UNIVERSITY COLLEGE ROOSEVELT
Lange Noordstraat 1, Middelburg, The Netherlands

Bachelor of Liberal Arts & Sciences

SENIOR PROJECT

Predictive Modeling of Suitable Habitat for Scleractinia Deep-Sea Corals

Validating coral occurence data and exploring possible habitats

Author:
E. A. S. Engh

Student Number:
6965997

Supervisor:
Dr. F. van der Stappen

External supervisor:
Dr. A. van der Kaaden

Second examiner:
Dr. M. Jansen

July 23, 2022

Abstract

Habitat suitability models have become an important tool for insights into the distribution of species for conservation efforts. Due to the high costs of surveying and sampling the deep sea, habitat suitability models are essential to estimate the full extent of deep-sea habitats and aid conservation management of deep-sea species, including deep-sea corals. The initial focus of this project was the creation of a global dataset of deep-sea framework-forming (Scleractinia) corals. A large proportion of the observations in this dataset were subjectively considered as unreliable due their old sampling age and/or lack of sampling information. The observations were therefore classified as reliable or unreliable based on available sampling information and/or sampling age. Several habitat suitability models were created to investigate the reliability of the unreliable-classified observations. By comparing the locations of the unreliable-classified observations with areas of predicted suitable habitat, one would get an idea of the true reliability of the unreliable-classified observations, and the sensibleness of the reliability classification scheme. Also, creating a well-performing habitat suitability model was in itself an aim of the project. The study area was a region surrounding the continental shelf offshore the US east coast. This region contained plentiful of Scleractinia observations both classified as reliable and unreliable. A diverse group of predictor variables was used, concerning seafloor terrain and oceanography. These variables were obtained from various different source data, and processed to create a gridded dataset with a resolution of 15 arc-sec for the study area. The dataset was used to train models on reliable coral observations, and the models were tested on both reliable and unreliable observations. The model results gave an indication on the different spatial patterns of the unreliable and reliable observations, and hence, the reliability of the unreliable observations. Model performance differed on the unreliable and reliable test sets. The difference in model performance showed that the reliable and unreliable observations had different spatial patterns, and that the unreliable observations could be considered truly unreliable. The reliability classification scheme was shown to be sensible. The predicted areas of suitable habitats for Scleractinia corals were on the continental shelf offshore the US east coast. The continental shelf offshore Northeastern USA saw a narrow areas of suitable habitats, and the area of suitable habitats widened as one moved south along the continental shelf to the Florida coast.

Contents

1 Introduction

Coral reefs form some of the most biodiverse ecosystems on Earth. An estimated 25% of all marine life are believed to depend on coral reefs throughout their life cycle; an estimate which highlights the importance of conservation efforts towards coral reef systems [1]. Deep-sea corals are non-photosynthesizing corals that exists at depths between 50-3000m on solid substrates of continental shelves, canyons or sea mounds [2]. They form biodiverse habitats in the deep-sea, and provide ecosystem services for benthic communities such as fishes and invertebrates. Several taxonomic groups of corals are represented in the deep-sea; including the order Scleractina, which consist of the coral species that are framework-forming (reef-building) [3]. Although deep-sea corals are relatively well researched in comparison to other deep-sea ecosystems, they are susceptible to anthropogenic impact such as ocean acidification, bottom trawling, offshore drilling and deep-sea mining [4]. Furthermore, deep-sea corals are vulnerable due to their slow recovery rates. The susceptibility and vulnerability of deep-sea corals to anthropogenic activities have led to increased awareness and planning of conservation efforts [5].

Habitat suitability modeling is an important tool for insights of the spatial distribution of deep-sea corals for conservation efforts. Collecting spatial data of deep-sea corals is a costly procedure, due to the operation of remotely operated vehicles (ROVs) and ships with multibeam capacity [5]. The high costs associated with data collection of deep-sea corals, highlights the need for well-planned coordination of deep-sea surveying and refined habitat suitability models for conservation management and the establishment of Marine Protected Areas (MPAs) [4] [5].

The majority of habitat suitability models are correlative models that link the spatial distribution of species with spatial niches given by environmental data. When a spatial pattern between species occurrences and environmental data is identified, predictions of habitat suitability are made based on new environmental data input [3]. Several machine learning methods have been used for habitat suitability modeling of deep-sea corals, including maximum entropy (maxent), boosted regression trees, random forest and deep nerual networks. Maxent has been the most common model type for habitat suitability modeling of deep-sea corals, and has performed well in past research [3]. Further details of habitat suitability modeling is described in the Literature Review (Section 2).

The aim of this project is to use habitat suitability modeling to validate presence occurrences, and explore suitable habitats of Scleractinia corals offshore the US east coast. The first phase of this project was to create a global database of deep-sea corals in the order Scleractinia. After the creation of the database, a large proportion of observations were observed to lack important information about sampling accuracy and sampling method. Furthermore, many observations derived from old records, dating back 70-154 years. Observations lacking information on either sampling accuracy or sampling method (or both), and observations that derived from old records, were considered as unreliable.

In order to investigate the reliability of the unreliable-classified observations, a maxent habitat suitability model was created based on observations classified as reliable. If unreliable-classified observations were within regions of high habitat suitability for Scleractinia coral, the observations could be reconsidered as reliable. However, if the unreliable-classified regions fell into areas con-

sidered to have low habitat suitability, the unreliable-classified observations would continue to be considered as unreliable. If many unreliable-classified observations were located in areas predicted to have low habitat suitability, the classification scheme of reliability/unreliability would be confirmed to be sensible. Considering the aim of the project, the research question is the following: *how reliable are the unreliable-classified Scleractinia observations within the global Scleractinia dataset?* Also, its important to note that creating a well-performing habitat suitability model was a goal of itself in this project.

An area surrounding the continental shelf offshore the US east coast was chosen as the study area. This area contained an abundant amount of both reliable and unreliable observations. A restricted region was chosen for model predictions, due to lack of computational power for ocean-scale predictions. The resolution of the maxent model was 15 arc-sec. Compared to habitat suitability modeling of deep-sea corals in past research, 15 arc-sec is relatively high resolution for modeling (see Table 1 for resolutions used in past research).

2 Literature Review

A literature review of habitat suitability modeling was conducted prior to modeling efforts. The focus of the review was on the methodology of past research; to get an understanding of the common procedures taken for habitat suitability modeling of deep-sea corals. Each subsection reviews important aspects of habitat suitability modeling of deep-sea corals: the sourcing of data, data processing, model types, common predictor variables, variable selection and model validation.

2.1 Data Sources

The deep sea has long been a data-poor environment due to lack of light penetration, costs needed for sampling and difficult sampling logistics. However, in recent years there has been better access to deep-sea data due to improvements in remote sensing and centralized oceanographic databases [6]. Studies in the existing literature have used datasets from Gebco for bathymetry (ocean depth) data. From bathymetry data, different seafloor terrain metrics can be computed with GIS software. The World Ocean Atlas (WOA) has been used to obtain seafloor nutrient, temperature and salinity data [3, 5]. The WOA data contains temperature and nutrient data on a grid of 1^o or $1/4^o$, which is partitioned into standardized depth levels throughout the world's oceans down to depths of 5500m [7]. Hydrodynamic data utilized in past research has been modeled with the Regional Ocean Modeling System (ROMS) [3, 8].

2.2 Data Processing

Existing Deep-sea species records tend to lack information on the absence of species. Species records containing data on presence and absence derive from systematic data collection, often in the form of biological surveying of a restricted study area. For most regions, survey data of deep-sea species is sparse and/or limited in extent. Most deep-sea species data derives from past recordings of species occurrences; a process of data collection spanning well over a century. In order to utilize the available deep-sea occurrence data, algorithms using presence-only data is most commonly used for habitat suitability modeling [9]. Common practice in habitat suitability modeling of deep-sea

species is to use species presence data, and environmental data in the study area where presences are lacking (often called background data). Creating background data helps distinguish suitable and unsuitable environmental factors for the occurrences of deep-sea corals. The background data can be generated at random or by the user within the geographical area of the analysis [6]. In the review of the topic by Vierod et al, only one research paper based their habitat suitability model on presence-absence data [6]. The models using presence-absence data have been geographically restricted and only used for local predictions [10]. The remaining studies that were reviewed by Vierod et al used presence- background or presence only data [6].

Variables that are used in habitat suitability models derive from different sources and are geo-tagged with different grid sizes [3, 5, 8, 11, 12]. In order to create a dataset ready for modeling, the variables used in the model need to be interpolated to a common grid. The grid size depends on the area covered by model predictions. Habitat suitability models that predict over a large area tend to have lower resolution, due to limited data and computational power [6]. In Table 1, one can see research papers concerning habitat suitability models on a regional and global scale and the corresponding resolution of each model.

Table 1: The resolution used in habitat suitability of some past studies.

Spatial Extent	Author	Resolution (approximate)
Global	Davies et al [11]	3600 arc-sec
	Davies & Guinotte [5]	30 arc-sec
Regional	Hu et al [3]	15 arc-sec
	Kinlan et al [2]	15 arc-sec
	Rooper et al [8]	6 arc-sec
	Rengstorf et al [12]	2.6 arc-sec
	Georgian et al [13]	0.9 arc-sec

In past research concerning habitat suitability modeling of deep-sea corals, common procedure has been to retain one observation within a grid cell that contains multiple observations. This procedure is relevant when the habitat suitability model is created with presence-background data. Davies et al, and Hu et al created habitat suitability models for the prediction of deep-sea Scleractinia, and followed this procedure [3, 5]. Multiple datapoints within one cell make the model depend heavily on the environmental conditions within this cell for prediction. The environmental conditions within cells with many datapoints might be especially suitable for particular species of Scleractinias. Removing all but one datapoint within cells removes bias towards certain Scleratinia species and reduces overfitting [5]. In the habitat suitability model created by Hu et al, grid cells containing a presence of deep-sea corals were encoded as "1", while the background cells containing no corals were encoded with "0" in the raster data layer containing coral occurrences [3].

