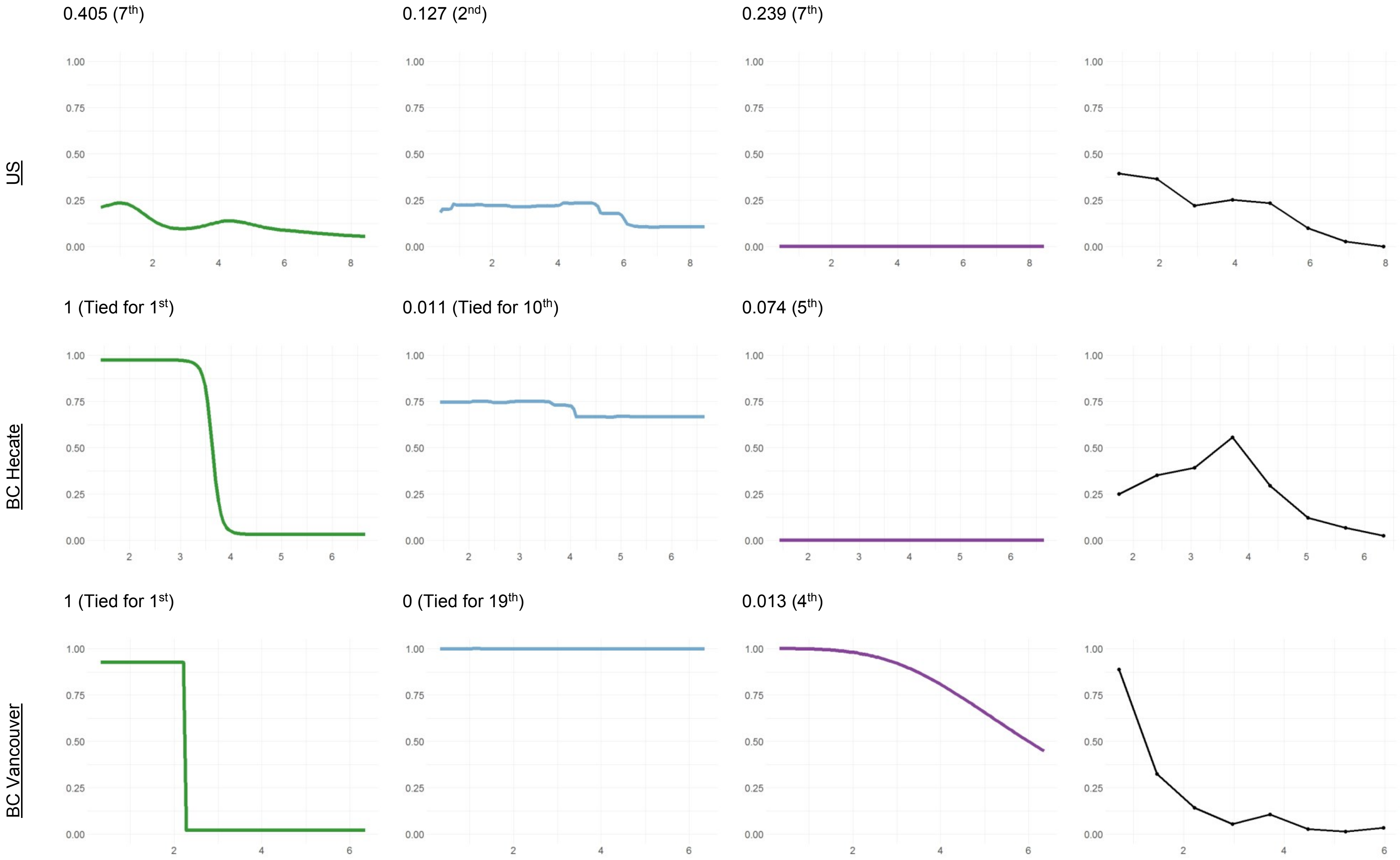
0.111 (6th)

Alaska

0.113 (11th)

0.108 (2nd)

0.134 (2nd)



0.405 (7th)

0.127 (2nd)

0.239 (7th)

1 (Tied for 1st)

0.011 (Tied for 10th)

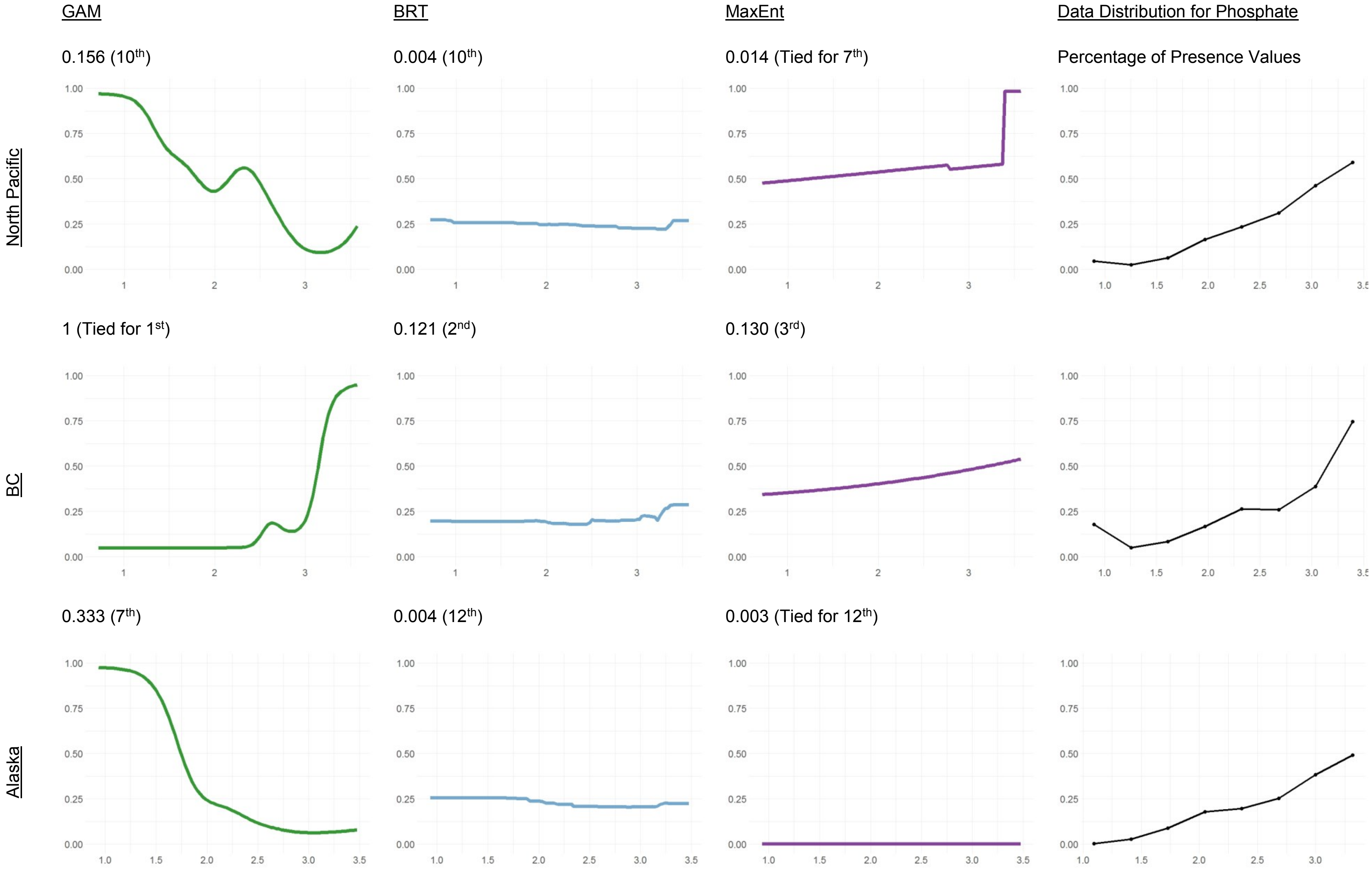
0.074 (5th)

1 (Tied for 1st)

0 (Tied for 19th)

0.013 (4th)

Figure 10. Partial dependence plots for phosphate in all areas and models with ranked variable importance.



GAM

BRT

MaxEnt

Data Distribution for Phosphate

North Pacific

0.156 (10th)

0.004 (10th)

0.014 (Tied for 7th)

Percentage of Presence Values

BC

1 (Tied for 1st)

0.121 (2nd)

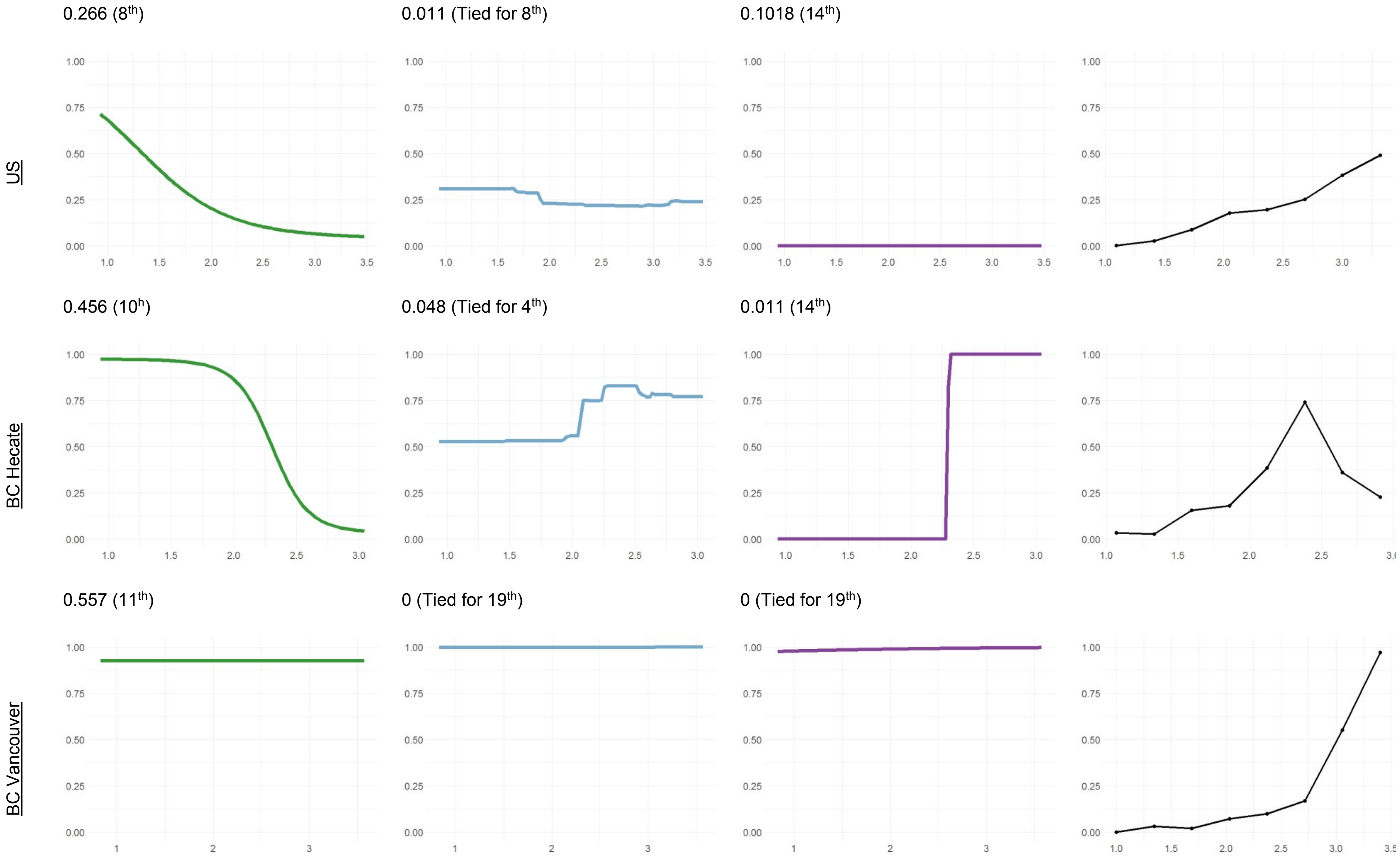
0.130 (3rd)

