Title: The implementation of rare events logistic regression to predict the distribution of mesophotic hard corals across the main Hawaiian Islands

Author (corresponding): Lindsay Veazey; Hawai‘i Institute of Marine Biology, 46-007 Lilipuna Rd., Kane‘ohe, HI, 96744. lindsayv@hawaii.edu

Co-author: Erik C. Franklin; Hawai‘i Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawai‘i, 46-007 Lilipuna Rd., Kane‘ohe, HI, 96744. erik.franklin@hawaii.edu

Co-author: Christopher Kelley; The Hawai‘i Undersea Research Laboratory, 1000 Pope Rd., Marine Science Building 229, Honolulu, HI 96822. ckelley@hawaii.edu

Co-author: John Rooney; University of Hawai‘i, Joint Institute for Marine and Atmospheric Research, 1845 Wasp Blvd., Building 176, Honolulu, HI 96818. john.rooney@noaa.gov

Co-author: L. Neil Frazer; Department of Geology and Geophysics, School of Ocean and Earth Science and Technology, University of Hawai‘i at Mānoa, 1680 East West Road, Honolulu, HI 96822. neil@soest.hawaii.edu

Co-author: Robert J. Toonen; Hawai‘i Institute of Marine Biology, 46-007 Lilipuna Rd., Kane‘ohe, HI, 96744. toonen@hawaii.edu
Abstract

Predictive habitat suitability models are powerful tools for cost-effective, statistically robust assessment of the environmental drivers of species distributions. The aim of this study was to develop predictive habitat suitability models for two genera of scleractinian corals (*Leptoseris* and *Montipora*) found within the mesophotic zone across the main Hawaiian Islands. The mesophotic zone (30 – 180 m) is challenging to reach, and therefore historically understudied, because it falls between the maximum limit of SCUBA divers and the minimum typical working depth of submersible vehicles. Here, we implement a logistic regression with rare events corrections to account for the scarcity of presence observations within the dataset. These corrections reduced the coefficient error and improved overall prediction success (73.6% and 74.3%) for both original regression models. The final models included depth, rugosity, slope, mean current velocity, and wave height as the best environmental covariates for predicting the occurrence of the two genera in the mesophotic zone. Using an objectively selected theta ("presence") threshold, the predicted presence probability values (average of 0.051 for *Leptoseris* and 0.040 for *Montipora*) were translated to spatially explicit habitat suitability maps of the main Hawaiian Islands at 25 m grid cell resolution. Our maps are the first of their kind to use extant presence and absence data to examine the habitat preferences of these two dominant mesophotic coral genera across Hawai‘i.

Introduction

Consistent and pervasive deterioration of marine ecosystems worldwide highlights significant gaps in current management of ocean resources (Foley et al. 2010, Douvere 2008,
Crowder and Norse 2008). One such gap is the data required for informed marine spatial planning, a management approach that synthesizes information about the location, anthropogenic use, and value of ocean resources to achieve better management practices such as defining marine protected areas and implementing harvesting restrictions (Jackson et al. 2000, Larsen et al. 2004). The creation of spatial predictive models for improved marine planning is a relatively low-cost and non-invasive technique for projecting the effects of present-day human activities on the health and geographic distribution of marine ecosystems.

Defining and managing the biological and physical boundaries of ecosystems is a complicated but essential component of marine spatial planning (McLeod et al. 2005). The heterogeneous nature of ecological datasets can require the time-intensive development of problem-specific ecosystem models (Cramer et al. 2001, Tyedmers et al. 2005). Scientists frequently use straightforward, easy-to-implement regression methods to analyze complex datasets. The development of software accessible to relative novices has contributed to the growing popularity of regression methods (e.g., Lambert et al. 2005, Tomz et al. 2003).

Here, we employ a logistic regression with rare events corrections (King and Zeng 2001) to analyze the presence and absence data of two coral genera (Leptoseris and Montipora) and, thus, develop a predictive framework for the geographic mapping of mesophotic coral reef ecosystems (MCEs) across the main Hawaiian Islands. Mesophotic coral ecosystems, located at depths of 30–180 meters, are considered to be extensions of shallow reefs, because they harbor many of the same reef species present at shallower depths, and are also oases of endemism in their own right (Grigg 2006, Lesser et al. 2010, Kane et al. 2014, Hurley et al. 2016). MCE habitats are formed primarily by macroalgae, sponges, and hard corals tolerant of low light levels (Lesser et al. 2009). Corals of genus Montipora colonize primarily the shallow reef zone (< 30 m).
m), but some species, particularly Montipora capitata (Rooney et al. 2010), are able to extend their settlement into mesophotic depths. Corals of genus Leptoseris construct extremely efficient, light-capturing skeletons that facilitate their habitation of the lower mesophotic zone (Kahng et al. 2012) and are considered to be exclusively mesophotic dwellers (Kahng and Kelley 2007).