2.3 Habitat Suitability Model Types

Several algorithms have been developed for habitat suitability models in the deep sea. Most of these models use presence-only data, with a few exceptions of models using presence-absence data. Common models in past research is Environmental Niche Factor Analysis model (ENFA), Maxi-

mum Entropy (maxent), Random Forests and Boosted Regression Trees [3, 14]. Other models that have been used are Support Vector Machines (SVM) and Neural Networks [3].

Early use of habitat suitability models for predictions of deep-sea corals were mostly based on ENFA models. This model is based on presence-background data. The predictions are made by relating the mean environmental conditions of sites with species occurrences to the environmental conditions in the background data [6]. Later research applied the maxent model due to its stronger predictive power. Recent research has also combined predictions from various machine learning models. Generally, there has been no preferred model. When several models are used, one can compare overlapping results, which contributes to confidence in the predictions. Hu et al created a model for predictions of Scleratinian coral occurrences in the deep sea by combining maxent, support vector machine, random forest and deep neural network models. Combining various machine learning models to create robust predictions has increased in popularity for habitat suitability modeling of deep-sea corals [3].

The maxent model is the most common habitat suitability model for coral occurrence predictions. The model has proven to perform well for various different data situations [6]. Furthermore, user-friendly habitat suitability softwares utilizing the maxent model have also been developed, making the model approach accessible to the public. The underlying assumption of the maxent model is that the prediction of an unknown distribution (in this case the distribution of deep-sea corals) is based on maximum entropy constraints of the predictor variables (in this case environmental variables). The constraints of the environmental variables are found prior to modeling by investigating the environmental variables needed for a uniform distribution of the coral occurrence data [3, 5]. A uniform distribution of deep-sea corals can be interpreted as a situation of maximum entropy. The final output of the model is then the distribution of coral occurrences with environmental variables satisfying the constraints of maximum entropy. The output of the model is a probability of habitat suitability within each cell grid [6].

2.4 Predictor Variables

Common predictor variables that have been used for habitat suitability models are bathymetric, oceanographic, chemical and biological. In past research, environmental predictor variables differed depending on the study area, coral taxa, model resolution, availability of data and correlation among variables [3]. Table 2 of the Appendix (see section A) shows the predictor variable that were used in some of the research papers that were investigated for this literature review.

2.5 Variable Selection

In habitat suitability models of deep-sea corals, covariance between predictor variables is common [3, 5, 12, 15]. Covariance does not affect machine learning accuracy, but causes inaccurate estimates of variable importance scores [3]. In past research of deep sea coral habitat suitability models, common practice has been to investigate covariance with a correlation matrix or by using variance inflation factors (VIF) for each variable [5, 15]. When using correlation matrices, one of the two correlating variables are omitted from the modeling dataset. Davies & Guinotte excluded one of the two variables when the correlation coefficient was above 0,7 [5]. Morato et al excluded one of the two variables when the correlation coefficient was above 0,85. Common practice is to retain

the variables that are believed to have highest ecological significance [5]. In the research conducted by Morato et al, the variables depth, dissolved oxygen concentration and pH were excluded prior to modeling [15]. When VIFs are calculateds, variables exceeding a certain threshold are excluded. A threshold value of 10 was chosen by Hu et al, causing the exclusion of nutrient and bathymetric variables (nitrate, phosphate, silicate, plane curvature, and profile curvature) [3]. Rooper et al chose a threshold factor of 5, with all variables scoring below this value, meaning all variables could be retaiend for modeling [8].

2.6 Variable Importance

Past research of habitat suitability modeling has commonly used statistical methods to evaluate the importance of predictor variables [3, 15, 12]. Rengstorf et al assessed variable importance by generating combinations of the six predictor variables ($2^6 = 64$ models), and testing the performance of each model using the Aikaike Information Criterion (AIC). The AIC quantifies the relative performance between different models [12]. .

Hu et al estimated the importance of each variable in the maxent model by using the standard variable importance metrics contained in the *maxent* function in R [3]. The standard variable importance metrics are *percent contribution* and *permutation importance*. The *percent contribution* metric quantifies the contribution of each predictor variable in the fitting of the model on the training set. The *maxent* function keeps track on the model gain caused by respective predictor variables for each iteration in the model fitting process. After the fitting process is complete, the gain caused by the predictor variables are calculated to percentages, giving the *percent contribution* metric. It is important to note that the *percent contribution* metric depends on the particular path that the maxent model takes when fitting the training data. A different maxent algorithm can arrive at the same result by taking a different path, giving different *percent contribution* scores [16]. On the other hand, the *permutation importance* metric only depends on the final model, not the path taken during the training process. *Permutation importance* scores are calculated by randomly permuting one of the predictor variables. The model is then re-fit on the training set, which now includes the permuted variable. The drop in training AUC due to the permuted variable gives the *permutation importance* score, normalized to percentages, for the variable. Every predictor variable is permuted one-by-one, and the model is re-fitted for every permuted variable. The result is *permutation importance* scores for every predictor variable [16].

2.7 Model Evaluation

For models that give a probability of occurrence as output, a threshold value (ranges from 0-1) is needed to classify presence/absence of species prior to evaluation. Generally in habitat suitability studies, a model is evaluated on a testing dataset, which is independent of the dataset used to train the model. When the habitat suitability model is applied on the test set, the proportion of correctly and incorrectly predicted coral occurrences are reported in the form of a correlation matrix or as the area under the receiver operating characteristic curve (AUC) score [6].

The AUC metric is the most common measure of model performance for habitat suitability models, and was used in most studies investigated in this review [3, 4, 5, 6, 8, 12, 13]. The exception was the study by Davies et al, where the ENFA model was used as a habitat suitability model and the

model performance measure was the absolute validation index (AVI) [11]. In the literature review by Bowden et al, all investigated studies used the AUC as a model performance metric [14]. The AUC is a single value metric of model performance, and is the calculated area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of all correctly predicted occurrences in the test set (y-axis), against all wrongly predicted occurrences (x-axis) for every threshold value (continuous variable).

3 Materials

This section describes the different datasets that were used for this project. The *initial datasets* were source data obtained online or from the existing literature. The *global Scleractinia dataset*, and the *dataset for modeling* were datasets that were created from the source data.

3.1 Initial Datasets

Coral species occurrence datasets and datasets concerning bathymetry and oceanography were obtained from online databases in order to create a global Scleractinia coral dataset, and a dataset for habitat suitability modeling of Scleractinia coral. The datasets were obtained from the following databases: Ocean Biodiversity Information System (OBIS), National Oceanographic and Atmospheric Administration (NOAA), International Council for the Exploration of the Sea (ICES), General Bathymetric Chart of the Oceans (GEBCO), The World Ocean Atlas (WOA) and the Nauru Environmental Data Portal [17]. Table 9 in the Appendix shows all datasets that were used for the creation of the global Scleractinia dataset.

3.2 Global Scleractinia Dataset

In the initial part of the project, a global deep-sea Scleractinia coral dataset was created for further research and exploratory data analysis. The dataset contained information on the taxonomy, location, life stage, reliability and sampling of each coral observation. The variable *coral reliability* was a binary variable that classfied each observation as either *reliable* or *unreliable*. The variable was created to distinguish observations based on available sampling information. Observations were labelled as unreliable if there was no available information for at least one of the following variables: *sampling method*, *sampling accuracy* and *sampling year*. Also, observations sampled before the year 1950 were labelled as unreliable. The year 1950 was chosen as an arbitrary cut-off point.

3.3 Dataset for Modeling

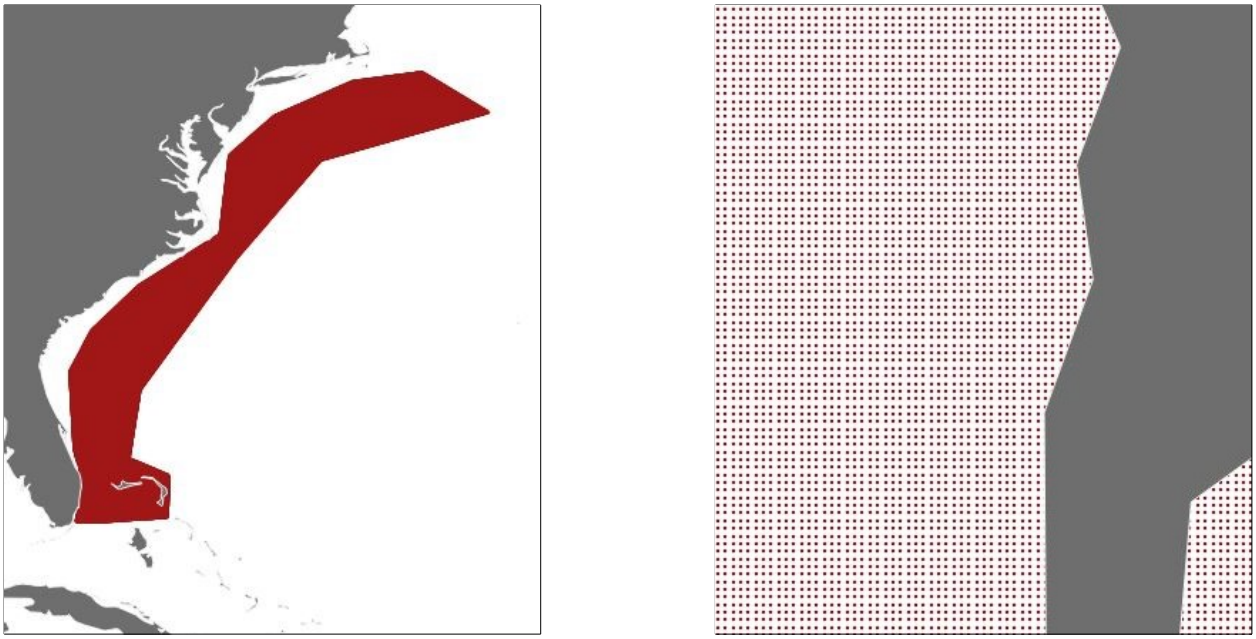
The dataset used for modeling contained variables concerning coral occurrence, reliability of coral occurences, seafloor terrain and oceanography. Table 2 shows the variables in the dataset. The *coral presence* variable is the response variable of the model, while the seafloor terrain and oceanography variables are predictors. The coordinate variables give the location of each observation, which are respective grid points in the study region. The distance between each grid point in the horizontal and vertical direction is 15 arc-seconds. Therefore, the dataset can be interpreted as a gridded dataset with a resolution of 15 arc-seconds. The red area in Figure 1a illustrates the grid points covering the entire study region offshore the US east coast. Due to the high resolution of the grid points, the area is shown as a continuous surface. There are in total 2,294,843 grid points, which