Alaska

0.333 (7th)

0.004 (12th)

0.003 (Tied for 12th)



0.266 (8th)

0.011 (Tied for 8th)

0.1018 (14th)

0.456 (10h)

0.048 (Tied for 4th)

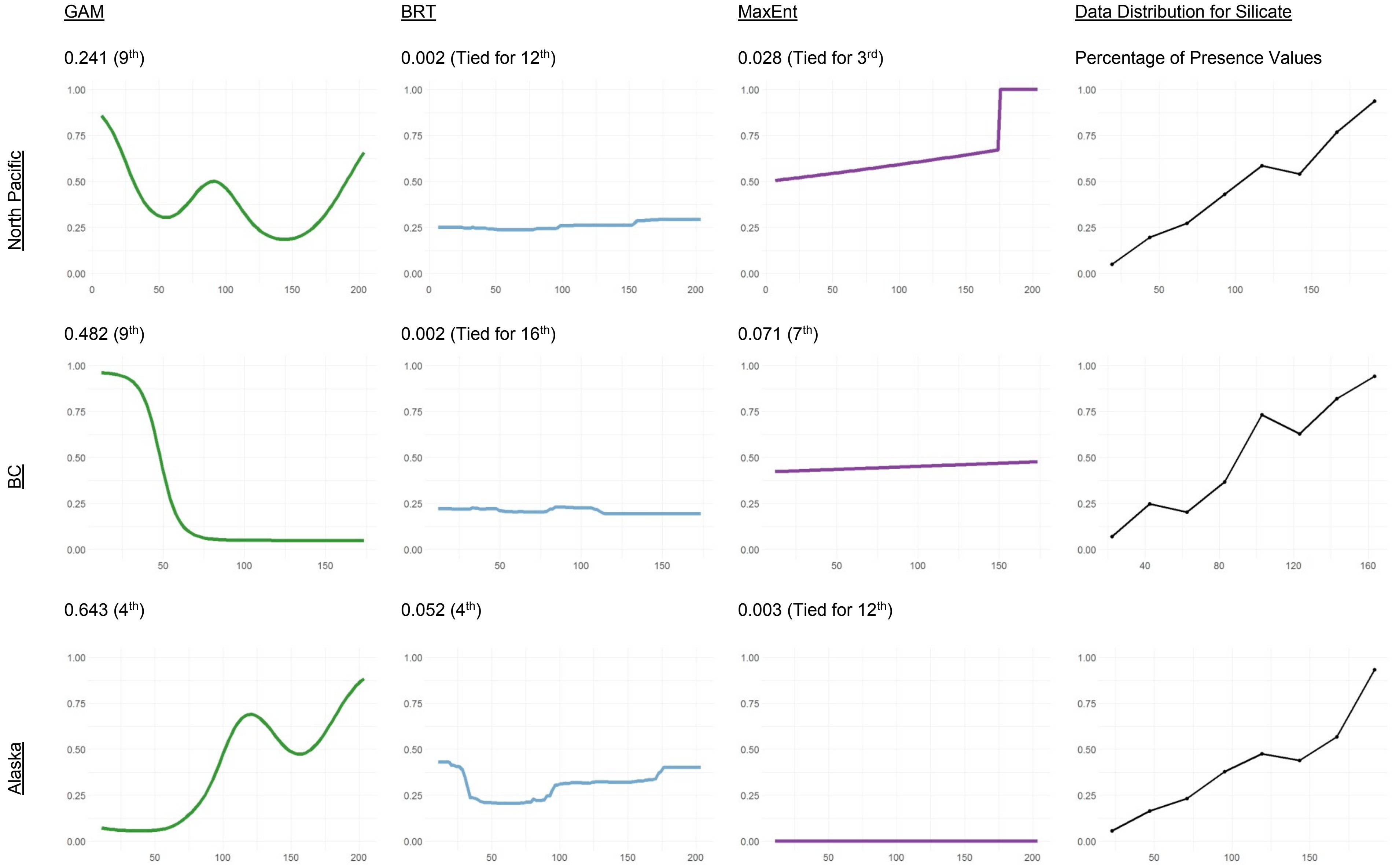
0.011 (14th)

0.557 (11th)

0 (Tied for 19th)

0 (Tied for 19th)

Figure 11. Partial dependence plots for silicate in all areas and models with ranked variable importance.



GAM

BRT

MaxEnt

Data Distribution for Silicate

0.241 (9th)

0.002 (Tied for 12th)

0.028 (Tied for 3rd)

Percentage of Presence Values

0.482 (9th)

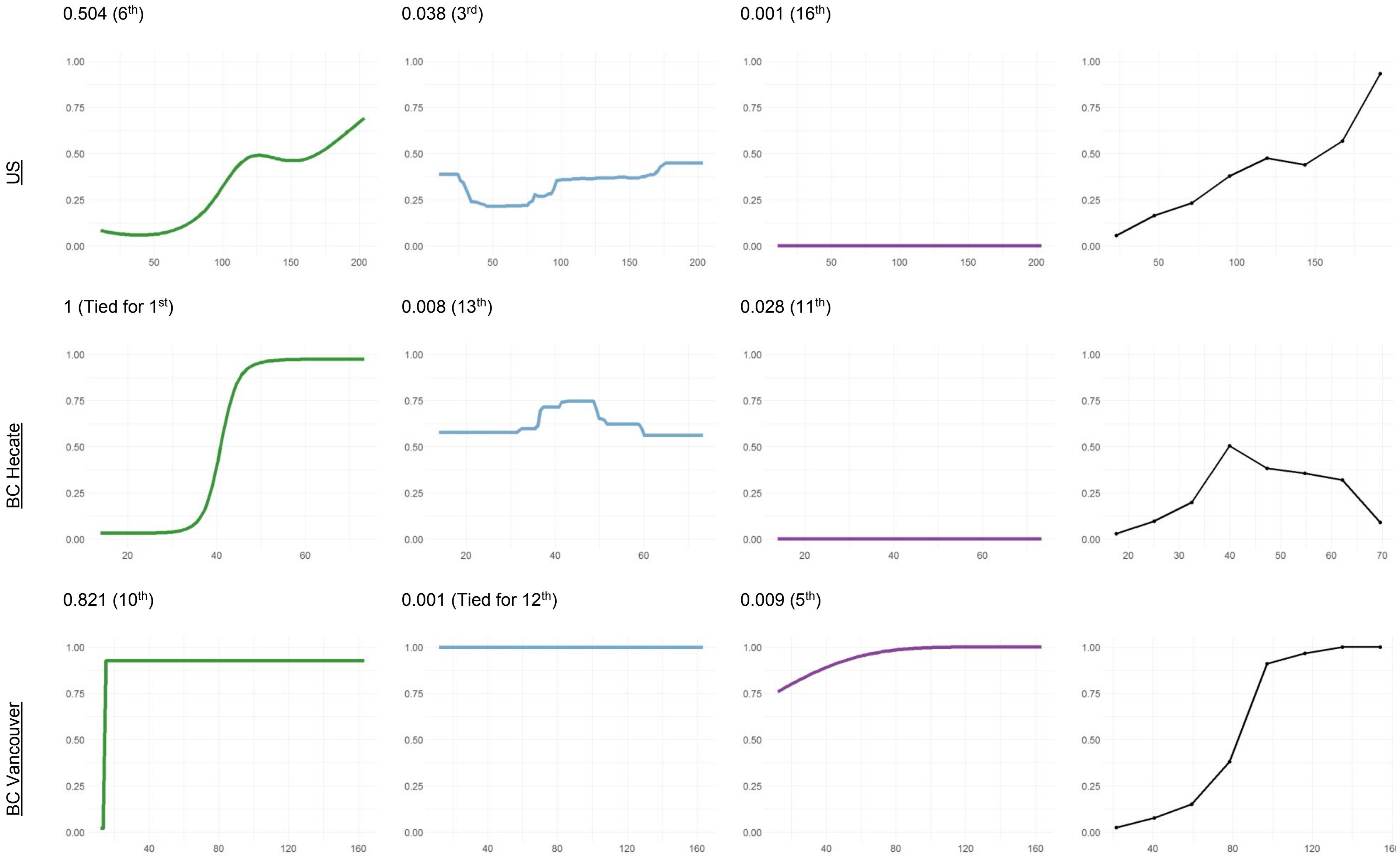
0.002 (Tied for 16th)

0.071 (7th)

0.643 (4th)

0.052 (4th)

0.003 (Tied for 12th)



0.504 (6th)

0.038 (3rd)

0.001 (16th)

1 (Tied for 1st)

0.008 (13th)

0.028 (11th)

0.821 (10th)

0.001 (Tied for 12th)

0.009 (5th)

2.4.3 Partial Dependence Plots

Partial dependence plots generated for alkalinity, oxygen, phosphate and silicate can be seen in Figures 8-11. Table 5 presents depth as a frequently important variable in this analysis of glass sponges. Partial dependence plots from multiple models and areas show that as the taxa encounters depths shallower than 1000 meters, the probability of presence decreases, confirming they are more likely to be found in deep waters (Fig. 12). Figure 13 shows the taxa data in the BC Vancouver sub area, it can be easily noted here that the majority of the presence values are in the deeper waters. It is important to note that glass sponges also exist in shallow waters, as evidenced by the glass sponge reefs of coastal British Columbia (Fig. 15).

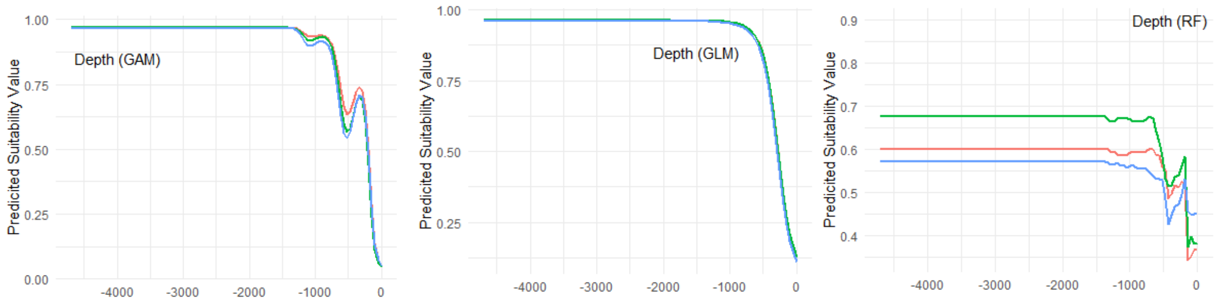


Figure 12. Partial dependence plots for depth from GAM, GLM, and random forest (RF) models. Each plot has three lines for each time the evaluation was run.