Ecological studies in the mesophotic zone are sharply limited in contrast to the shallower photic zone more accessible by open circuit SCUBA, but steady advances in diving, computing, and remotely operated vehicle technologies continue to facilitate interdisciplinary mesophotic research (Pyle 1996, Puglise et al. 2009). Mesophotic research in Hawai‘i has been conducted primarily in the ‘Au‘au Channel, Maui, a relatively shallow, semi-enclosed waterway between the islands of Maui and Lāna‘i that is among the most geographically sheltered and accessible areas in the Hawaiian Archipelago, and, as a result, much of the existing video and photo records of MCEs are from this area. This concentration of historic surveys highlights the importance of creating a pan-Hawai‘i predictive habitat model to identify likely areas of MCEs across unexplored areas of Hawai‘i’s mesophotic zone. Increasing our knowledge about the habitat preferences of the deep extensions of shallow coral species is critical given that approximately 40% of shallow (< 20 m) reef-building corals face a heightened extinction risk from the effects of climate change (Carpenter et al. 2008). Here, we model the habitat associations of mesophotic scleractinian corals because of both their intrinsic biological value as well as their potential to recolonize globally threatened shallow reef areas and serve as a refuge to mobile reef organisms (Bongaerts et al. 2010, Kahng et al. 2014).

Previous research about the environmental variables driving mesophotic scleractinian colonization in Hawai‘i suggests that distinct variation in community structure exists between the upper (30 – 50 m) and mid to lower mesophotic (50 – 180 m) depths (Rooney et al. 2010, Kahng et al. 2012).
et al. 2010, Kahng et al. 2014). Potentially influential environmental variables include photosynthetically active radiation (PAR) levels (Goreau and Goreau 1973, Fricke et al. 1987, Kahng and Kelley 2007, Kahng et al. 2010), isotherms (Grigg 1981, Kahng and Kelley 2007, Rooney et al. 2010), and hard substrate availability (Kahng and Kelley 2007, Costa et al. 2012). Rooney et al. (2010) noted that hard coral abundance declined dramatically below 100 m despite high (> 25%) availability of colonizable substrate; this sudden reduction in coral cover occurs at increasingly shallower depths across the northwestern Hawaiian Ridge and may be driven by the synchronously shallower occurrence of isotherms. Light and temperature intensity (Jokiel and Coles 1977, Rogers 1990), physical stress (e.g., wave energy or uncontrolled tourism) (Dollar 1982, Nyström et al. 2000, Franklin et al. 2013), and availability of colonizable substrate (Jokiel et al. 2004, Franklin et al. 2013) are known drivers of shallow (< 30 m) reef coral distributions across the world. We expect that our model will capture the influence of these abiotic variables on the distribution of mesophotic corals, especially in the shallower mesophotic zone. We speculate that our model may detect unexpected drivers of Leptoseris distribution, particularly because Leptoseris is known to colonize deeper depths that bear little resemblance to shallow reefs (Lesser et al. 2009, Rooney et al. 2010). Finally, previous predictive modeling research about the drivers of Hawaiian mesophotic coral colonization identified depth, distance from shore, euphotic depth, and sea surface temperature as potentially influential environmental variables (Costa et al. 2012, Costa et al. 2015). Our novel modeling approach utilizes all observational data (corals present and absent), which we believe will offer more insight into the dynamics that facilitate and inhibit coral colonization across the mesophotic zone.
Materials and methods