corresponds to the same amount of observations in the dataset. Figure 1b is a visual aid to comprehend the resolution of the grid points. The enlarged area in the figure area allows to visualize the grid points as discrete. The enlarged area is in the bottom-right area of the study region, on the coast of Great Abaco Island, The Bahamas.

The dataset used for modeling is a *presence-background* dataset, meaning observations labelled to have coral absence (in the *coral presence* variable seen in Table 2) are merely areas within the study region with no coral data; it is unknown if the coral absence entries are areas of actual coral absence. The *coral presence* variable is binary, meaning each observation has coral presence (the variable entry is 1), or lack of data (the variable entry is *0*). The *coral presence* and coral reliability variables are derived from the *global Scleractinia dataset*, which only contained coral presence observations in certain areas of the study region. If a cell in the 15 arc-sec grid overlapped with areas with coral observations from the *global Scleractinia dataset*, the corresponding grid point had coral presence entries. On the contrary, grid cells that did not overlap with coral observations from the *global Scleractinia dataset* had coral absence entries in the corresponding grid point. In total, 464 observations were presence observations, while the remaining 2,294,379 observations were background observations. The details of creating the modeling dataset is described in subsection 4.1.

Table 2: Variables in the dataset used for modeling

Type of data	Variable	Data source
Coral data	Coral reliability	Categorical variable created by author
	Coral presence Coordinates	OBIS, NOAA and ICES
Seafloor terrain	Bathymetry Aspect (15 arc-second scale) Aspect (60 arc-second scale) Aspect (200 arc-second scale) Slope (15 arc-second resolution) Slope (200 arc-second resolution) Curvature (15 arc-second resolution) Curvature (60 arc-second resolution) Curvature (200 arc-second resolution) Roughness (15 arc-second resolution) Roughness (60 arc-second resolution)	GEBCO 2021 Grid [18]
Oceanography	Apparent oxygen utilization Dissolved oxygen Percent oxygen utilization Nitrate Phosphate Salinity Silicate Temperature	The World Ocean Atlas [7]
	Chlorophyll a	Nauru Environmental Data Portal [17]



(a) Grid points covering the study region

(b) Grid points in an enlarged area

Figure 1: Each observation in the dataset is located at respective grid points. The grid points are showed for the whole region (a), and for an enlarged area (b).

4 Methods

This section documents the methodology of the project. The methodology of this project can be divided into five main parts: data processing, variable selection, maxent modeling, model validation and variable importance evaluation. The methodology of these parts are described in more detail in this section.

4.1 Data Processing

The data processing part of the project involved processing coral, bathymetry and oceanography data to make a dataset ready for modeling. The softwares R and QGIS were used for data processing. All datasets were downloaded as raster files, or converted to the raster file format. The resolution of the raster datasets was increased using cubic spline interpolation. Cubic spline was chosen as the interpolation method to create smooth transitions between cells in the interpolated grid. The raster datasets were then resampled to the grid points shown in Figure 1. These grid points were approximately located in the centre of each grid cell in the raster datasets. After this re-sampling procedure, the datasets were merged based on their joint coordinate variable. More specific processing steps of the datasets are described in the subsections.

4.1.1 Scleractinia Dataset

A deep-sea Scleractinia dataset was created in the initial part of the project. The dataset was based on deep-sea coral data obtained from OBIS, NOAA and ICES. All datasets that were used

from these sources are shown in Table 9 in the Appendix. The source data was filtered in R to only contain the variables concerning taxonomy, location, life stage, reliability and sampling for each observation. If the sourced datasets missed some of these variables, the variable was created with *not available (NA)* entries. Also, the variable *coral reliability* was created, and each observation in the Scleractinia dataset was classified as either reliable or unreliable based on the classification scheme presented in subsection 3.2.

4.1.2 Dataset for modeling

The gridded dataset used for modeling contained Scleractinia presence observations and background observations for the study area. The *gridRecords* function from the *fuzzySim* library was used to transform the Scleractinia dataset into a presence-background dataset [19]. The function takes a raster file (gridded data file) and species occurrence coordinates as input, which gives a gridded dataset containing a presence-absence variable (binary variable) as output. Species abundance within each grid cell is not given in the output; only the presence/absence of the species. The raster file is necessary as input to specify the geographical extent and resolution of the presence-background output dataset. The bathymetry raster file was used as input, in order to get a presence-background dataset with a resolution of 15 arc-seconds for the study area. Each grid cell that contained unreliable coral observations were re-labelled as reliable, while grid cells containing unreliable observations were re-labelled as unreliable.

4.1.3 Bathymetry and Seafloor Terrain Data

The seafloor terrain metrics were calculated from the bathymetry with QGIS. Datasets were produced for each terrain metric at different resolutions, to capture small scale and large scale variations in the seafloor terrain. For the terrain metrics capturing large-scale variations, the resolution of the bathymetry data was lowered to 60 arc-seconds or 200 arc-seconds. Then, the raster analysis tools in QGIS were used to create aspect, slope, roughness and curvature datasets at lower resolutions. The lower-resolution datasets were then resampled to grid points with a resolution of 15 arc-seconds.

4.1.4 Environmental Data From The World Ocean Atlas

The environmental datasets (temperature, salinity, silicate, phosphate, nitrate, percent oxygen utilization, dissolved oxygen, oxygen saturation) from the World Ocean Atlas were multilayered gridded (raster) datasets for standardized depth levels. The datasets had a resolution of either 900 or 3600 arc-seconds. Each layer represented depth levels that in total spanned the ocean depths from 0 – 5000m. The resolution of the environmental datasets was increased to 15 arc-seconds. Then, the datasets were clipped using QGIS to only cover the study area (see Figure 1a). Deep-sea corals only live on the seafloor. Therefore the seafloor values were extracted from each grid cell of the gridded dataset. The result was a 15 arc-second gridded dataset that only contained the seafloor values of the study area.

4.1.5 Chlorophyll-a Data

The chlorophyll-a dataset was created from mean monthly chlorophyll-a datasets between the years 2010 to 2019. The datasets were in raster format, and were obtained from the Nauru Environment

Data Portal [17]. Using QGIS, the raster datasets were re-formatted to a single, multi-layer raster in QGIS, and resampled to the grid points shown in Figure 1. The resulting dataset contained observations with coordinates of the respective grid point, and monthly variables of chlorophyll-a concentration between the years 2010-2019. Using R, the mean concentration across all months were calculated for each observation, and stored in a new variable. Only three variables were retained in the final chlorophyll-a dataset: the coordinate variables and the calculated mean value across all months.

4.2 Variable Selection

Covariance among predictor variables does not affect machine learning model performance but causes unreliable calculations of variable importance scores [3]. Therefore, a variable selection procedure was done prior to modeling. A Pearson's correlation coefficient of 0.7 was chosen as a threshold to omit covarying variables. Variables with high covariance were grouped together, and a single variable from each group was retained. Variables were retained based on variable importance scores in past research. Table 3 shows the groups of variables with correlation coefficients above 0.7 among themselves, and the retained variable from each group. The retained variables from each group were temperature, slope (15 arc-sec) and dissolved oxygen. These variables were retained based on variable importance scores from Hu et al, Kinlan et al and Davies & Guinotte [3, 2, 5]. The following variables were omitted: silicate, percent oxygen saturation, salinity, nitrate, phosphate, apparent oxygen utilization, slope (15 arc-sec scale), roughness (60 arc-sec scale) and curvature (200 arc-sec scale).

Table 3: Variable groups with high correlation among themselves (Pearson correlation coefficent above 0.7), and the variable of each group that was retained.

Covarying variables	Retained variable
Temperature Silicate Percent oxygen saturation Salinity	Temperature
Dissolved oxygen, Nitrate Phosphate Apparent oxygen utilization	Dissolved oxygen
Slope (15 arc-sec scale) Rougness (15 arc-sec scale) Rougness (60 arc-sec scale) Curvature (200 arc-sec scale)	Slope (15 arc-sec scale)

4.3 Maximum Entropy Modeling

The maximum entropy model was chosen as the habitat suitability model to validate unreliable coral occurrences, and explore suitable habitats for Scleractinia coral offshore the eastern United States. First, a model was trained and tested on reliable coral presences. This model was created to investigate model performance for predictions of of reliable coral coral presences. A second model

was trained on all reliable coral presences and tested on all unreliable coral presences. The purpose of the second model was to investigate model performance for predictions of unreliable coral presences. If the first model performed well, the performance of the second model would give an indication of the reliability of the unreliable-classified coral presence observations. A third model was trained on reliable coral presences and tested on unreliable coral presences separately on the lower and upper half of the study region. The purpose of the third model was to investigate if the spatial relationship between reliable and unreliable observations differed in the upper half and lower half of the region.