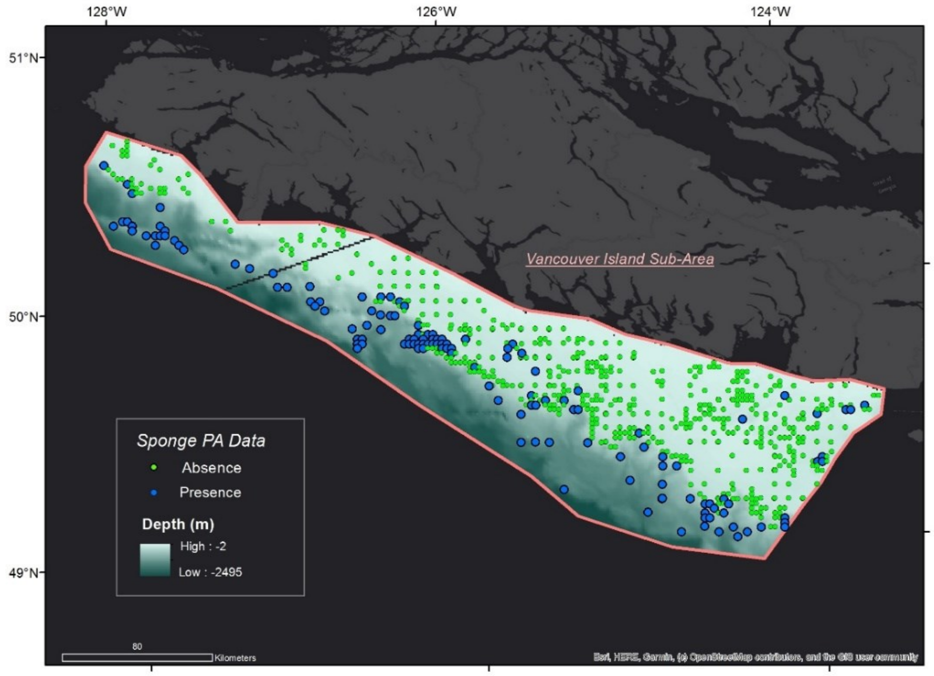

128°W
126°W
124°W
51°N
50°N
49°N
Vancouver Island Sub-Area
Sponge PA Data
Absence
Presence
Depth (m)
High : -2
Low : -2495
80
Kilometers
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Figure 13. Bathymetry and species data points within BC Hecate Strait sub-area (grey line is missing data).

2.4.3.1 Alkalinity

Figure 8 shows partial dependence plots and ranked variable importance for alkalinity, for each of the 18 model-area combinations. Within the North Pacific area, all three models suggest a high probability of glass sponge presence within highly alkaline waters. In the British Columbia area, the GAM model suggests an increase in glass sponge presence probability with alkalinity values higher than 2.2 μmol l^{-1}, a trend that is also present in each of the sub-areas. The BC Hecate area, in particular, suggests an increased probability of presence for glass sponges in alkalinity values of 2.2 μmol l^{-1} and higher. All three models produced comparable partial dependence plots for this area, where variable importance values for alkalinity rank 1st out of 19 variables for all three models. Finally, the data distribution plot for alkalinity values in the BC Hecate area shows a high percentage of presence values in alkalinity ranges of between 2.15 and 2.25 μmol l^{-1}. The BC Vancouver area has a comparable GAM response curve as BC Hecate, however the BRT and MaxEnt models for BC Vancouver have lower variable importance ranks and do not show an increase in probability of presence with increased alkalinity values. The BC Hecate Strait area has a range of alkalinity values of 1.855 – 2.308 μmol l^{-1}, while most other areas have a maximum of closer to 2.4 or 2.5 μmol l^{-1}.

2.4.3.2 Oxygen

Partial dependence plots for oxygen (Fig. 9) are interestingly varied as well; the general trend seen in GAM models from several of the areas suggests an increased probability of presence with lower oxygen values, except that the GAM model for the BC area suggests the opposite. Oxygen was 10th in variable importance in the GAM model for the BC area, which means there is less indication in the model that glass sponges are strongly influenced by oxygen. The GAM plot for the BC area is interesting because it presents a pattern opposite to the oxygen plots for other areas, opposite to the plots for other models, and opposite to the data distribution itself. The variable importance values for oxygen in these models are not as consistently high as for alkalinity. The data distribution plots for all of the areas show a greater proportion of presence values in area with relatively low oxygen concentrations.

2.4.3.3 Phosphate

Response curves for phosphate can be seen in Figure 10, and present a wide variety of possible responses of probability of sponge presence in relation to phosphate content. Due to the lower variable importance values, it is more difficult to find strong environmental trends in the data. Phosphate was within the top three variables influencing the GAM, BRT and MaxEnt models within the BC area. For the remaining areas, the variable importances range from 4th to last (19th). The plots from the BC area indicate that probability of sponge presence increases with phosphate levels of roughly 3 μmol l^{-1} and higher. The next highest variable importance values are a result of the BRT and MaxEnt models for the BC Hecate area. Phosphate was 4th in variable importance for these two models and both indicate a slight increase in probability of presence between 2.0 and 2.5 μmol l^{-1}.

2.4.3.4 Silicate

Models which indicated that higher silicate content is more suitable for sponges included MaxEnt (North Pacific, BC Hecate and BC Vancouver), and GAM (Alaska, US, BC Hecate and BC Vancouver) (Fig. 11). The data distribution plots largely indicate a greater proportion of presence values with increasing silicate value. The two partial dependence plots with the highest variable importance values (3rd) area the MaxEnt model in the North Pacific area and the BRT model in the US area. While both these plots indicate a general increase in probability of presence in relation to an increase in silicate levels, the BRT model in the US also indicates an increased probability of presence with very low silicate contents. Generally, because silica plots have lower variable importance values than alkalinity or oxygen plots, less weight can be placed on their accuracy. The silicate GAM plot for the BC area produced an opposite result to the remaining plots.

2.4.3.5 Spatial Predictions

Figure 14 shows the predicted probability of glass sponge presence from the BRT, GAM and MaxEnt models for Alaska, in the form of a raster prediction probability of presence.

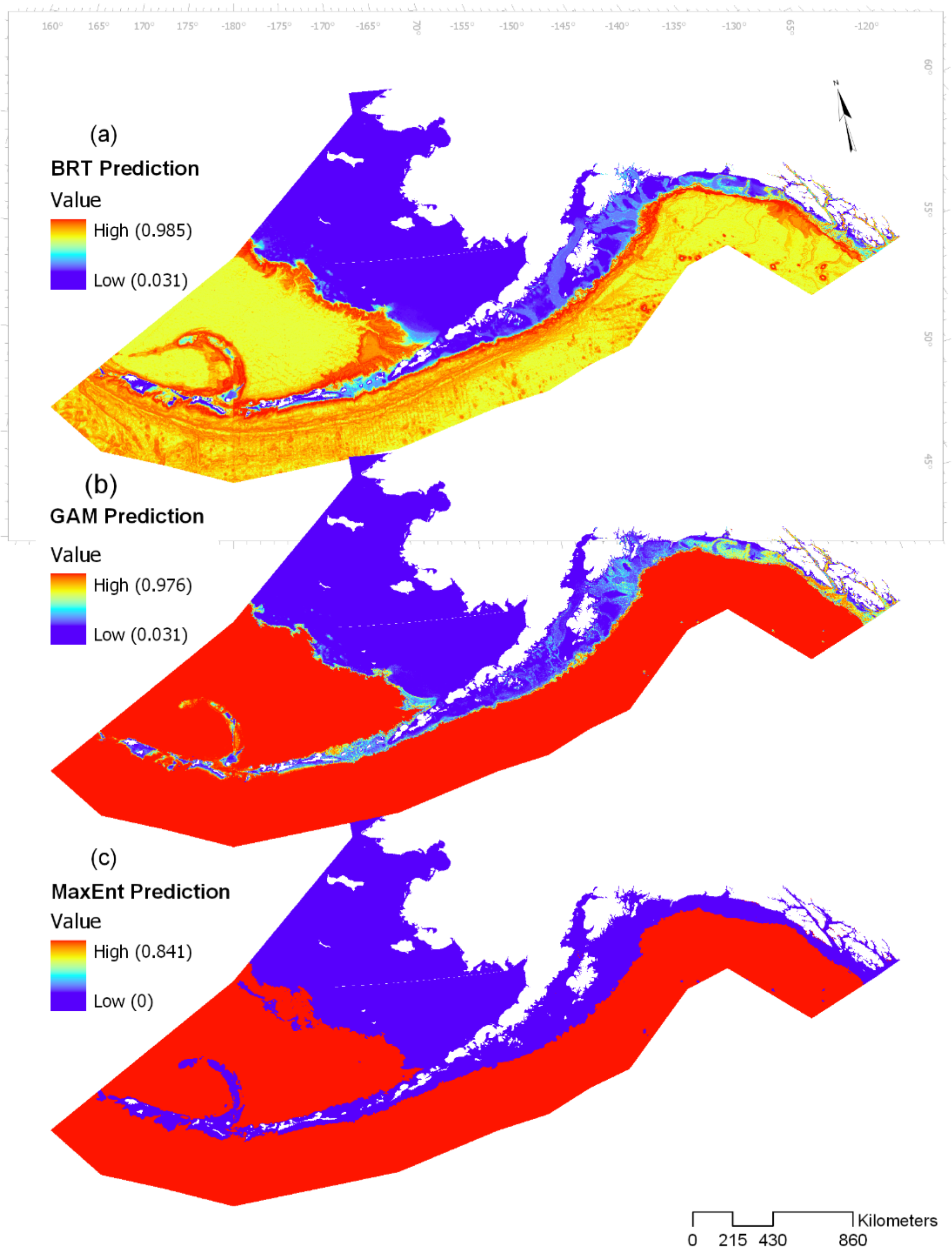


Figure 14. Model predictions from BRT, GAM, and MaxEnt models for the Alaska sub-area.

Alaska was selected for this section because it is a smaller area than the North Pacific (which is a large area, thus making predictions difficult to see in detail) but larger than the BC and US areas (which are quite small and have less variation in predictions of probability of presence). The AUC values for these three models respectively are 0.816, 0.804, and 0.428, meaning the BRT and GAM models performed very well and the MaxEnt model performed poorly. The MaxEnt model for the Alaska area is the model that performed least well across all models and areas, and as can be seen in Figure 14c, the area is divided into red (high probability of presence) and blue (low probability of presence) without much variation between those two predictions. Figure 14a and b show much more variation of probability of presence across the prediction.

2.4.4 Uncertainty Metrics: BC Hecate Strait

Uncertainty refers to a lack of sureness or confidence about something (Elith et al., 2002b). Most outputs of SDM work are presented with confidence, with no indication of uncertainties, but it has been proposed that maps of uncertainty would help in the interpretation of these predictions (Elith et al., 2002b). The Hecate Strait sub-area in BC was used for the uncertainty metric analysis because of its high environmental variation and interesting patterns of alkalinity, oxygen and silicate distributions. The prediction from a GLM run on this area can be seen in Figure 15, along with the outlines of the Hecate Strait/Queen Charlotte Sound Glass Sponge Reefs Marine Protected Area (MPA).