Organismal and environmental data

The Hawai‘i Undersea Research Laboratory (HURL) and the Pacific Islands Fisheries Science Center (PIFSC) provided video and photo records from MCEs in the Hawaiian Islands for our analyses. This imagery came from 19 dives conducted using submersibles, remotely operated vehicles (ROVs), autonomous underwater vehicles (AUVs), and tethered optical assessment devices (TOADs) in the ‘Au‘au Channel, Maui (13 dives) and two other geographically distinct regions: south O‘ahu (5 dives) and southeast Kaua‘i (1 dive). These dives were conducted between 2001 – 2013. We analyzed dive video using the Coral Point Count with Excel extensions (CPCe) tool (Kohler and Gill 2006) in combination with a modified PIFSC 2011 mapping protocol (PIBHMC 2015). PIFSC has used this type of combined analysis, referred to as the random five point overlay method (RFPOM), to process coral reef ecosystem benthic imagery throughout the U.S. Pacific Islands Region since August 2011, and our use of it ensures database consistency with regions processed prior to this study. The CPCe software placed five points randomly on each snapshot, which we then assessed for coral presence. If any of the five points was on coral, the observation was recorded as a “presence”. In an effort to evaluate the accuracy of RFPOM, we counted all corals in 200 randomly selected screen grabs and found that this method missed 2.4% of coral observations recorded in these images. We categorized corals by genus, as both Montipora (Forsman et al. 2010) and Leptoseris (Luck et al. 2013) contain species complexes that remain the subject of taxonomic uncertainty which prevent us from being able to reliably identify corals to the species level from photographs.
We recorded snapshots every 30 seconds for the duration of each dive video. In addition to an existing database of 40,193 records from dives in the ‘Au’au Channel, 3517 new snapshots were collected from the additional dives across south O‘ahu and Kaua‘i (Fig. 1). Of these 43,710 total images, 20,980 were discarded because either: 1) crucial observational data were absent, 2) they were redundant due to an extended stationary period, or 3) they fell outside the study depth range of 30 – 180 m. Of the remaining 22,714 records, we analyzed 2757 unprocessed images using the RFPOM (Table 1).

We selected our environmental covariates, listed in Table 2, based on the sufficiency of the data and the potential significance of each variable as an indicator of hard coral habitat suitability (e.g., Dolan et al. 2008; Rooney et al. 2010; Costa et al. 2012). We defined summer and winter seasons as May – September and October – April, respectively (Kay 1994, Rooney et.al. 2010). We delineated significant wave height estimates and mean current velocities by season and direction. We extracted and averaged significant wave height data from 144 days per season of twenty-four hour PacIOOS Simulating WAves Nearshore (SWAN) regional wave models estimated values for 2011–2015 (see website: http://oos.soest.hawaii.edu/las/). Mean current velocity values were available from 0:00 – 21:00 every three hours for all months from 2013-2015; for each season and direction, 48 mean current velocity values were extracted and averaged from the PacIOOS Regional Ocean Modeling System (see website: http://oos.soest.hawaii.edu/las/). This model has a 4 km horizontal resolution with 30 vertical levels across seafloor terrain. We sourced monthly MODIS Aqua Chlorophyll a averages for the year 2012 from the NOAA PIFSC OceanWatch Live Access Server (see website: http://oceanwatch.pifsc.noaa.gov/). Using the Morel (2007) method, we applied the following cubic polynomial equation to obtain logged euphotic depth:
\[
\log_{10} Z_{eu} = 1.524 - 0.436x - 0.0145x^2 + 0.0186x^3, \quad (1)
\]

where \( x \) represents the measured Chlorophyll \( a \) concentrations (mg/m\(^3\)) at sea surface. Euphotic depth is the depth at which the level of photosynthetically active radiation (PAR), a limiting factor for many heterotrophic mesophotic corals, is at 1% of surface PAR. In total, we used 14 environmental predictor variables to shape our model (Table 2) (Supplementary material, Figs. A1 – A5).

The spatial resolution of the bathymetry data was 50 m x 50 m for all islands. We resampled the bathymetry raster to a cell size of 25 m x 25 m consistent with a conservatively estimated \( \pm 25 \) m positioning error margin observed at a depth of \( \sim800 \) m. We estimated an average camera swath value of 3.24 m (range 2.45 – 4.54 m) based on previous measurements from 19 still image screenshots taken when the submersible was located at different heights above the seafloor. Our geopositional error for the images is \( \pm 5 \) m and we can expect that the location data are within a circle with a 10 m diameter. Our observation sampling area is projected out from the location area a distance of \( \leq 5 \) m. Addition of a conservative 5 m observation area buffer to the location error area produces an observational data position of \( \pm 20 \) m from the given coordinates of a data point.