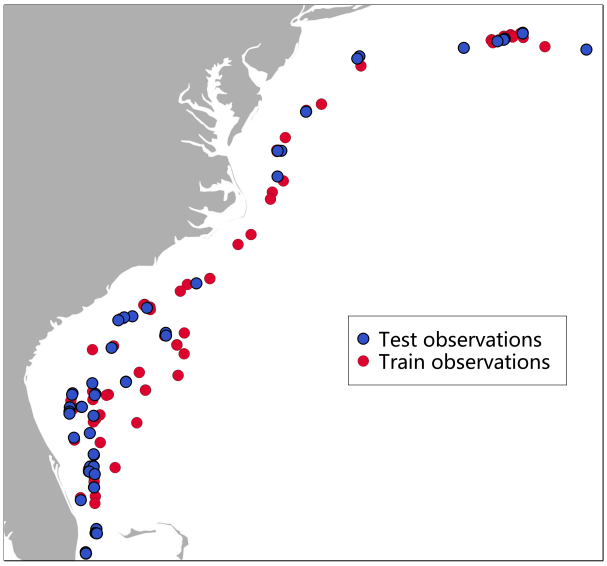
The *maxent* function from the *dismo* package in R was used to build the model. The input of the function is a binary presence-background variable, and the corresponding environmental variables of respective presence-background observations. The result is a model object that can be used to predict the habitat suitability of new locations where the environmental variables are present [20]. The default model parameters were used, as these parameters have performed well in past studies [5].

4.3.1 Model 1

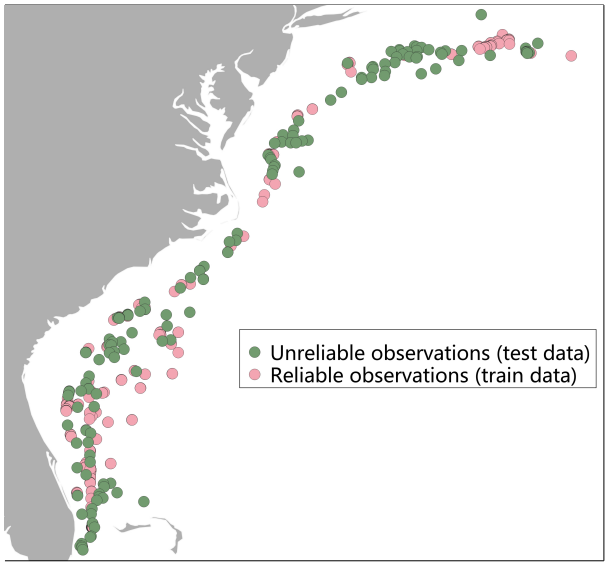
A maxent model was first trained and tested on reliable coral data. 80% of reliable presence observations were randomly sampled (without replacement) and used in the training dataset. The remaining 20% were used in the test set. The model was trained and tested for different ratios of background and presence observation, and the ratio that gave the best model performance (model evaluation metrics are described in section 4.4) on the test set was chosen for the final model. The ratio of background-to-presence observations giving the best model performance was 158.8 (background/presence). The same ratio of coral presence observations and background observations was used for the training and test sets. Therefore, the final training set contained 252 presence observations, and 40,014 background observations, while the final test set contained 63 presence observations, and 10,004 background observations. Figure 2a shows the spatial distribution of the presence observations in the training and test sets. The background points in the training and test sets were randomly sampled from the 2,294,379 background observations in the modeling dataset (the dataset is described in subsection 3.3).

4.3.2 Model 2

In the second model run, all reliable presence occurrences were used in the training set, and the unreliable presence occurrences were used in the test set. The same ratio of background-to-presence observations (158.8 background/presence) was used as in model 1 (see subsection 4.3.1). Therefore, the training set contained 315 presence observations (all reliable observations) and 50,018 background observations, while the test set contained 149 presence observations (all unreliable observations) and 23,659 background observations. Figure 2b shows the spatial distribution of the reliable presence observations in the test set, and the unreliable observations in the training set. The background points in the training and test sets were randomly sampled (without replacement) from the 2,294,379 background observation in the modeling dataset (the dataset is described in subsection 3.3).



(a) Distribution of the presence observations in the training and test sets for model 1.

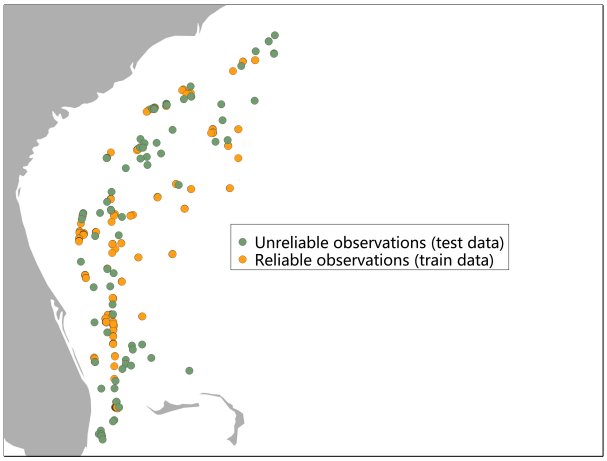


(b) Distribution of the reliable presence observations (training data), and the unreliable presence observations (test data) for model 2.

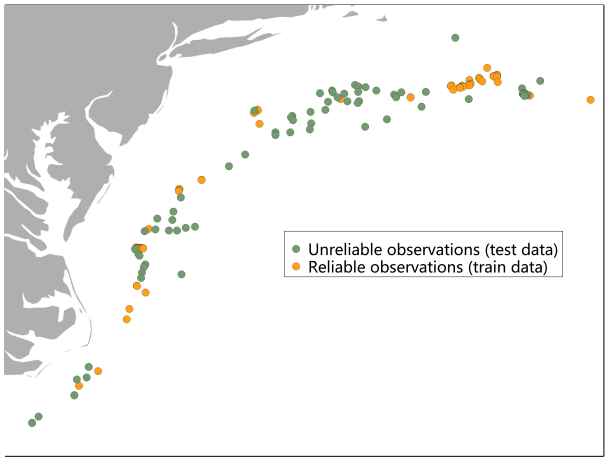
Figure 2: Distribution of the presence observations in the training and test sets of model 1 and model 2.

4.3.3 Model 3

The third model run consisted of two models. The first model was trained on reliable presence observations and tested on unreliable presence observations; all within the lower half of the study area. The second model was trained on reliable observations and tested on unreliable observations within the upper half of the study area. The same ratio of background-to-presence observations (158.8 background/presence) was used for both models. Therefore, for the lower half of the study area, the training set consisted of 37,000 background observations and 233 reliable presence observations, and the test set contained 12,704 background observations and 80 unreliable presence observations. The background observations were randomly sampled from the lower half of the study area. In the upper half of the study area, the training set contained 82 reliable presence observations, and 13,022 background observations. The test set contained 11,434 background observations, and 72 unreliable presence observations. The background observations were randomly sampled from the upper half of the study area. Figure 3a shows the distribution of unreliable (test data) and reliable (training data) presence observations in the lower half of the study area. Figure 3b shows unreliable (test data) and reliable observations (training data) in the upper half of the study area.



(a) Distribution of reliable and unreliable presence observations in the lower half of the study area.



(b) Distribution of reliable and unreliable presence observations in the upper half of the study area.

Figure 3: Distribution of reliable and unreliable presence observations in the lower and upper part of the study area respectively.

4.4 Model Validation

The accuracy of the models was assessed using a threshold-independent and threshold-dependent statistical metric. The AUC score of the models was used as a threshold-independent assessment. For the threshold-dependent assessment, the omission rate was used with the 10th percentile training presence threshold (P10). The P10 threshold is found by investigating the distribution of predicted probabilities of the presence observations in the training set. The 10th percentile of predicted probabilities of training presence records is used as a threshold [5, 21]. The P10 threshold was used to predict coral presence and absences in the test set. The proportion of presence records that were predicted as unsuitable (omission rate) was then calculated and used as a model assessment metric.

4.5 Variable Importance

The importance of predictor variables in the models was assessed using the standard variable importance metrics in the *maxent* function. The standard variable importance metrics in the maxent function are the *percent contribution* and *permutation importance* metrics. For a more comprehensive explanation of these variable importance metrics, see section 2.6 of the Literature Review.

5 Results

The modeling results are described in this section. The results from the habitat suitability predictions of the entire region is first documented, followed by the model validation metrics for model 1, model 2 and model 3. The importance of predictor variables in model 1 and 2 is also documented.

5.1 Habitat Suitability

Figure 4 shows the predicted habitat suitability of Scleractinia coral for the entire study region. All reliable presence occurrences (model 2) were used for prediction. The areas predicted to have high suitability are along the continental shelf of the United States East Coast; a sensible result based on prior observations and predicted habitats of Scleractinia corals in the region [2]. The areas predicted to have high habitat suitability are also more concentrated towards the southern regions of the study region.

Figure 5 and Figure 6 shows the location of the unreliable coral observations, and the predicted habitat suitability for the lower and upper half of the study region respectively. In the lower half of the study region, the unreliable coral observations seem to be located in areas with close proximity to areas predicted to have high habitat suitability. In the upper half of the study region, observations seem to be located further from areas predicted to have high habitat suitability in comparison to the lower half of the study region.