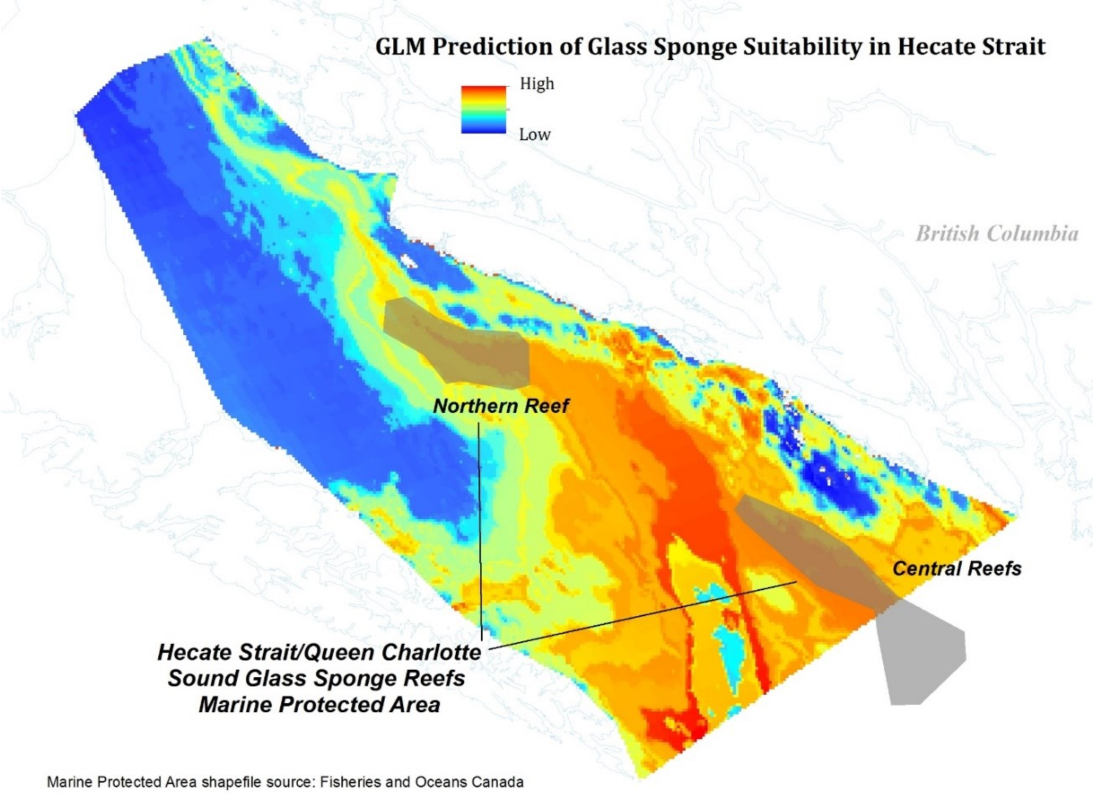

GLM Prediction of Glass Sponge Suitability in Hecate Strait
High
Low
British Columbia
Northern Reef
Central Reefs
Hecate Strait/Queen Charlotte Sound Glass Sponge Reefs Marine Protected Area
Marine Protected Area shapefile source: Fisheries and Oceans Canada

Figure 15. GLM prediction of glass sponge probability of presence in Hecate Strait with MPA boundaries.

The Northern Reef and part of the Central Reef are contained within the Hecate Strait boundaries employed for this study. The MPA boundaries overlap with moderately high suitability for glass sponges; providing some confidence in the model predictions and their real-world accuracy, despite the MPA area boundaries not falling within the highest probability of presence areas (red areas).

2.4.4.1 Standard Error of GLM Prediction

Figure 16 shows the partial dependence plots for alkalinity, oxygen and silicate for the GLM model of the BC Hecate Strait area, with model-based confidence intervals added. Generally the confidence intervals are narrow, corresponding to a low expected error, for predictor value ranges with many data points, shown in the figure as a high density of red/blue lines. Value ranges with wider intervals have fewer data points. If the areas with wide confidence intervals overlap spatially, the relevant areas are likely to produce less certain predictions.

The standard error of the GLM fit was written to a raster and can be seen in Figure 17a. Area 2 in Figure 17a has high uncertainty.

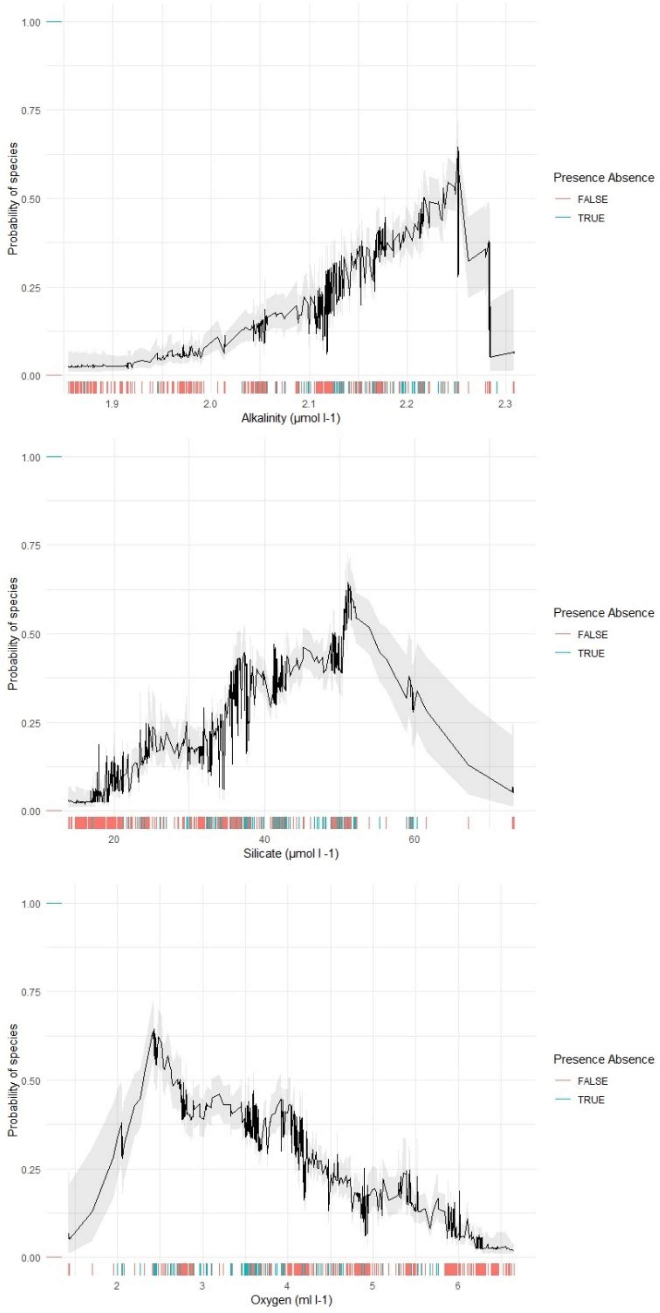


Figure 16. Partial dependence plots for alkalinity, silicate, and oxygen with estimated confidence intervals.

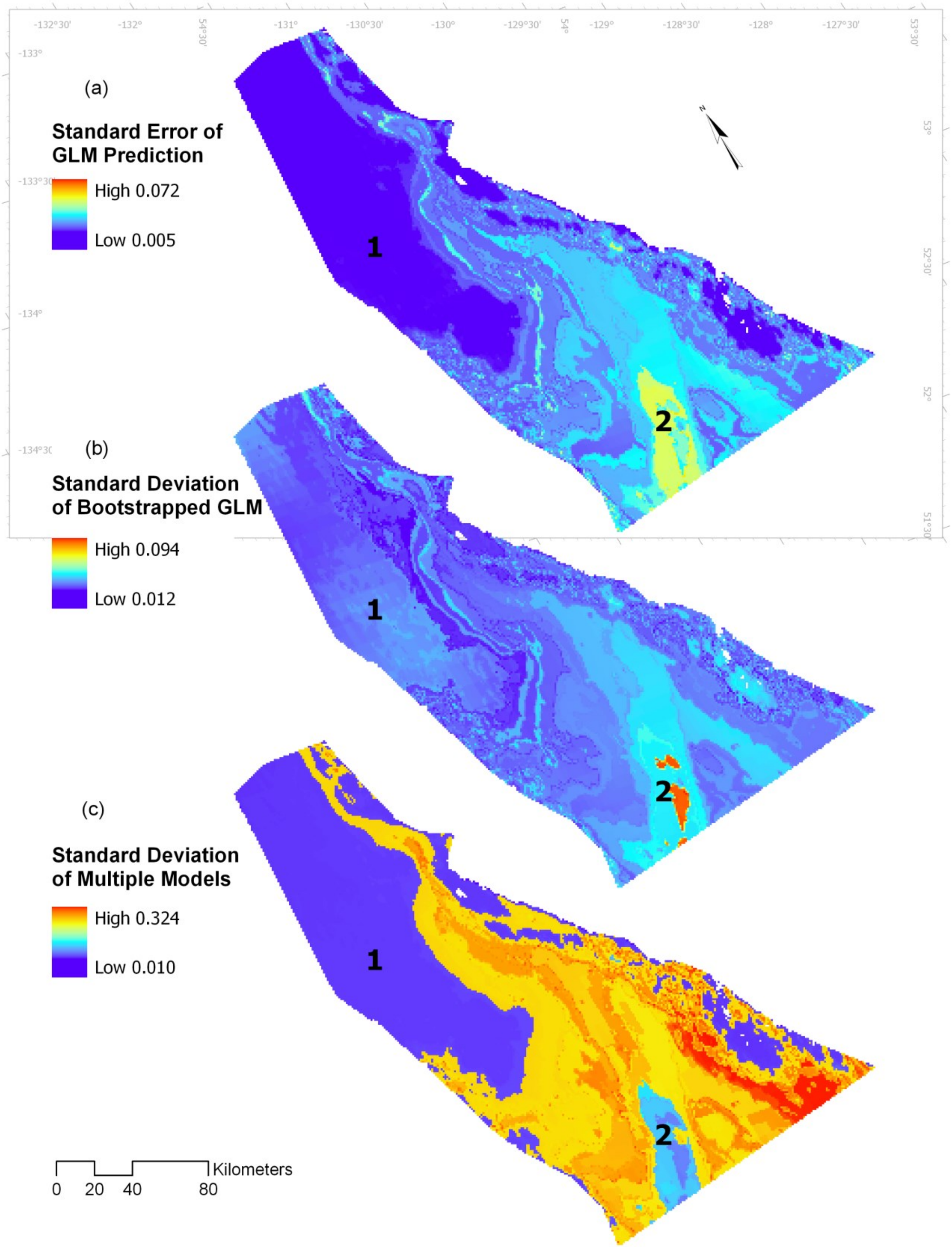


Figure 17. Uncertainty metrics mapped to Hecate Strait: a) standard error of GLM prediction, b) standard deviation of bootstrapped GLM, and c) standard deviation of multiple models.