We removed all subsampling within cells due to slight variations in camera angles or vessel speed through a point-to-raster conversion. We categorized all cells with \( \geq 1 \) presence observation as "present" cells and all cells with only absence observations as "absent" cells. This removal of multiple observations within the same 25 x 25 m pixel effectively eliminated pseudoreplication within the data. We used ArcToolbox and the Benthic Terrain Modeler Toolbox to calculate slope, curvature, rugosity, and aspect (compass direction) values (Wright et al. 2012). We performed a spatial join based on proximity to observation point data to assign
values for surface Chlorophyll a concentration, mean current velocities, distance from shore, and significant wave heights.

**Regression methods**

In describing the relationship between a response variable and one or more predictor variables, we use a logistic regression model because the response variable is dichotomous (Hosmer and Lemeshow 2004). The ordinary logistic regression (OLR) model is defined as:

$$\theta = \expit(\mu) = \frac{1}{1 + \exp(-\mu)}$$  \hspace{1cm} (2)

where $\theta$ is the probability that the species of interest is present ($y = 1$), and $1 - \theta$ is the probability it is absent ($y = 0$). The logit function is the inverse of the expit function, and

$$\text{logit}(\theta) = \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$  \hspace{1cm} (3)

is the linear sum of predictor variables, $x_1, x_2, \ldots, x_n$, with intercept $\beta_0$ and regression coefficients $\beta_1, \beta_2, \ldots, \beta_n$. In the language of generalized linear models (GLM), OLR is said to have the logit function as its link function and the expit function as its inverse link function.

Logistic regression provides a straightforward, meaningful interpretation of the relationship between a dichotomous dependent variable $y$ and a set of predictor variables (Allison 2001).

Despite the popularity of OLR, it may yield extremely biased results when an imbalance exists in the proportion of the response variable data (e.g., such as in our case, when $y = 0 >> y = 1$) (Van Den Eeckhaut et al. 2006). King and Zeng (2001) coined the term "rare events logistic regression" to describe their corrective methodology in dealing with unbalanced
1. The first step requires the selection of a representative sample. Though researchers generally prefer to work with more uniform response data (e.g., Liu et al. 2005), selection of an unusually high proportion of the rare event (in this case, \( y = 1 \)) to "balance" the dataset and increase \( \theta \) estimates will yield nonsensical results. We divided the data in half to create our training and testing datasets and checked that each set of observations had an approximately equal proportion (\( \bar{y} \)) of presence observations to better reflect the "true state" of the full dataset.

2. The second step rectifies any bias that might be introduced when dividing the dataset. This prior correction on the intercept (\( \beta_0 \)) can be calculated as:

\[
\hat{\beta}_0 = \beta_0^0 - \ln \left( \frac{1 - \tau}{\tau} \right) \left( \frac{\bar{y}}{1 - \bar{y}} \right) ;
\]  

(4)

here, \( \hat{\beta}_0 \) is the corrected intercept, \( \beta_0^0 \) is the uncorrected intercept, \( \tau \) is the true proportion of 1s in the population; and \( \bar{y} \) is the observed proportion of 1s in the training sample.

3. The third step rectifies any underestimation of the probabilities of the independent variables \( \beta_{i,...} \) from the substitution of the intercept value, obtained as:

\[
P(y_i = 1) = \theta^\prime C_i ,
\]  

(5)

where the correction factor \( C_i \) is given by:

\[
C_i = (0.5 - \theta^\prime \theta 0 - \theta^\prime X V (\theta^\prime X)^\prime ,
\]  

(6)

where \( X \) is a 1\(\times\)(\(n + 1\)) vector of values for each independent variable \( \beta_i \), \( X^\prime \) is the
transpose of $X$, and $\nabla(\beta)$ is the variance covariance matrix. We obtained the improved probability estimates through estimation of $\beta_i$ via $\hat{\beta_i}$, thereby considered "mostly" Bayesian (King and Zeng 2001). Our priors in this case would be uninformative, which means that we lack sufficient knowledge to estimate the probability distributions of our data and our parameter of interest, $\theta_i$. This is often the case when working with sparse ecological datasets. As the uninformative prior for a regression coefficient with domain $(, )$ is uniform, a full Bayesian estimation with uninformative priors is equivalent to a traditional logistic regression. Therefore, this correction is effectively a correction to the approximate Bayesian estimator, and its addition improves our regression by lowering the mean squared error of our estimates. We implemented this rare events logistic regression using the 'Zelig' package run in R (Imai et al. 2008, Choirat et al. 2015).