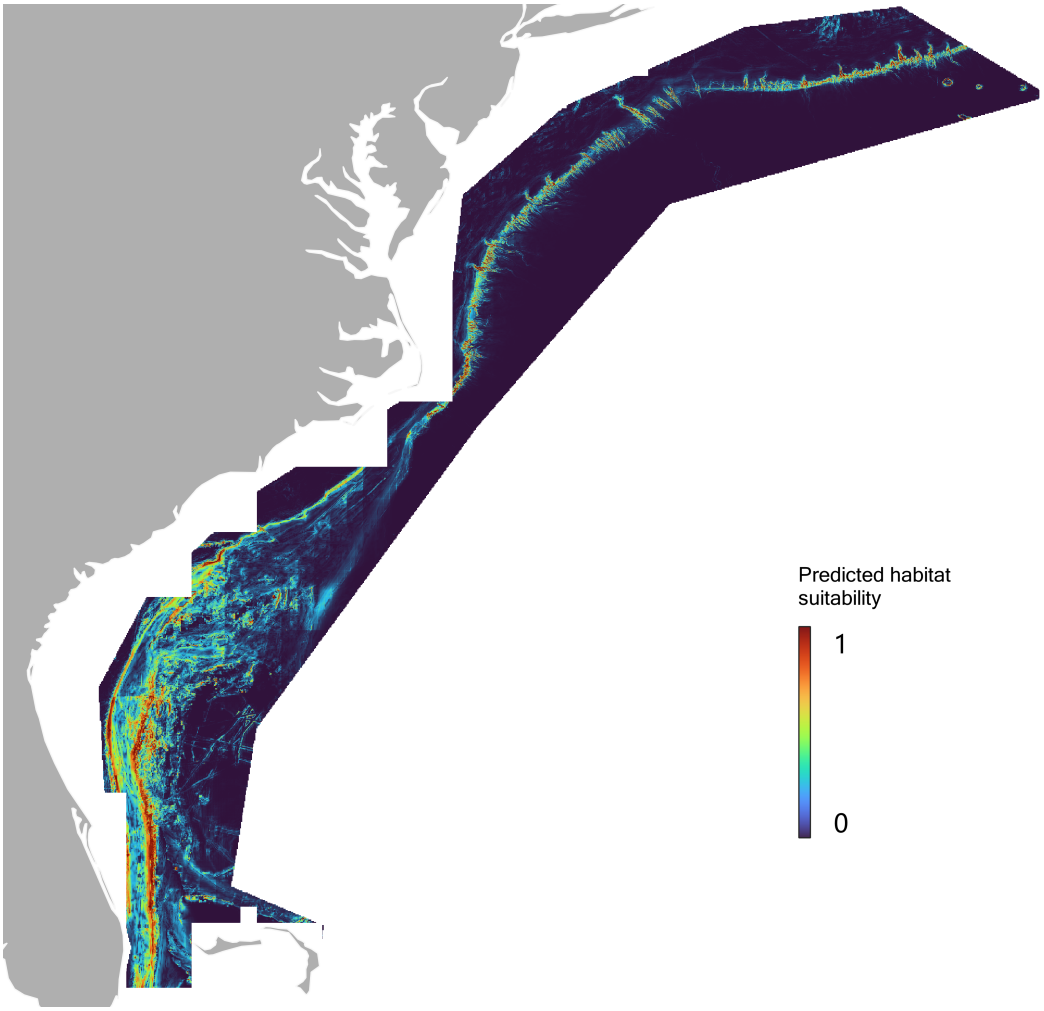


Figure 4: Map of predicted habitat suitability of the region.

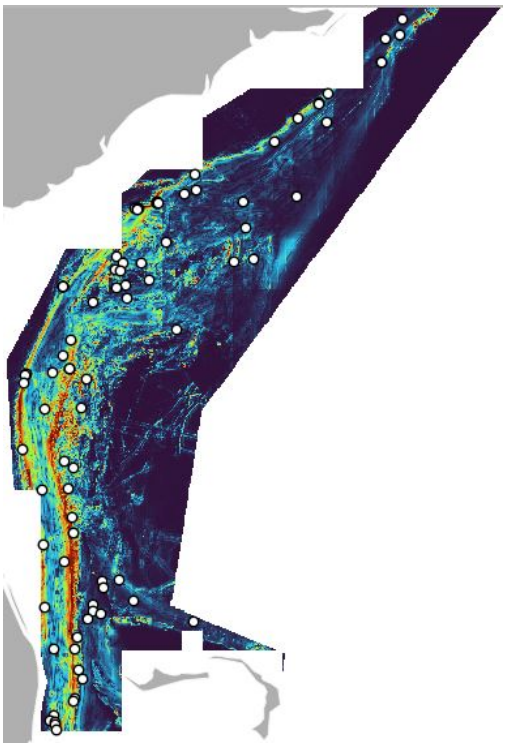


Figure 5: Predicted habitat suitability and unreliable coral occurrences (white dots) for the lower half of the study area.

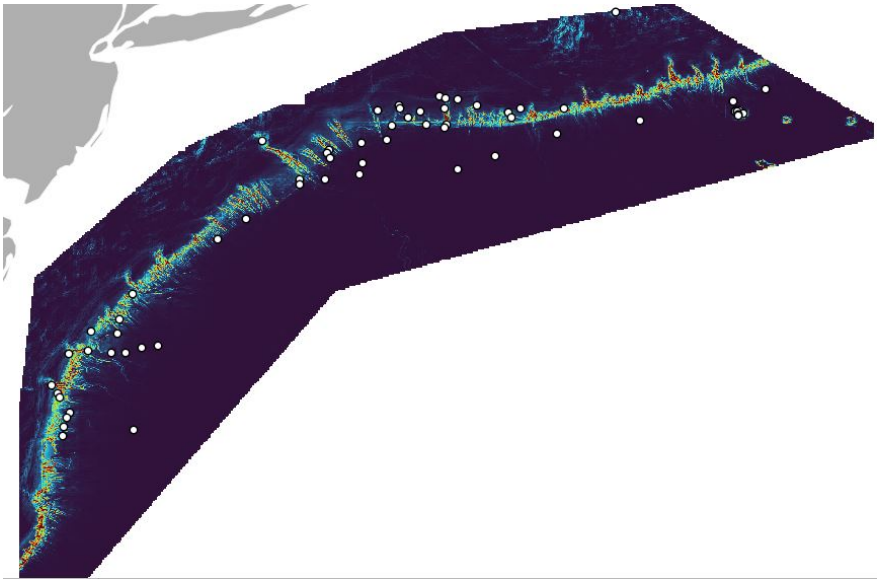


Figure 6: Predicted habitat suitability of the region, and unreliable coral occurrences (white dots) for the upper half of the study area.

5.2 Model Validation

The difference in model performance of model 1 on 2 on the test sets was larger than their difference on the training sets (see Table 4). Both models had nearly the same training AUC score, with model 2 having a 0.004 higher score than model 1. This small difference in AUC score between model 1 and model 2 gives indication of a higher model fit on the training set containing all reliable presence points (model 2 training set), than on the training set containing 80‘% of presence points (model 1 training set). The test AUC of model 1 was 0.196 higher than than the test AUC of model 2. Furthermore, model 2 had a significantly higher omission rate than model 1. Section 5.2.1 and section 5.2.2 summarises the validation metrics of model 1 and model 2 respectively.

Table 4: Validation metrics of model 1 and 2

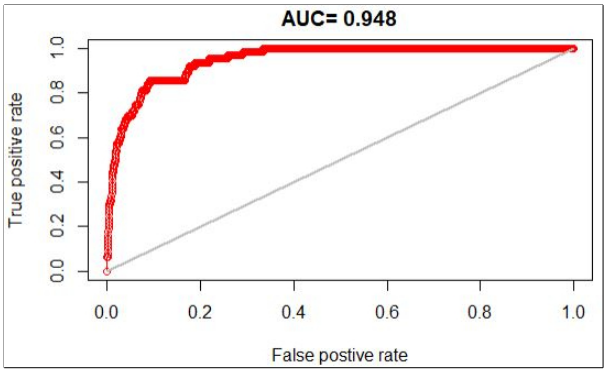
	Training AUC	Test AUC	10th percentile training presence threshold	Test omission rate
Model 1	0.953	0.948	0.196	0.076
Model 2	0.955	0.767	0.181	0.543

5.2.1 Model 1

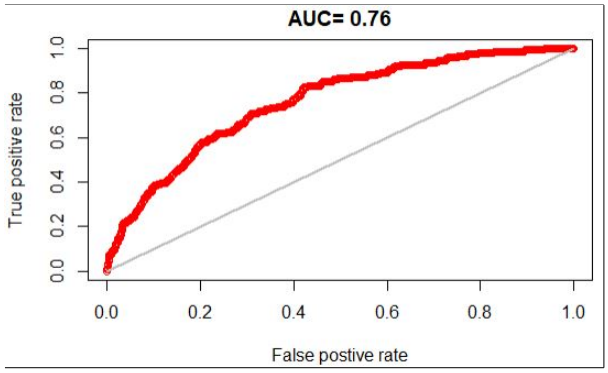
Model 1 had good performance scores across the metrics used to validate model output. The training and test AUC scores were 0.953 and 0.948 respectively, which are higher scores than random predictions (AUC = 0.5). Figure 7a shows the receiver operating characterstic (ROC) curve for the test set. The 10th percentile training presence threshold was 0.196, giving the threshold-dependent omission rate of 0.076. An omission rate of 0.076 of model output means that 7.9% of presence occurrences in the test set were wrongly predicted to have coral absence.