There are very clear environmental gradients throughout Hecate Strait which can be visually confirmed to have an influence on the uncertainty metrics. Area 1 in Figure 17a has low alkalinity and silicate levels with high oxygen levels (Figs. 19-21). This combination of environmental values generally coincides with absence data for glass sponges, which the models interpret as unsuitable habitat. The standard error in Fig. 17a is low, indicating high certainty in the prediction of low probability of sponge presence. Area 2 in Figure 17a has high uncertainty values. Area 2 corresponds with opposite niche environmental characteristics to Area 1; very high alkalinity levels, very high silica levels, and very low oxygen levels, all of which are value ranges that are poorly represented in the data. As can be seen in the partial dependence plots (Fig. 16), these value ranges are associated with low data density and high CIs. The GLM is forced to make predictions for these areas based on a combination of few data points with similar values and extrapolation from more data-dense value ranges, which leads to extreme predictions and higher uncertainty.

2.4.4.2 Standard Deviation of Bootstrapped GLMs

Figure 17b presents the result of bootstrapping the binomial GLM 200 times, calculating the standard deviation for each cell and then writing this result to a raster. This method of spatially showing prediction uncertainty yields similar results to the initial method of calculating the standard error from the GLM. It shares an area of high uncertainty with the first method (Area 2), which was noted as having extreme values of all three input predictors. This method, as well as the first method, does not show Area 1 to have high standard error, indicating consistency across methods.

2.4.4.3 Standard Deviation of Multiple SDMs

Finally, Figure 17c presents the result of running multiple SDMs and mapping the standard deviation of the model predictions. The models used and their individual predictions can be seen in Figure 18. This method shows the highest uncertainty to exist in the lower right corner of Hecate Strait, which corresponds to medium uncertainty in the first two methods. While taking a different approach from the first two methods, this final method is equally as important for determining spatial uncertainty from predictive models and yields interesting results that could aid policy makers in making informed decisions based on SDMs. Area 1 in Figure 17c has low standard deviation because the majority of the SDMs produced a similar probability of presence for this area (Fig. 18). This is a consistent result from all three methods, indicating low uncertainty in environments considered unsuitable for the species in question. The areas that are yellow and orange yield higher error values because the models produced different results, despite having been calibrated with the same data. Each of the specific model outputs showed the highest probability of species to be in the area which has high standard deviation in Figure 17c. The resultant high standard deviation is a result of this being the area of the model output that changes the most with each different SDM.

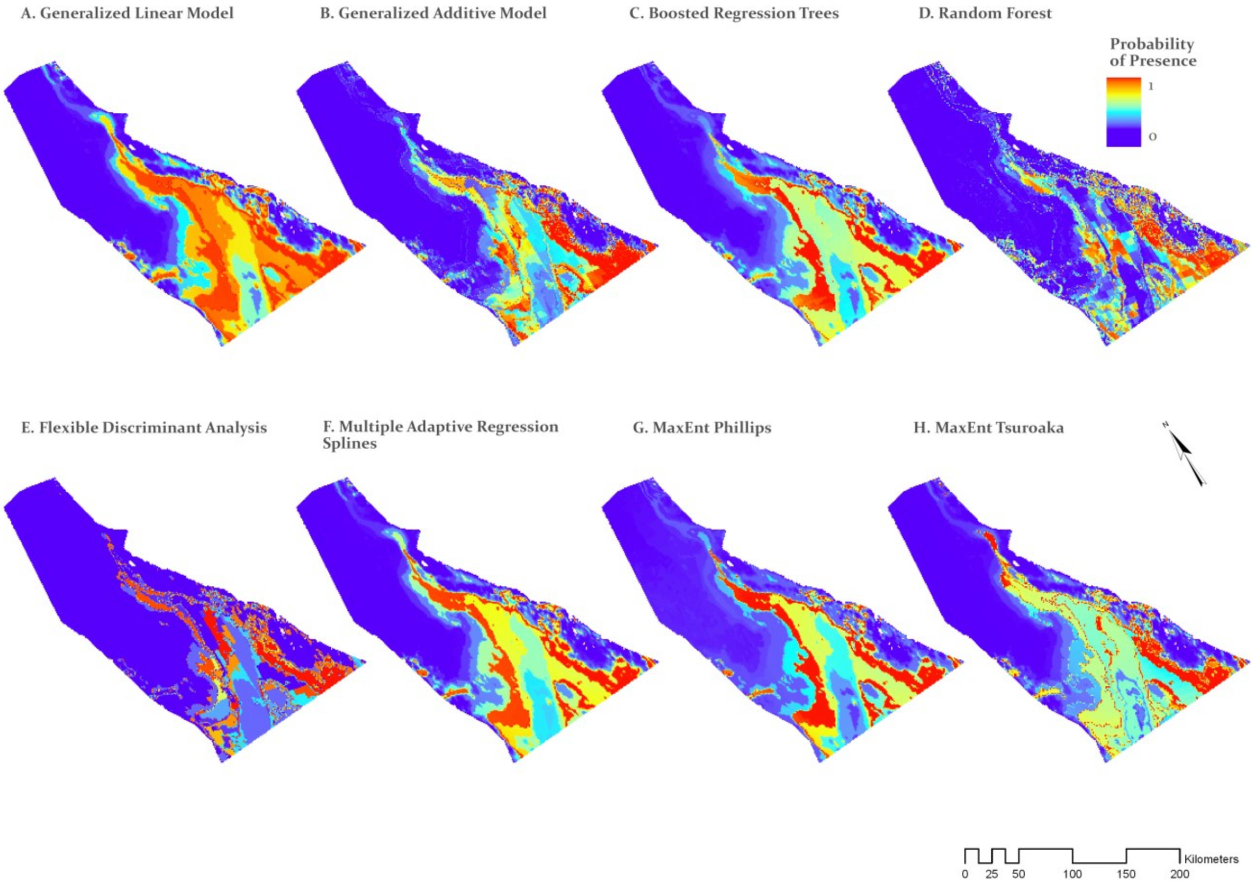

A. Generalized Linear Model
B. Generalized Additive Model
C. Boosted Regression Trees
D. Random Forest
Probability of Presence
1
0
E. Flexible Discriminant Analysis
F. Multiple Adaptive Regression Splines
G. MaxEnt Phillips
H. MaxEnt Tsuroaka
N
0 25 50 100 150 200
Kilometers

Figure 18. Predictions of glass sponge probability of presence in Hecate Strait: a) GLM, b) GAM, c) BRT, d) RF, e) FDA, f) MARS, g) MaxEnt Phillips, and h) MaxEnt Tsuroaka.

2.5 Discussion

2.5.1 Aspatial Model Predictions: Partial Dependence Plots and Variable Importance

As a method of assessing the validity of models, partial dependence plots and variable importance rankings were presented and analyzed. The results presenting partial dependence plots and variable importance values from the GAM, BRT and MaxEnt models can provide information about 1) the ability of the model to describe the environment-species relationships, and therefore 2) the potential for using the model to make inferences about the ecology of glass sponges and characteristics of their habitats.

In addition to standard model performance metrics such as AUC, TSS and Kappa, ranked variable importance and partial dependence plots can inform how certain or uncertain an SDM-based prediction is. If a multi-model and multi-area approach has been used for the modeling, and a given environmental variable has high importance values across multiple model types and areas, a higher confidence can be placed in that variable having a non-spurious effect on the distribution of the response variable, e.g. glass sponges in the present case. It is likely that the use of more models, and more environmentally distinct areas would serve to strengthen the multi-model, multi-area approach even more. The expected result from performing a multi-area, multi-model approach was that strong habitat preferences will be reflected similarly in partial dependence plots from different models and areas, while weak habitat preferences will not.

Using this approach, the partial dependence plots for alkalinity strongly suggest a causal relationship between alkalinity and glass sponge presence. According to the model results, glass sponge probability of presence increases in conjunction with higher alkalinity values, specifically at concentrations of 2.1 μmol l^{-1} or 2.2 μmol l^{-1} and higher. While several of the areas modeled produced this trend, the BC Hecate area displays arguably the most consistent result based on the fact that all of the three models for this area ranked alkalinity as first out of nineteen other variables. Less confidence can be placed in certain areas and models where alkalinity is ranked lower and the associated response curves are inconsistent with those of high variable importance. For example, MaxEnt models in BC and US areas have alkalinity importance rankings of 14th and 15th respectively, and neither of the corresponding partial dependence plots provide any useful ecological information about how sponges respond to alkalinity levels. Although the models are too complex to provide a definitive explanation, it is likely that the BC and US areas have other environmental variables which are more influential for the model, and therefore the relationship between sponge presence and alkalinity is more difficult for the model to identify. Every highly ranked (top 25 percentile) partial dependence plot of alkalinity shows an increased probability of presence associated with high alkalinity values, generally above 2.2 µmol l^{-1}. Areas of Hecate Strait which meet these alkalinity values are relatively few and can be seen in Figure 19.

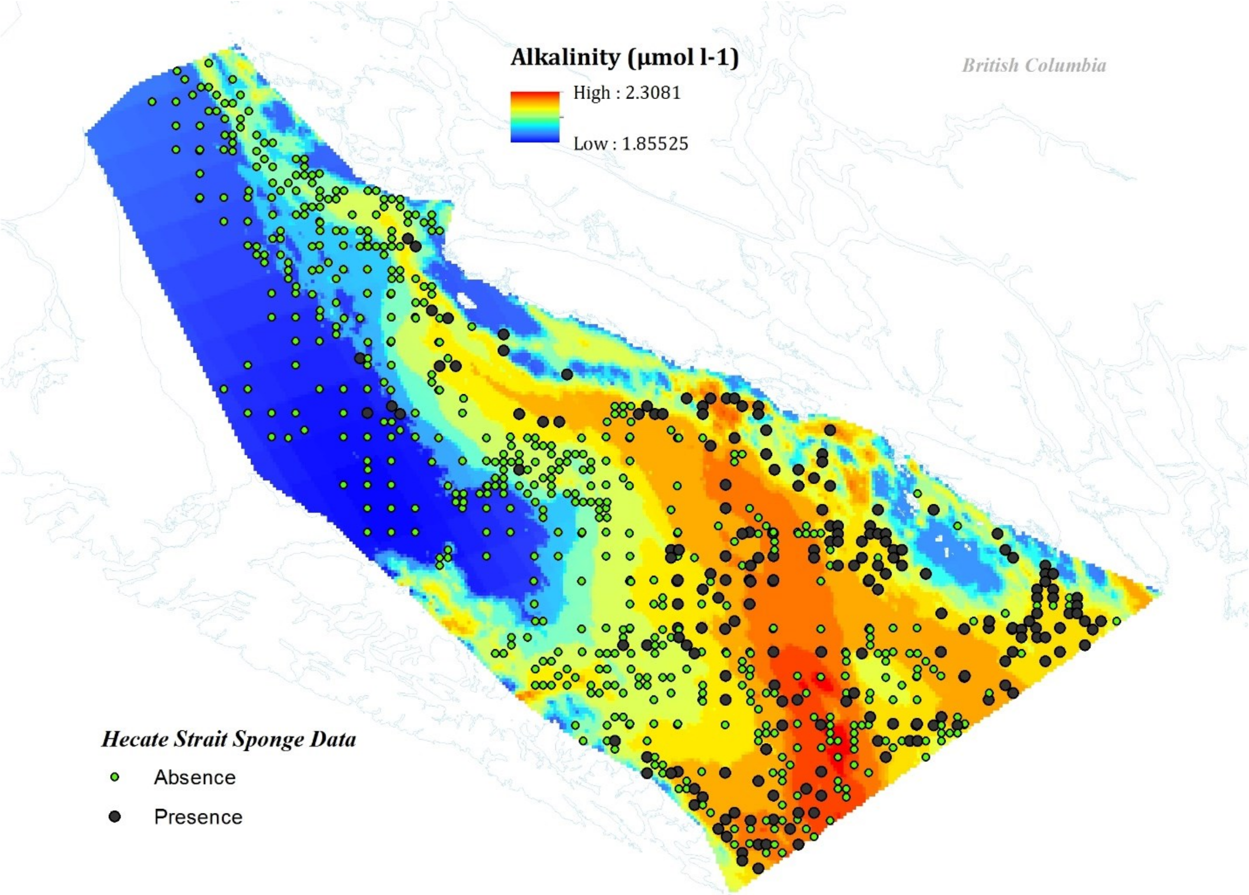


Figure 19. Alkalinity values across Hecate Strait.