We constructed a correlation scatterplot matrix per coral genus to observe correlation levels between all variables. In choosing which highly correlated variables to exclude from the analyses, we followed the criteria outlined by Dancey and Reidy (2004) and Tabachnick and Fidell (1996), who suggest a cutoff correlation value of 0.7. Only mean significant wave height parsed by season consistently overreached this threshold; the covariate that was least correlated with the response variable was removed. We excluded predictors that lacked a clear distribution pattern or correlated minimally (< 0.05) with the response variable.

One of the more studied habitat preferences of *Leptoseris* and *Montipora* is the influence of depth on their distribution (Rooney et al. 2010, Costa et al. 2012, Kahng et al. 2010). Increasing depths often correlate with greater distance from shore. The inclusion of squared terms (e.g., $x_2 = x_1^2$) in our regression equation $\text{expit}(\theta) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ permits the
logistic curve to reflect the bell curve shape expected in plotting the distribution of these animals across a range of depths or distance from shore. In order to account for these trends, we added Depth \( \text{Squared} \) and Distance \( \text{Squared} \) as potential variables for consideration in our final model. As depth or distance increases, its square increases even more rapidly, allowing the squared term to eventually dominate and "pull down" the probability curve.

We withheld 50% of our information per genus as testing (i.e., validation) data. Using the remaining 50% (our training data), we performed the rare events corrected logistic regression described above. Using an exhaustive iterative algorithm (Calcagno and Mazancourt 2010), we modeled all possible combinations of included covariates. We ranked models using the corrected Akaike information criterion (AICc) (Hurvich and Tsai, 1989), which is considered an excellent comparative measurement of model strength, especially for sparse datasets. For both genera, the models with the lowest (lowest = best) AICc scores were lower than the "second best" AICc scores by at least 2 (i.e., \( \Delta \text{AICc} \geq 2 \)), indicating strong preference for the best model (e.g., Hayward et al., 2007).

In an ideal and unrealistic study, all biotic and abiotic components of a model would be homogenous and evenly distributed across a sampling space. Our sampling design includes overlapping submarine dive tracks and the inherent heterogeneity of the marine environment, which could problematically violate our model’s underlying assumption regarding the independence of our biological and environmental data. We removed all instances of pseudoreplication (multiple observations in one grid cell) when we assigned each grid cell to a category of “corals present” or “corals absent”. After we removed subsampling within our observational data, we checked for the presence of clustering, or spatial autocorrelation, within these data. Uncorrected spatial autocorrelation between observational data points confounds and
undermines any biological inferences drawn from model predictions.

We checked small-scale, local spatial autocorrelation using Geary’s C statistic (Geary 1954), based on the deviations in the responses of observation points with one another:

\[ C = \frac{n-1}{2S_0} \sum \sum w_{ij} (x_i - x_j)^2 \left/ \sum (x_i - \bar{x})^2 \right. \]

(7)

Here, \( x \) is the variable of interest, \( i \) and \( j \) are locations (where \( i \neq j \)), \( w_{ij} \) represents the components of the weight matrix, and \( S_0 \) is the sum of the components of the weight matrix.

Geary’s C ranges from 0 (maximal positive autocorrelation) to 2 for high negative autocorrelation. In the absence of autocorrelation, its expectation is 1 (Sokal and Oden 1978).

We also examined global spatial autocorrelation using Moran’s I statistic, which measures cross-products of deviations from the mean (Moran 1950):

\[ I = \frac{n}{S_0} \sum \sum w_{ij} (x_i - \bar{x})(x_j - \bar{x}) \left/ \sum (x_i - \bar{x})^2 \right. \]

(8)

Moran’s I values generally range from \(-1\) to 1, with 0 as the expectation when no spatial autocorrelation is present.

We also verified the spatial independence of our observational point data using a semivariogram, which is a graphical method of quantifying spatial correlation in a set of points (Figs. 2–3). We selected our theoretical semivariogram to fit the empirical semivariance using the ordinary least squares (OLS) method (Jian et al. 1996, Kendall et al. 2005). The spherical model had the best quantitative fit based on OLS estimates (Table 3). For each dataset, the low thresholds at which semivariance stopped increasing indicated the almost complete absence of spatial autocorrelation for each genus.
Model assessment