5.2.2 Model 2

Model 2 had lower performance scores on the test set relative to model 1. The training AUC for model 2 was high with the score of 0.955. The test AUC of model 2 was 0.767, which is higher than the score for random predictions (AUC = 0.5). Figure 7b shows the ROC curve for model prediction on the test set. The 10th percentile training presence threshold was 0.181, giving the threshold-dependent omission rate of 0.543. An omission rate of 0.543 means that 54.3% of presence occurrences in the test set were wrongly predicted to have coral absences.



(a) ROC curve for model 1 predictions on the test set



(b) ROC curve for model 2 predictions on the test set.

Figure 7: ROC curve and AUC score for predictions on the test for model 1 and model 2 respectively.

5.2.3 Model 3

As shown in Table 5, the model that was trained and tested on the upper half of the study area performed worse (test AUC = 0.693, test omission rate = 0.438) than the model that was trained and tested on the lower half of the study area (test AUC = 0.784, test omission rate = 0.889). 43.8% of unreliable presence observations were wrongly predicted to have absences in the lower half of the study area. In the upper half, 88.9% of unreliable presence observations were wrongly predicted to have coral absences. The model that was trained and tested on the lower half of the study area performed better than model 2 (test AUC = 0.767, test omission rate = 0.543), while the model that was trained and tested on the upper half of the study area performed worse than model 2.

Table 5: Validation metrics for the models that were trained (reliable coral observations) and tested (unreliable coral observations) separately on the lower and upper half of the study area.

Study area	Training AUC	Test AUC	10th percentile training presence threshold	Test omission rate
Lower half	0.935	0.784	0.253	0.438
Upper half	0.984	0.693	0.1998	0.889

5.3 Variable Importance

For model 1 and model 2, the variables with highest scores for *percent contribution* and *permutation importance* were the same. For both models, the following variables had highest *percent contribution* scores: slope (15 arc-sec), dissolved oxygen ($\mu mol/kg$), curvature (60 arc-sec), depth and temperature. These variables had highest contribution to the fitting of both models. The following four variables had significantly highest permutation importance: slope (15 arc-sec), curvature (60 arc-sec), depth and temperature. These variables had the highest importance for maintaining the training AUC score.

Although both models had the same variables of highest importance, the variable importance scores for each variable differed between the two models. Model 2 had higher scores than model 1 for slope (15 arc-sec) and temperature, and lower scores for dissolved oxygen, curvature (60 arc-sec) and depth. As explained in section 4.3, the difference between model 1 and model 2 was the amount of sampled reliable presence observations for the respective training sets. Model 1 was trained on 80% of reliable presence observations (randomly sampled without replacement), while model 2 was trained on all reliable presence observations. The change in variable importance scores between model 1 and model 2 is due to the different amount of reliable presence observations used in the model training of both models. Therefore, the increased importance of slope (15 arc-sec) and temperature, and the decreased importance of dissolved oxygen, curvature (60 arc-sec) and depth in model 2 (compared to model 1) is due to the additional 20% of reliable presence occurrences that were used for training.

Table 6: Variable importance in model 1

Variable	Percent contribution	Permutation importance
Slope (15 arc-sec)	28.1	32.6
Dissolved oxygen	23	4.4
Curvature (60 arc-sec)	17.3	16.4
Depth	13.1	22
Temperature	7.8	11.6
Aspect (60 arc-sec)	3.6	4.5
Chlorophyll-a	2.4	3.8
Aspect (200 arc-sec)	1.8	3
Aspect (15 arc-sec)	1.8	0.8
Curvature (200 arc-sec)	1.1	0.8

Table 7: Variable importance of model 2

Variable	Percent contribution	Permutation importance
Slope (15 arc-sec)	32.1	42.3
Dissolved oxygen	20.5	3.5
Temperature	13.1	16.5
Curvature (60 arc-sec)	12.7	6.5
Depth	12.3	19.4
Aspect (60 arc-sec)	3.5	5.4
Aspect (200 arc-sec)	2.5	3.2
Curvature (200 arc-sec)	1.5	0.8
Chlorophyll-a	1.3	1.7
Aspect (15arc-sec)	0.5	0.6

6 Discussion

This sections provides an interpretation of model results, and discusses general limitations concerning *maxent* modeling.

6.1 Interpretation of Model Results

Figure 4 shows the predicted habitat suitability of the study area. The southern half of the study area have a higher concentration of predicted suitable habitats compared to the northern half of the area. The high concentration of suitable habitats in the southern half of the area somewhat matches the modeling results of Davies & Guinotte. Within the the study area, the predictions of Davies & Guinotte gave an expanded area of habitat suitability in the southern half, and a thinned-in area of suitability on the northern half [5]. In general, the habitat suitability maps of Davies & Guinotte predicted wider areas of suitability than the results found in this research; a result that most likely is affected by the different resolutions of the predictions. Davies & Guinotte used a resolution of 30 arc-sec, while this research used a resolution of 15 arc-sec. Kinlan et al created a habitat suitability model for Scleractinia corals offshore the Northeastern USA with a resolution of approximately 15 arc-seconds. Their model predicted habitat suitability along the continental shelf a region corresponding more or less with the upper half of the study area in this research. Kinlan et al predicted that the suitable habitats followed the continental shelf; a result found in this research as well. However, the area of predicted suitable habitat found by Kinlan et al was wider than the area found in this research [2].

Model 1 was trained and tested on reliable presence observations, while model 2 was trained on reliable coral observations, and tested on unreliable observations. The results from model validation (see section 5.2) showed that model 1 performed better (AUC = 0.963) than model 2 (AUC = 0.767). This result demonstrates that predictor variables at reliable coral presence observations predict other reliable coral occurrences more accurately than unreliable coral occurrences. Therefore, the reliable and unreliable coral observations follow, to a certain extent, different spatial patterns with respect to environmental (predictor) variables.

Two apparent explanations may have caused the different spatial patterns of the reliable and unreliable coral observations shown in the model output. The first explanation (hypothesis 1.) is based on the assumption that the coral training data reflects well the spatial pattern of Scleractinia coral in the study area. Therefore, the model will give accurate predictions for suitable habitats of Scleractinia coral. If this is the case, one can estimate the reliability of unreliable-classified coral occurrences based on the model results.

The second possible explanation (hypothesis 2) for the different spatial pattern of reliable and unreliable coral occurrences, is based on the assumption that the reliable presence observations (aka. the training set) does not sufficiently reflect the spatial patterns of Scleractinia in the study area. Therefore, the reliable and unreliable observations have different spatial patterns. The model will perform well when trained and tested on reliable coral presences (model 1), as the model training and testing is isolated to the spatial patterns of the reliable coral occurrences. However, when the model is trained on reliable observations, and tested on unreliable observations (model

2), the model will overfit on the spatial patterns of the reliable observations, which leads to poor performance when the model is tested on the unreliable observations. If this explanation is true, the reliability of unreliable-classified observations cannot be inferred.

The model validation scores of model 3 (see section 5.2.3) gave poorer results than the model validation scores of model 1. This result indicates that reliable and unreliable observations had different spatial patterns; also within sub-regions of the study area. As a result, the second explanation (hypothesis 2) can be rejected; the differing spatial patterns between reliable and unreliable observations cannot be explained by the fact that model 2 only captures spatial patterns of reliable occurrences in a restricted region.

Based on the model validation metrics and habitat suitability explorations done in this research, the unreliable-classified observations are considered to be truly unreliable. Model validation metrics showed that models trained on reliable observations performed worse when tested on unreliable observations compared to reliable observations. This result also applied to model tests in subregions of the study area. These validation metrics confirmed that the reliable and unreliable observations had different spatial patterns.

6.2 Limitations

Habitat suitability models of deep-sea corals are affected by sample bias. Models created with presence data will only capture spatial patterns that exist within the coral training data. This means that coral predictions will be limited to areas with similar environments as where corals previously have been observed. Areas outside of the observed environments of deep-sea coral habitats will never be predicted to have suitability of deep-sea corals [2]. Future surveying of deep-sea corals should include areas with more uncertainty of deep-sea coral occurrences; to explore possible environmental habitats for deep-sea corals outside of known areas of occurrence and document areas with presence and absences. Having data on coral presence and absences in *grey areas* will allow for a more delicate discrimination between areas of coral presence and absence with a presence-absence model [2].

Some reliable coral presence observations had a higher sampling inaccuracy than the spatial resolution (15 arc-sec) of the model dataset. The sampling inaccuracy of the coral observations varied between 0 - 1000m, while 15 arc-sec varied approximately between 350-400 meters for the study region. Grid cells containing higher sampling inaccuracy than the length of a grid cell run the risk of giving grid cells with false coral presences, and therefore wrong environmental data for the presences of coral. However, environmental variables rarely change drastically within a length scale of 1000m, given the original length scale the environmental source data (see section 4.1 for further details of the original resolutions of the source data). Therefore, a neighboring grid cell to a coral presence observation might also capture coral spatial patterns. This fact is also shown in Figure 4, as grid cells predicted to have high habitat suitability often neighbor grid cells that also have high predicted suitability. The model that was trained and tested on reliable coral data (model 1) performed well, despite the sampling inaccuracies of some coral observations.