According to the partial dependence plots with high variable importance rankings, a high probability of glass sponge presence is associated with low oxygen values, generally below 2 ml l^{-1}. Deep-water sponges, and many other invertebrates, use little oxygen and have adapted to live in low-oxygen environments, for example during low tide or in benthic sediments (Leys et al., 2018). Leys and Kahn (2018) note that glass sponges tolerate long-term hypoxic conditions by reducing their filtration rate and feeding activity. Filtration, they concluded, is costly to glass sponges and attempting to slow their filtration has driven innovations in their morphology and physiology (Leys et al., 2018). Chu et al. (2019) also found that dissolved oxygen was a highly ranked positive predictor of habitat for cold-water coral and sponge grounds in the Canadian northeast Pacific Ocean. As a result of this finding, Chu et al. (2019) predicted that cold-water coral and sponge taxa would have lower oxygen requirements in comparison to highly mobile taxa such as fish. It was also found, in an attempt to validate the models predicting that cold-water corals and sponges are likely to occur in severely low oxygen environments, that these taxa exist in oxygen levels as low as 0.2 ml l^{-1} at the Union and Dellwood seamounts (both are southwest of the southern point of Haida-Gwaii). Figure 20 shows the distribution of oxygen content across Hecate Strait. It is important to note how related oxygen and alkalinity are to each other in this area. With many sponges existing in high-alkalinity, low-oxygen waters, it is difficult to know whether this is because of the high alkalinity, the low oxygen, some combination of the two, or a third variable that is also correlated with both alkalinity and oxygen. This illustrates the benefit of using a multi-area approach, because these two variables may be less related in other areas which have also been modeled.

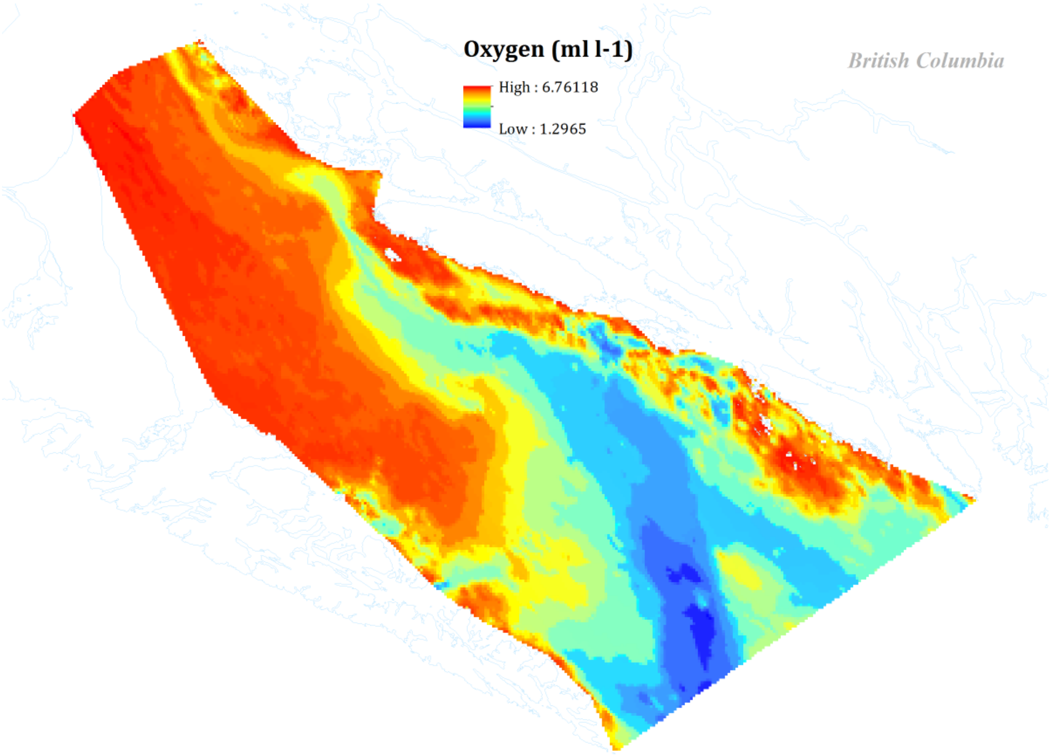


Figure 20. Oxygen values across Hecate Strait.

Silicate was not often returned from the models as one of the top 25% of variables, however certain models and areas did produce silicate as the most important variable. These included MaxEnt for the North Pacific area, GAM and BRT for the Alaska area, BRT for the US area, and GAM for the BC Hecate area. Out of these, most show an increased probability of glass sponge presence with high levels of silicate. High levels of silicate often overlap with high levels of alkalinity within Hecate Strait (compare Fig. 19 and Fig. 21).

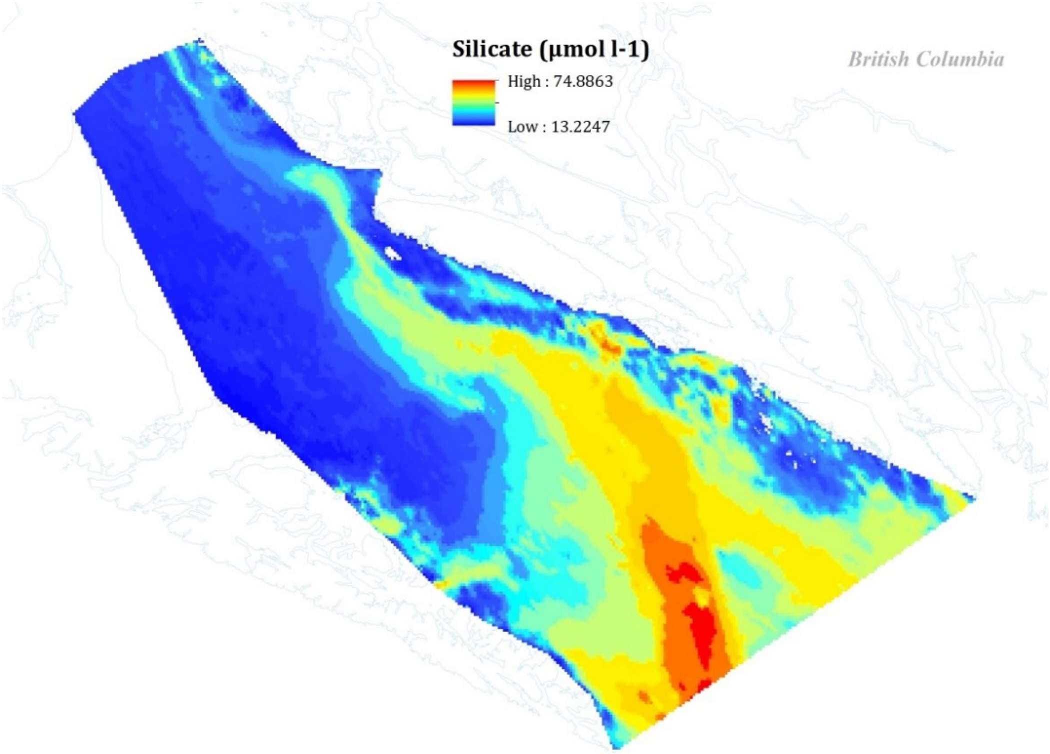


Figure 21. Silicate values across Hecate Strait.

In published literature, it has been indicated that glass sponges need high levels of dissolved silica (Leys et al., 2004; Austin., 1984; Chu et al., 2019). Chu et al. (2019) found that silicic acid was a top predictor for sponge groups because biogenic silica (biogenic silica occurs when dissolved silicate transforms to particulate skeletal matter (Treguer et al., 1995)) can constitute over 90% of the biomass of cold-water sponges. Silicate levels are high in both the Antarctic Ocean as well as the coastal northeast Pacific Ocean, which are both regions of high glass sponge abundance (Leys et al., 2004; Treguer et al., 1995). Interestingly, the Hecate Strait area shows an increased probability of glass sponge presence at much lower levels of silicate than the remaining areas tested within this thesis. Many areas indicate high probability of glass sponge presence in areas with silicate values of 150 μmol l^{-1}, however Hecate Strait indicates high probability of presence beginning where silicate values reach over 40 μmol l^{-1}. The highest level of silicate within the Hecate Strait subarea is 74 μmol l^{-1}. Whitney et al. (2004) identified silicate levels of over 40 μmol l^{-1} around sponge reefs in Hecate Strait, thus confirming this result.

Evaluating partial dependence plots in addition to model accuracy metrics as an additional way to assess SDM outputs is a descriptive and largely qualitative exercise. The challenge is that there can be valuable information concerning ecological relationships, but also nonsensical and spurious relationships, presented in these plots. But there is value in producing multiple models for multiple areas because oceanic environments subject to different currents and water masses and different levels of terrestrial influence can vary drastically in their biogeochemistry. This thesis presented one approach to disentangling the two; by looking for species-environment relationships that are strong (as indicated by high variable importance), consistent between model types and consistent between different areas, it is possible to extract only those relationships most likely to be caused by ecological processes. The two anomalous plots mentioned in the results section, the oxygen and silicate plots from the GAM model for the BC area, presented the opposite relationships of what the remaining models presented. These two plots are an excellent example of why it is important to not make inferences about ecological relationships based on single-model and single-area partial dependencies.

Using ranked variable importance values and selecting a threshold for a confidence cut-off can provide a quantitative measure of accuracy. Providing a measure of probable accuracy alongside model outputs can be helpful for environmental managers and stakeholders who require numerical models to estimate species distribution to design effective spatial management measures for conservation and protection.

2.5.2 Spatial Model Uncertainty Predictions

Spatially examining the uncertainty of model predictions is not commonly done in SDM studies, however it is important that potential users of SDM products have an understanding of the predictive accuracy of models and how this may vary across geographic space (Elith et al., 2005). Most evaluation metrics of predictive performance use a comparison of predictions against observations at a particular set of sites (Fielding et al., 1997). As also done in this thesis, statistics such as kappa and AUC values are widely used to assess whether predictions are suitably accurate for their intended use, however these statistics are somewhat restricted because they do not assess the predictions in geographic space and do not allow for exploration of spatial errors (Elith et al., 2005; Elith et al., 2002a; Fielding et al., 1997).

2.5.2.1 Confidence Intervals and Standard Error

It is suggested in SDM literature that plotting confidence intervals (CI) around model predictions could be crucial to the interpretation of the models' performance, particularly mapping confidence intervals of these predictions (Elith et al., 2002b; Elith et al., 2005). CIs around plotted responses (such as partial dependence plots) help show where species-predictor variable relationships are most uncertain (Ferrier et al., 2002). Adding these error metrics is instrumental for producing models which can be understood as ecological realities (Elith et al., 2005). Figure 16 shows CIs added to partial dependence plots of alkalinity, oxygen and silicate. The largest CI on each plot correlates with the lowest density of data points in each variable. Sponge data where silicate values are greater than 55 µmol l^{-1} have the highest uncertainty, as there are fewer data points for silicate values over this threshold. As mentioned earlier, silicate levels in Hecate Strait have been documented to be lower than surrounding areas, yet sponges remain in great abundance in Hecate Strait.

According to Figure 15, the highest probability of sponge presence within Hecate Strait coincides with areas comprised of high alkalinity, very low oxygen, and medium-high silicate levels. Hecate Strait is a shallow asymmetric channel between Haida Gwaii and the northern mainland of British Columbia (Perry et al., 1994). It is a unique area due to its shape; it is roughly 140km wide at its southern end and narrows to 48km in the north, covering around 23,000km^2 with depth values reaching down to 494m. The shallowest part is the northwest area, which has low alkalinity levels, high oxygen levels and low silicate levels. This is also the area for which the lowest probability of presence for sponges was predicted (Fig. 15). This NW area (labelled as Area 1 in Fig. 17) shows consistently low uncertainty with all three methods. All models used predicted low probability of sponge presence in this area, due to its physical characteristics mentioned previously, and all methods of quantifying uncertainty show low uncertainty in this area, indicating it is highly probably this area is unsuitable for sponges.

Using confidence intervals as the only means of quantifying the uncertainty of SDM predictions is not a complete method, according to Elith et al. (2002b), who mention that uncertainty in model outputs is not explicitly accounted for in the CIs of GLMs. They suggest that bootstrapped CIs can better account for different sources of uncertainty rather than simply applying CIs to GLM predictions. This is an interesting avenue for further work on spatially quantifying model uncertainty, as only the SE of the GLM fit and SD of the bootstrapped GLMs were calculated in the work that underlies this thesis. Area 2 was identified as an area of highly uncertain predictions by both the SE of the GLM fit and the SD of the bootstrapped GLMs. The bootstrapped runs of the GLM produce very consistent predictions in the northwest part of Hecate Strait (Area 1). The third method of measuring uncertainty involved running eight SDMs on the same data used for the prior methods and then calculating SE of all eight predictions. Figure 17c presents high SE values around the southeast corner of Hecate Strait, this area has medium uncertainty in the first two methods, indicating slight differences in model predictions. The eight models produced consistent predictions for the NW area of Hecate Strait, suggesting with a low level of uncertainty that sponge probability of presence in this area is low.

One method of comparing these three metrics of estimating uncertainty is by looking at the original biological input data. Area 1 mostly contains absence values and almost no presence values (Fig. 19), therefore it seems that when every model shows low probability of presence, one can assume with some confidence it is likely correct. Additionally, Area 1 is shallow, has high levels of oxygen and low levels of both silicate and alkalinity, which are environmental conditions that are the opposite of what generally models consider suitable habitat for sponges. Figure 17a, b and c present Area 1 as having low uncertainty, indicating all assign low uncertainty to the prediction of low probability of presence in this area.

These results allow for the conclusion that if the model predicts low probability of presence, it has higher certainty in this prediction than in predictions of high probability of presence. The first two methods differ from the last method in what they show, but the first two methods show medium uncertainty in the same areas which are highly uncertain in the last method (the areas where the models all predict relatively high probability of presence, but of varying values and slightly different geographic spreads of this high probability of presence). This leads to the conclusion that uncertainty is generally lowest where the models predict the species not to be, and highest where the models predict the highest probability of species presence to be.

Both bootstrapping a model and running multiple SDMs are useful methods of calculating prediction uncertainty, and both these methods could be extremely useful for providing planners with information to consider when employing the predictions in conservation planning and decision making (Ferrier et al., 2002).

2.5.3 Limitations

Numerous limitations exist within SDM work; not all models are transposable to distinct environments, they are strongly dependent on the considered scale, they are difficult to implement in a management context, many models are not easily interpretable, and software is not always available to practitioners (Guisan et al., 2005). Alongside all of these limitations, a consistent limitation is that fact that any model will rely heavily on the quality of the input data. This study used a spatial data thinning method based on local environmental variation to eliminate the spatial sampling bias that was present in the original data set. Spatial bias is a common limitation in the SDM field because it may cause biased model results and it is difficult to tell if the species-environment relationships in the model are representative of the real world or if they are a function of how the data was sampled. When using biological data from another organization, such is the case here, it can be challenging to find sufficient details about the data to ensure its quality. There are also limitations in the interpretation of regression based models and machine learning models. The two methods produce different results. For example calculating uncertainty metrics is more easily done from regression models as opposed to machine learning models. This thesis compared a regression based model with two machine learning models. A final limitation often overlooked in SDM studies is the spatial dependency of accuracy of the model outputs. Presenting the error spatially is an important aspect of SDM moving forward, as it will be easily understood by those in environmental management who are unfamiliar with the modeling methods. Spatially presented error metrics add value to the already used aspatial error metrics.

Chapter 3. Conclusion

By analyzing a variety of commonly used SDMs and examining different spatial and aspatial metrics to quantify model accuracy and uncertainty, this thesis has shown how applying a multi-model and multi-area approach can improve the interpretation of the modeled species-environment relationships. It has also shown how different methods of uncertainty mapping can provide increased insight as to which areas are predicted by the model to have high/low levels of uncertainty.

Running three models on six areas showed that partial dependence plots can differ substantially between model types and adjacent geographical areas. It is therefore necessary to not overstate the ecological results presented in individual plots, and to be careful while interpreting them ecologically. One way to assess the ecological interpretability of partial dependence plots is to perform a multi-model, multi-area study, and compare plots across models and areas prior to drawing ecological inferences.

Based on the results presented in this thesis, it appears that glass sponges are most likely to be found in areas with alkalinity values greater than 2.2 μmol l^{-1} and oxygen values lower than 2 ml l^{-1}. While silicate was also an important environmental predictor, the results for the probability of sponge presence in relation to silicate are more variable. Every area except Hecate Strait indicated that glass sponges are more likely to exist in areas with silicate values of 150 μmol l^{-1} and over, however lower values in Hecate Strait confirm sponges can exist in areas with silicate values of 40 μmol l^{-1} and over.

While model accuracy metrics like AUC and TSS contain important information about the ability of a model to produce good predictions, spatial uncertainty metrics can outline areas where predictions are more or less likely to be correct. There is a small area in the south of Hecate Strait (Area 2) that both the GLM and the bootstrapped GLM indicate as being subject to highly uncertain predictions. Such areas should be treated cautiously regardless of the overall accuracy of the model as indicated by the accuracy metrics, and such areas could be targeted for future data collection.

Finally, it was shown that different approaches to estimating prediction uncertainty can yield different results but still important results. This can be seen in Hecate Strait: predictions for the shallow, low-alkalinity area in the NW part of Hecate Strait (Area 1) are consistent between models (all models give low probability of sponge presence), consistent between bootstrapped runs of the GLM models and the SE of the GLM model (which also all predict low probability of sponge presence), and are also consistent with the data points from that area. This indicates we are very confident in the model prediction of low probability of sponge presence in Area 1. The highest uncertainty corresponds to areas where models have presented high probability of presence, since these areas do not all overlap neatly, the uncertainty arises from these varying predictions.

Chapter 4. References

Aarts, G., J. Fieberg, and J. Matthiopoulos. (2012). Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution, 3*(1), 177-187.

Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology, 43*, 1223-1232.

Araújo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology, 11*(9), 1504-1513.

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution, 22*(1), 42-47.

Atwater, D., & Fautin, D. G. (2001). "Hexactinellida". University of Michigan, Museum of Zoology. On-line resource: https://animaldiversity.org/accounts/Hexactinellida/

Austin, W. C. (1984). Underwater birdwatching. *Can Tech Rep Hydrogr Ocean Sci 38*, 83–89.

Austin, W. C. (1999). The relationship of silicate levels to the shallow water distribution of hexactinellids in British Columbia. *MEMOIRS-QUEENSLAND MUSEUM*, 44, 44-49.

Beazley, L. I., Kenchington, E. L., Murillo, F. J., & Sacau, M. D. M. (2013). Deep-sea sponge grounds enhance diversity and abundance of epibenthic megafauna in the Northwest Atlantic. *ICES Journal of Marine Science, 70*(7), 1471–1490.

Becker, J. J., Sandwell, D. T., Smith, W. H. F., Braud, J., Binder, B., Depner, J., Fabre, D., Factor, J., Ingalls, S., Kim, S. H., Ladner, R., Marks, K., Nelson, S., Pharaoh, A., Trimmer, R., von. Rosenberg, J., Wallace, G., & Weatherall, P. (2009). Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS. *Marine Geodesy, 32*(4), 355–371.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). Bioclim: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions, 20*(1), 1-9.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Cañadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E., & Hammond, P. S. (2005). Habitat preference modelling as a conservation tool: Proposals for marine protected areas for cetaceans in southern Spanish waters. *Aquatic Conservation: Marine and Freshwater Ecosystems, 15*(5), 495–521.

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling, 210*(4), 478-486.

Chu, J. W. F., Nephin, J., Georgian, S., Knudby, A., Rooper, C., & Gale, K. S. P. (2019). Modelling the environmental niche space and distributions of cold-water corals and sponges in the Canadian northeast Pacific Ocean. *Deep-Sea Research Part I: Oceanographic Research Papers, 151*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Conway, K. W., Barrie, J. V., & Krautter, M. (2005). Geomorphology of unique reefs on the western Canadian shelf: sponge reefs mapped by multibeam bathymetry. *Geo-Marine Letters, 25*(4), 205-213.

Dempsey, S.J., E.M. Gese, G.M. Kluever, R.C. Lonsinger, and L.P. Waits. (2015). Evaluation of Scat Deposition Transects versus Radio Telemetry for Developing a Species Distribution Model for a Rare Desert Carnivore, the Kit Fox. *PLOS One 10*(10):e0138995.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubius, A., Pellissier, L., Mateo, R. G., Hordijk, W., Salamin, N., & Guisan, A. (2017). ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography, 40*(6), 774-787.

Džeroski, S., & D. Drumm. (2003). Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Islands. *Ecological modeling, 170*, 219–226.

Elith, J., & Burgman, M.A. (2002a). Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: *Predicting Species Occurrences: Issues of Accuracy and Scale,* Editors: Scott, J.M., Heglund, P. J., Morrison, M. L., Raphael. M. G., Wall, W. A., & Samson, F. B. Island Press, Covelo, CA, 303-314.

Elith, J., Burgman, M. A., & Regan, H. M. (2002b). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling, 157*, 313–329.

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling, 186*, 280–289.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R. , Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S., Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), 129–151.

Elith, J., Leathwick, J, R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*, 802-813.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography, 32*(1), 66–77.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions, 17*(1), 43-57.

Ferrier, S., Watson, G., Pearce, J., & Drielsma, M. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation, 11*, 2275–2307.

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. Journal of *Applied Ecology, 43*(3), 393-404.

Fieldling, A., & Bell, J. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation, 24*(1), 38-49.

Findley, J. (2010). *Presence-only modeling with Maxent* (Rep.). Syracuse University.

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. *PLoS ONE, 9*(5).