In the processing of coral data, only one coral observation was retained if multiple coral obser-

vations existed within the 15 arc-second grid cells. Multiple coral observations within grid cells contributes to weighing the model in favour of the environmental conditions in these cells [5]. The potential bias caused by high clustering of observations was tackled by retaining one observation from grid cells containing multiple coral observations. This filtering of coral observations reduced 41,703 coral observations to 464 observations; a result that highlighted the high spatial clustering of observations in the study region, and lack of dispersed sampling. Maxent models are capable of performing well despite few presence observations, as shown by Davies et al, who used 1,667 observations for a global habitat suitability model of Scleractinia coral, and Kinlan et al, who used 167 observations for a habitat suitability model of Scleractinia coral in Northeastern USA [2, 5]. However, maxent models tend to overpredict habitats of high suitability when few presence occurrences are used relative to other methods [2, 22, 23]. Based on the low number of coral occurrences that remained after data processing, availability of extensive data on deep-sea coral are needed for more accurate prediction models. This issue has already been recognized in past research [5]. Offshore the US East Coast, sampling efforts to obtain more extensive deep-sea coral data are currently in progress for the creation of the next generation of habitat suitability models [2].

The resolution of the bathymetry data was chosen as the resolution for the modeling dataset. The bathymetry data had the highest original resolution (15 arc-sec) of the predictor variables. It is common practice in habitat suitability modeling to use the same a resolution for modeling as the finest resolution of your predictor variables, in order to capture fine-scale spatial patterns [3]. Furthermore, deep-sea corals are distributed in areas with fine-scale topographic features, such as canyons and seamounts. It is valuable to capture such fine-scale topographical variation by using high-resolution bathymetry data [3, 5]. All predictor variables that did not concern seafloor terrain were originally data of lower resolution, and were interpolated to a resolution of 15 arc-sec to match the resolution of the bathymetry data for the final dataset. Cubic spline interpolation was chosen to increase the resolution of the predictor variables rather arbitrarily. The method caused the low resolution data to have smooth transitions in the high resolution grid. The values of the interpolated predictor variables were not controlled with empirical measurements, as was done in the research by Davies & Guinotte [5]. Therefore, the interpolation method may introduce errors.

The resolution of the predictor variables fail to capture environmental variation which occurs at a finer resolution than 15 arc-seconds. This effect is even stronger for the predictor variables that does not concern seafloor terrain, as they have source data at a lower resolution than 15 arc-seconds. These predictor variables will fail to capture strong environmental gradients, like e.g, areas of upwelling [5]. However, the limitation of too coarse resolution is assumed to be strongest for seafloor terrain data [2, 6]. Fine-scale variation in seafloor terrain, such as canyons and seamounts, are important features of suitable habitats for deep-sea corals, and exists at finer resolutions than 15 arc-seconds [3]. Furthermore, fine-scale variation in seafloor terrain function as effective proxies for seafloor substrata and fine-scale current conditions [6]. The common size at which these proxies exist is not mentioned in the review by Vierod et al [6].

7 Conclusion

Maxent models based on oceanography, bathymetry and Scleractinia presence data were successfully created offshore the US east coast. The model that was trained and tested on reliable presence

observations (model 1) performed well on the test set, making it a useful tool to predict Scleractinia occurrences based on the spatial patterns that are captured in the training set. A habitat suitability prediction was conducted for the entire region, and areas predicted to have high habitat suitability were located at the continental shelf along the US east coast. Areas with highest concentration of predicted suitability were located in southern areas of the study area; offshore the Florida coast. The models that were trained on reliable presence observations, and tested on unreliable observations performed worse on the test set in comparison to model 1. This result indicates that unreliable and reliable presence observations have different spatial patterns. With sub-regional model training and testing, the reliable presence observations were assumed to reflect the general spatial pattern of Scleractinia in the study area. As the reliable presence observations are considered to be reliable and reflect the spatial pattern of the region well, the unreliable observations located outside of areas of high habitat suitability can be considered as unreliable. The reliability classification scheme can be considered as sensible. For future work, an R code could be implemented that omits or retains unreliable presence observations based on predicted habitat suitability.

A Additional Tables

Type of data	Predictors	Rengstorf et al [12]	Davies et al [11]	Hu et al [3]	Davies and Guinotte [5]	Rooper et al [8]	Anderson et al [4]
Bathymetric	Depth	X	X	X	X		
	Slope	X	X	X	X	X	X
	Aspect		X	X		X	
	Rugosity	X		X	X		
	Roughness						
	Curvature (plane or profile)						
	Position index	X		X		X	
	Geomorphology zones			X			
	Substrate			X			
	Hydrocarbon seeps/pockmarks		X				
	Iceberg plough-mark area		X				
Chemical	Dissolved oxygen		X	X	X		X
	Salinity	X	X	X	X		X
	Aragonite saturation state		X	X	X		X
	Calcite saturation state				X		
	Sigma Theta						
	Nitrate		X	X	X		X
	Phosphate		X	X	X		X
	Silicate		X	X	X		
	Apparent oxygen utilization				X		
	Percent oxygen saturation				X		
	Alkalinity		X		X		
	Carbonate ion concentration				X		
	Dissolved inorganic carbon		X		X		
	pH				X		
Oceanographic	Temperature	X	X	X	X	X	X
	Current speed	X	X		X	X	
	Current direction	X					
	Vertical flow	X			X		
	Bottom stress	X					
	Maximum Tidal Current					X	
Biological	Surface chlorophyll a		X	X	X		
	Particulate organic carbon				X		X
	Primary productivity export				X	X	

Table 8: The predictor variables that were used in some of the habitat suitability studies that were investigated as part of the literature review.

Table 9: The datasets that were used to create teh global Scleractinia dataset.

Dataset name	Database	Citation
NOAA Global Deep-Sea Coral and Sponge Data	NOAA National Database for Deep-Sea Corals and Sponges	NOAA [24]
ICES Vulnerable Marine Ecosystems Data	ICES Vulnerable Marine Ecosystems	ICEES [25]
Cold Water Corals	OBIS	Rogers et al [26]
Deep-water azooxanthellate Scleractinia from Vanuatu, and Wallis and Futuna Islands	OBIS	Cairns, S.D [27]
Scleractinia distribution data from: Deep-sea fauna of European seas - an annotated species check-list of benthic invertebrates living deeper than 2000 m in the seas bordering Europe	OBIS	Keller, N.B [28]
Scleractinia specimens of Kuroshio Biological Research Foundation	OBIS	Imhara & Iwase [29]
Azooxanthellate Scleractinia Brazil	OBIS	Silveira & Rubens [30]
Cnidaria Anthozoa: Azooxanthellate Scleractinia from the Philippine and Indonesian regions	OBIS	Cairns & Zibrowius [31]
The Azooxanthellate Scleractinia (Coelenterata: Anthozoa) of Australia	OBIS	Cairns, S.D. [32]
Macroinvertebrate groups found on deep-sea volcanic habitats in the Galapagos Marine Reserve in the Eastern Tropical Pacific Ocean: Cnidaria - Anthozoa and Hydrozoa	OBIS	Buglass, S. [33]
Habitat-forming Cold Water Corals in the New Zealand region	OBIS	Tracey et al [34]

B Additional Figures

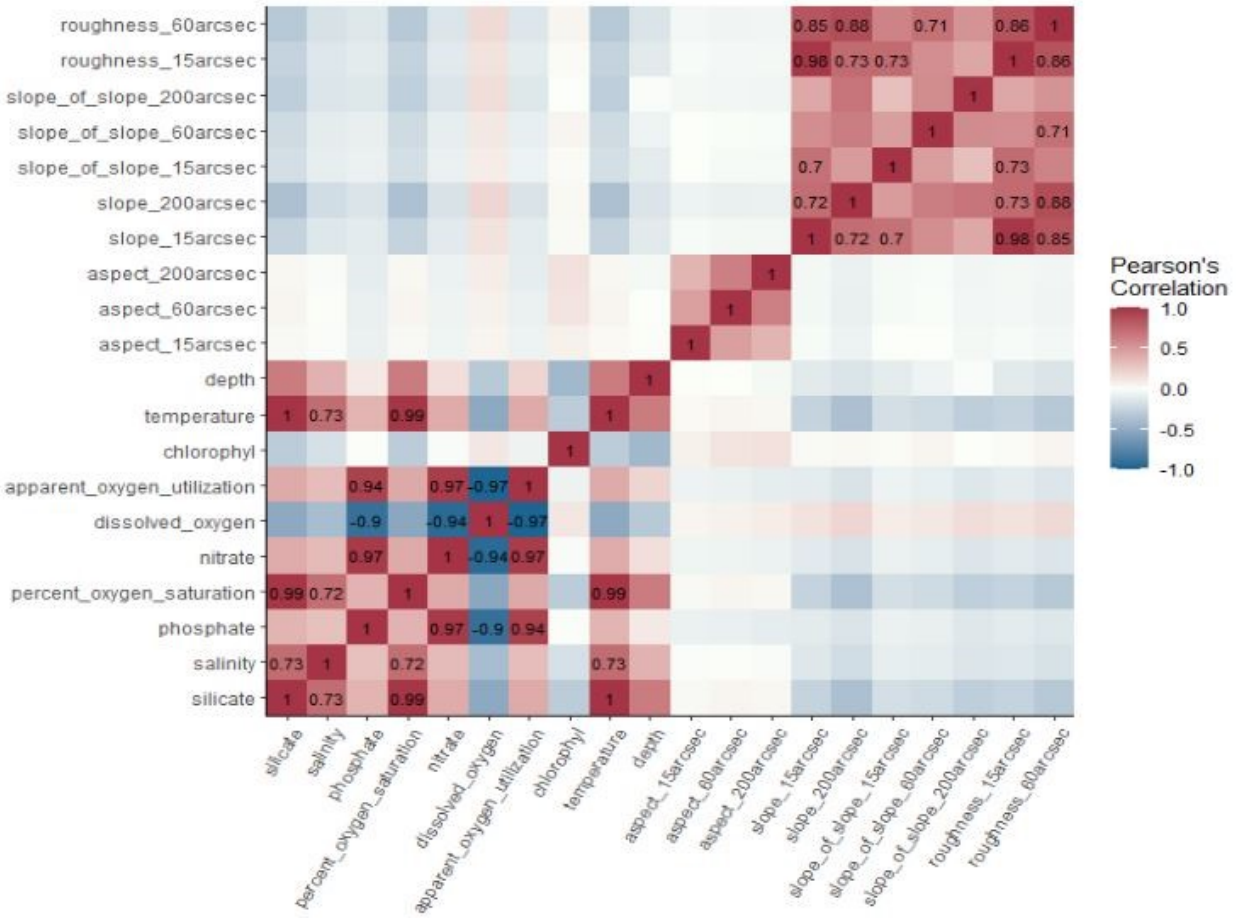


Figure 8: Pearson's correlation coefficient among predictor variables. Only correlation coefficients above the threshold of 0.7 is shown in the heatmap.