Franklin, J., & Miller, J. A. (2010). Mapping species distributions: spatial inference and prediction. Cambridge, UK: Cambridge University Press. 338 pp.

Friedman, J. H. Multivariate adaptive regression splines. (1991). *Ann. Stat 19,* 1–141.

Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, A. V. Mishonov, O. K. Baranova, M. M. Zweng, J. R. Reagan, D. R. Johnson, (2013a). World Ocean Atlas 2013. Vol. 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. S. Levitus, Ed.; A. Mishonov, Technical Ed. NOAA Atlas NESDIS 75, 27 pp

Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, O. K. Baranova, M. M. Zweng, J.R. Reagan, D. R. Johnson, (2013b). World Ocean Atlas 2013. Vol. 4: Dissolved Inorganic Nutrients (phosphate, nitrate, silicate). S. Levitus, Ed.; A. Mishonov, Technical Ed. NOAA Atlas NESDIS 76, 25 pp.

Guinotte, J. M., & Davies, A. J. (2014). Predicted Deep-Sea Coral Habitat Suitability for the U.S. West Coast. *PLoS ONE, 9*(4).

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*, 147–186.

Guisan, A., T.C. Edwards, Jr., and T. Hastie. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling 157*(2-3), 89–100.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters, 8,* 993-1009.

Guisan, A., & Rahbek, C. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography, 38,* 1433-1444.

Hijmans, R. J., & Elith, J. (2017). Species distribution modeling with R. R Package Vignette.

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: The importance of concepts in species distribution modelling. *Diversity and Distributions, 14*, 885–890.

Kent, R., & Carmel, Y. (2011). Presence-only versus presence-absence data in species composition determinant analyses. *Diversity and Distributions, 17*(3), 474-479.

Kestrup, Åsa M., Smith, D.L. and Therriault, T.W. (Eds.) (2015). Report of Working Group 21 on Non-indigenous Aquatic Species. PICES Sci. Rep. No. 48, 176 pp.

Knudby, A., Ledrew, E., & Brenning, A. (2010). Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine learning techniques. *Remote Sensing of Environment, 114*(6), 1230-1241.

Knudby, A., Kenchington, E., & Murillo, F. J. (2013). Modeling the distribution of geodia sponges and sponge grounds in the northwest Atlantic. *PLoS ONE, 8*(12).

Landis, J. R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Levesque, C., & Jamieson, G.S. (2015). Identification of Ecologically and Biologically Significant Areas in the Strait of Georgia and off the West Coast of Vancouver Island: Phase I - Identification of Important Areas. DFO Can. Sci. Advis. Sec. Res. Doc. 2014/100. viii + 68 p.

Leys, S. P. (2003). The significance of syncytial tissues for the position of the hexactinellida in the Metazoa. *Integrative and Comparative Biology, 43*(1), 19–27.

Leys, S., Wilson, K., Holeton, C., Reiswig, H., Austin, W., & Tunnicliffe, V. (2004). Patterns of glass sponge (Porifera, Hexactinellida) distribution in coastal waters of British Columbia, Canada. *Marine Ecology Progress Series, 283*, 133-149.

Leys, S. P., Yahel, G., Reidenbach, M. A., Tunnicliffe, V., Shavit, U., & Reiswig, H. M. (2011). The Sponge Pump: The Role of Current Induced Flow in the Design of the Sponge Body Plan. *PLoS ONE, 6*(12).

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., Paver, C. R., Reagan, J. R., Johnson, D. R., Hamilton, M., & Seidov, D. (2013). World Ocean Atlas 2013, Volume 1: Temperature. S. Levitus, Ed.; A. Mishonov, Technical Ed.; NOAA Atlas NESDIS 73, 40 pp.

Leys, S. P., & Kahn, A. S. (2018). Oxygen and the energetic requirements of the first multicellular animals. *Integrative and Comparative Biology, 58*(4), 666–676.

MacKenziem, D. I., (2005). Was it there? Dealing with imperfect detection for species presence/absence data. *Australian & New Zealand Journal of Statistics, 47*(1), 65-74.

Maldonado, M. (2006). The ecology of the sponge larva. *Canadian Journal of Zoology, 84*, 175–194.

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *Peer J, 5*:e2849

Monk, J., Ierodiaconou, D., Harvey, E., Rattray, A., & Versace, V. L. (2012). Are we predicting the actual or apparent distribution of temperate marine fishes? *PLoS ONE, 7*(4).

Monserud, R. A., Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling, 62*, 275–293

Nedjah, N. & Luiza de Macedo, M. (2005). Fuzzy Systems Engineering: Theory and Practice; Springer: New York, NY, USA.

Nigam, K., Lafferty, J., & Mccallum, A. (1999). Using Maximum Entropy for Text Classification. IJCAI-99 Workshop on Machine Learning for Information Filtering, 61–67.

Pagel, J., & Schurr, F. M. (2012). Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography, 21*(2), 293–304.

Pearce, J., & Ferrier, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling, 128*, 127–147.

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography, 12*(5), 361–371.

Pearson, R. G. (2010). Species' distribution modeling for conservation educators and practitioners. Network of Conservation Educators and Practitioners, Center for Biodiversity and Conservation, American Museum of Natural History. *Lesson in Conservation*, *3*, 54–89.

Pennino, M. G., Conesa, D., Lo´pez-Quı´lez, A., Mun˜oz, F., Ferna´ndez, A., & Bellido, J. M. (2016). Fishery-dependent and -independent data lead to consistent estimations of essential habitats. *ICES Journal of Marine Science, 73*(9), 2302–2310.

Perry, R. I., Stocker, M., & Fargo, J. (1994). Environmental effects on the distributions of groundfish in Hecate Strait, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, *51*, 1401–1409.

Phillips, S. (n.d.). A Brief Tutorial on Maxent (pp. 1-38, Rep.). AT&T Research.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modeling, 190*(3-4), 231-259.

Phillips, S.J., Dudik, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications, 19*, 181–197.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Redding, D. W., Lucas, T. C. D., Blackburn, T. M., & Jones, K. E. (2017). Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. *PLoS ONE 12*(11): e0187602.

Reiss, H., Cunze, H., König, K., Neumann, K., & Kröncke, I. (2011). Species distribution modelling of marine benthos: A North Sea case study. *Marine Ecology Progress Series, 442*, 71–86.

Reiswig, H. M., & Mackie, G. O. (1983). Studies on Hexactinellid Sponges. III. The Taxonomic Status of Hexactinellida within the Porifera. *Philosophical Transactions of the Royal Society B: Biological Sciences, 301*, 419-428.

Rivera, O. R., & Lopez-Quilez, A. (2017). Development and comparison of species distribution models for forest inventories. *ISPRS International Journal of Geo-Information, 6*(6), 176.

Roberts, J. M., Long, D., Wilson, J. B., Mortensen, P. B., & Gage, J. D. (2003). The cold-water coral Lophelia pertusa (Scleractinia) and enigmatic seabed mounds along the north-east Atlantic margin: are they related? *Marine Pollution Bulletin 46*, 7–20.

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.

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*(421).

Rooper, C. N., Zimmermann, M., & Prescott, M. M. (2017). Comparison of modeling methods to predict the spatial distribution of deep-sea coral and sponge in the Gulf of Alaska. *Deep Sea Research Part I: Oceanographic Research Papers*, *126*, 148–161.

Rowden, A., Anderson, O. F., Georgian, S. E., Bowden, D. A., Clark, M. R., Pallentin, A., & Miller, A. (2017). High-resolution habitat suitability models for the conservation and management of vulnerable marine ecosystems on the Louisville Seamount Chain, South Pacific Ocean. *Frontiers in Marine Science, 4*:335.

Rudnick, D., Beier, P., Cushman, S., Dieffenbach, F., Epps, C.W., Gerber, L., Hartter, J., Jenness, J., Kintsch, J., Merenlender, A.M., Perkle, R.M., Preziosi, D.V., Ryan, S.J., and S. C. Trombulak. (2012). The Role of Landscape Connectivity in Planning and Implementing Conservation and Restoration Priorities. *Issues in Ecology.* Report No. 16. Ecological Society of America. Washington, DC.

Shabani, F., Kumar, L., & Ahmadi, M. (2016). A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecology and Evolution, 6*(16), 5973–5986.

Steinacher, M., Joos, F., Frölicher, T. L., Plattner, G. K., & Doney, S. C. (2009). Imminent ocean acidification in the Arctic projected with the NCAR global coupled carbon cycle-climate model. *Biogeosciences*, *6*, 515–533.

Svensson, J.R., Jonsson, L., & Lindegarth, M. (2013). Excessive spatial resolution decreases performance of quantitative models, contrary to expectations from error analyses. *Marine Ecology Progress Series, 485*, 57-73.

Swets, K. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE, 8*(2).

Thuiller, W. (2003). BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology, 9*, 1353–1362.

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography, 32*(3), 369-373.

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7. https://CRAN.R-project.org/package=biomod2

Thurber, A. R., Sweetman, A. K., Narayanaswamy, B. E., Jones, D. O., Ingels, J., & Hansman, R. L. (2014). Ecosystem function and services provided by the deep sea. *Biogeosciences, 11*(14), 3941-3963.

Treguer, P., Nelson, D, M., Van Bennekom, A. J., DeMaster, D. J., Leynaert, A., & Queguiner, B. (1995). The silica balance in the world ocean: a reestimate. *Science, 268*, 375–379.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions, 13*, 397–405.

US Department of Commerce, & National Oceanic and Atmospheric Administration. (2018). Are glass sponges made of glass? Retrieved from https://oceanexplorer.noaa.gov/facts/glass-sponges.html

Václavík, T., J.A. Kupfer, and R.K. Meentemeyer. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography 39*(1), 42-55.

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11).

Vaughan, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species distribution models. *Journal of Applied Ecology, 42*, 720-730.

Vieilledent, G., Merow, C., Guelat, J., Latimer, A., Kery, M., Gelfand, A., Wilson, A., Mortier, F., & Silander, Jr. J. (2014) Hierarchical Bayesian species distribution models with the hSDM R Package. URL http://hsdm.sourceforge.net/wp-content/uploads/2014/07/hSDM-vignette.pdf

Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). mvabund - an R package for model- based analysis of multivariate abundance data. *Methods Ecol. Evol.,* 3, 471–474.

Ward, G. (2007). Statistics in ecological modeling: presence-only data and boosted mars. Stanford University, Palo Alto.

Ward, G., Hastie, T., Barry, S., Elith, J., & Leathwick, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics, 65*(2), 554–563.

Whitney, F., Conway, K., Thomson, R., Barrie, V., Krautter, M., & Mungov, G. (2004) Oceanographic habitat of sponge reefs on the Western Canadian continental shelf. *Continental Shelf Research.*

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., & Grehan, A. J. (2007). Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Marine Geodesy, 30*, 3-35.

Wintle, B. A., McCarthy, M. A., Parris, K. M., & Burgman, M. A. (2004). Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications, 14*(3), 703-712.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J. A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M. C., Normand, S., Öckinger, E., Schmidt, N. M., Termansen, M., Timmermann, A., Wardle, D. A., Aastrup, P., & Svenning, J. C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews, 88*, 15–30.

Zhang, L., Liu, S., Sun, P., Wang, T., Wang, G., Zhang, X., & Wang, L. (2015). Consensus forecasting of species distributions: The effects of niche model performance and niche properties. *PLoS ONE, 10*(3).