References

[1] "Coral reef ecosystems," Jul 2021. [Online]. Available: https://www.epa.gov/coral-reefs/basic-information-about-coral-reefs#:~:text=An%20estimated%2025%20percent%20of,point%20in%20their%20life%20cycle.

[2] B. P. Kinlan, M. Poti, A. F. Drohan, D. B. Packer, D. S. Dorfman, and M. S. Nizinski, "Predictive modeling of suitable habitat for deep-sea corals offshore the northeast united states," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 158, p. 103229, 2020.

[3] Z. Hu, J. Hu, H. Hu, and Y. Zhou, "Predictive habitat suitability modeling of deep-sea framework-forming scleractinian corals in the gulf of mexico," *Science of The Total Environment*, vol. 742, p. 140562, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720340845

[4] O. F. Anderson, J. M. Guinotte, A. A. Rowden, M. R. Clark, S. Mormede, A. J. Davies, and D. A. Bowden, "Field validation of habitat suitability models for vulnerable marine ecosystems in the south pacific ocean: Implications for the use of broad-scale models in fisheries management," *Ocean Coastal Management*, vol. 120, pp. 110–126, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0964569115300740

[5] A. J. Davies and J. M. Guinotte, "Global habitat suitability for framework-forming cold-water corals," *PLOS ONE*, vol. 6, no. 4, pp. 145–161, 2014. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018483

[6] A. D. Vierod, J. M. Guinotte, and A. J. Davies, "Predicting the distribution of vulnerable marine ecosystems in the deep sea using presence-background models," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 99, pp. 6–18, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967064513002403

[7] T. P. Boyer, H. E. Garcia, R. A. Locarnini, M. M. Zweng, A. V. Mishonov, J. R. Reagan, K. A. Weathers, O. K. Baranova, D. Seidov, and I. V. Smolyar, "World ocean atlas 2018," *NOAA National Centers for Environmental Information*, 2018. [Online]. Available: https://www.ncei.noaa.gov/archive/accession/NCEI-WOA18

[8] C. N. Rooper, M. Zimmermann, and M. M. Prescott, "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*, vol. 126, pp. 148–161, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967063717300407

[9] J. Elith, S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates, "A statistical explanation of maxent for ecologists," *Diversity and Distributions*, vol. 17, no. 1, p. 43–57, 2010.

[10] D. Woodby, D. Carlile, and Lee, "Predictive modeling of coral distribution in the central aleutian islands, usa," *Marine Ecological Progress Series*, vol. 397, pp. 227–240, 2009. [Online]. Available: https://www.int-res.com/articles/theme/m397p227.pdf

[11] A. J. Davies, M. Wisshak, J. C. Orr, and J. Murray Roberts, "Predicting suitable habitat for the cold-water coral lophelia pertusa (scleractinia)," *Deep Sea Research Part I:*

Oceanographic Research Papers, vol. 55, no. 8, pp. 1048–1062, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967063708000836

[12] A. M. Rengstorf, C. Mohn, C. Brown, M. S. Wisz, and A. J. Grehan, "Predicting the distribution of deep-sea vulnerable marine ecosystems using high-resolution data: Considerations and novel approaches," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 93, pp. 72–82, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967063714001368

[13] S. E. Georgian, W. Shedd, and E. E. Cordes, "High-resolution ecological niche modelling of the cold-water coral lophelia pertusa in the gulf of mexico," *Marine Ecology Progress Series*, vol. 506, pp. 145–161, 2014. [Online]. Available: https://www.int-res.com/abstracts/meps/v506/p145-161/

[14] D. A. Bowden, O. F. Anderson, A. A. Rowden, F. Stephenson, and M. R. Clark, "Assessing habitat suitability models for the deep sea: Is our ability to predict the distributions of seafloor fauna improving?" *Frontiers in Marine Science*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2021.632389

[15] T. Morato, J.-M. González-Irusta1, C. Dominguez-Carrió, C.-L. Wei, A. Davies, A. K. Sweetman, and G. H. Taranto, "Climate-induced changes in the suitable habitat of cold-water corals and commercially important deep-sea fishes in the north atlantic," *Global Change Biology*, pp. 2181–2202, 2020.

[16] S. Phillips, "A brief tutorial to maxent," *ATT Research*, 2006. [Online]. Available: http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc

[17] C. Hu, Z. Lee, and B. Franz, "Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference," *Journal of Geophysical Research: Oceans*, vol. 117, no. C1, 2012.

[18] GEBCO Bathymetric Compilation Group 2021, "Gebco 2021 grid," 2021. [Online]. Available: https://www.gebco.net/data_and_products/gridded_bathymetry_data/

[19] A. M. Barbosa, "Fuzzysim: Applying fuzzy logic to binary similarity indices in ecology," *Methods in Ecology and Evolution*, vol. 6, no. 7, p. 853–858, 2015.

[20] R. J. Hijmans, S. Phillips, J. Leathwick, and J. Elith, "dismo: Species distribution modeling," *R package version*, vol. 1, no. 4, pp. 1–1, 2017.

[21] C. Babich Morrow, "Thresholding species distribution models," Apr 2019. [Online]. Available: https://babichmorrowc.github.io/post/2019-04-12-sdm-threshold/

[22] R. G. Pearson, C. J. Raxworthy, M. Nakamura, and A. Townsend Peterson, "Original article: Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in madagascar," *Journal of Biogeography*, vol. 34, no. 1, p. 102–117, 2006.

[23] M. Papeş and P. Gaubert, "Modelling ecological niches from low numbers of occurrences: Assessment of the conservation status of poorly known viverrids (mammalia, carnivora) across two continents," *Diversity and Distributions*, vol. 13, no. 6, p. 890–902, 2007.

[24] NOAA Deep Sea Coral Research Technology Program, "Noaa national database for deep-sea corals and sponges (version 20220426-0)." [Online]. Available: https://deepseacoraldata.noaa.gov/

[25] International Council for the Exploration of the Sea, "Ices vulnerable marine ecosystems data portal." [Online]. Available: "https://vme.ices.dk/download.aspx"

[26] A. Rogers and J. Hall-Spencer, "Cold-water corals," *UNEP World Conservation Monitoring Centre (UNEP-WCMC)*, 2005.

[27] S. Cairns, "Cnidaria anthozoa: Deep-water azooxanthellate scleractinia from vanuatu, and wallis and futuna islands," *Southwestern OBIS, National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand, 1056 records*, 2015. [Online]. Available: http://nzobisipt.niwa.co.nz/resource.do?r=cairns1999corals

[28] N. Keller, "Scleractinia distribution data from: Deep-sea fauna of european seas - an annotated species check-list of benthic invertebrates living deeper than 2000 m in the seas bordering europe." [Online]. Available: http://ipt.vliz.be/eurobis/resource?r=deepsea_scleractinia

[29] M. Imahara and F. Iwase, "Scleractinia specimens of kuroshio biological research foundation," *National Museum of Nature and Science, Japan*, 2021. [Online]. Available: https://www.godac.jamstec.go.jp/ipt/resource?r=snet-1125-102

[30] F. L. da Silveira and R. M. L, "Wsa_obis_01_azooxanthellate scleractinia in brazil," *WSAOBIS, São Paulo*, 2008. [Online]. Available: http://ipt.iobis.org/wsaobis/resource?r=azooxanthellate_scleractinia_brazil_01

[31] S. Cairns and H. Zibrowius, "Cnidaria anthozoa: Azooxanthellate scleractinia from the philippine and indonesian regions," *Southwestern Pacific OBIS, National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand, 1396 records*, 2015. [Online]. Available: http://nzobisipt.niwa.co.nz/resource.do?r=cairns1997corals

[32] S. Cairns, "Data from: The azooxanthellate scleractinia (coelenterata: Anthozoa) of australia," *Southwestern Pacific OBIS, National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand, 1396 records*, 2004. [Online]. Available: http://nzobisipt.niwa.co.nz/resource.do?r=cairns2004corals

[33] S. Buglass, "Macroinvertebrate groups found on deep-sea volcanic habitats in the galapagos marine reserve in the eastern tropical pacific ocean: Cnidaria - anthozoa and hydrozoa. v1.2." *Charles Darwin Research Station*, 2019. [Online]. Available: http://ipt.iobis.org/obis-deepsea/resource?r=anthozoahydrozoa&v=1.2

[34] D. Tracey, A. Rowden, K. Mackay, , and T. Compton, "Habitat-forming cold-water corals in the new zealand region." *Southwestern Pacific OBIS, National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand*, 2014. [Online]. Available: http://nzobisipt.niwa.co.nz/resource.do?r=niwacoral