Evaluation of the rare events logistic regression model output is more complicated than for the typical linear model. For example, $R^2$ values, although calculated, have little applicability to logistic regressions and are therefore ignored (Menard 2000, Peng et al. 2002). Sample size and selected threshold largely influence the results of the Hosmer and Lemeshow goodness-of-fit test (Hosmer et al. 1997). Accordingly, we use model classification accuracy as a second measure of goodness-of-fit (in addition to $\Delta AIC_c$). We want to maximize true positives (TP) and true negatives (TN) while minimizing false positives (FP) and false negatives (FN). The sensitivity-specificity sum maximization approach (Cantor et al. 1999) therefore maximizes

$$SS_{\text{max}} = \frac{TP}{TP + FN} + \frac{TN}{TN + FP},$$

which is equivalent to finding the point on the ROC (receiver operating characteristics) curve at which the tangent slope is 1, indicating the optimal cutoff point at which "cost" (here, the number of FN and FP) and "benefit" (the number of TN and TP) is balanced. We chose this technique because we aim to identify regions devoid of hard corals as well as regions deemed potentially suitable for habitation.

ROC curves plot the true positive test rate against the false positive test rate across different theta cutoff points (Hadley and McNeil 1982). We calculated values for sensitivity and specificity for threshold increments of $0.005 \pm 1$ standard deviation of the rounded mean for each model. Because each theta threshold value varied based on the genus and model, the threshold independent area under the curve (AUC) test statistic best reflects the predictive accuracy of the
In addition to creating ROC curves, we also took into account the overall prediction success of each model, given as:

\[ OPS = \frac{TP + TN}{TP + TN + FP + FN}. \]  

Overall prediction success is a measure of total correct classification of both present and absent observations. While this is a good final assessment of model classification error, consideration of the prediction success alone is not a viable evaluation method when binary data is highly imbalanced, as a value given by this method may primarily represent model success in identifying the most common observation type (Fielding and Bell 1997). We plotted our sensitivity and specificity values on a ROC curve to show how each model performed relative to chance (Fig. 4). All models fall in the range \( 0.7 < \text{AUC} < 0.9 \), which indicates good discrimination and reliability of model predictions (Hosmer and Lemeshow 2004).

We also created maps of individual and summed predicted occurrence probabilities of both coral genera across the main Hawaiian Islands and ran a hotspot analysis using the ArcGIS Getis-Ord \( G^*_i \) Hotspot Analysis tool. We constructed a polygon fishnet composed of 1 x 1 km cells which encompassed all islands. We summed each 25 x 25 m raster cell value for probability of \( \text{Leptoseris} \) occurrence and probability of \( \text{Montipora} \) occurrence. We performed a spatial join of raster cell values within each polygon for an average value of summed probabilities. The Getis-Ord \( G^*_i \) statistic identifies clusters within these polygons that display values higher in magnitude than random chance would permit. The Getis-Ord local statistic is given as:
Here, \( w_{i,j} \) represents the spatial weights between features \( i \) and \( j \); \( n \) represents the total number of features; \( x_j \) is the attribute value for feature \( j \); \( X = \frac{1}{n} \sum_{j=1}^{n} x_j \); and \( S = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} x_j^2 - (X)^2} \).

### Results

Geary’s C test statistic is a measure of local (small–scale) spatial autocorrelation; in the absence of correlation, 1 is the expected value of Geary’s C. Moran’s I is a measure of global (large–scale) spatial autocorrelation; in the absence of correlation, a value of 0 is expected for the Moran’s I test statistic. For our Leptoseris dataset, Geary’s C = 0.990; for our Montipora dataset, Geary’s C = 0.996. For our Leptoseris dataset, Moran’s I = 0.006; for our Montipora dataset, Moran’s I = 0.003. These values indicate there was no local clustering or global spatial autocorrelation within either dataset. We observed negligible levels of autocorrelation up to ~100 m for Montipora (Fig. 3). By ensuring that spatial autocorrelation is not present in our data, we do not violate the assumption that our response data are independently observed, which enables us to draw robust conclusions about the ecological factors influencing the distribution of these coral genera within the mesophotic zone across the main Hawaiian Islands.

The OLR covariate coefficients were modified using the rare events corrections proposed by King and Zeng (2001), resulting in a change in predictive power (Table 4). Rare events corrected models usually performed better than the uncorrected models, in terms of improved
specificity and prediction success. Our sensitivity values for both corrected models were slightly lower than the corresponding OLR sensitivities, but in each case, specificity and prediction success values were improved. Additionally, standard errors of the coefficient estimates were lower for corrected models than for uncorrected models (Supplementary material, Tables A1–A4).

Leptoseris corals inhabit mesophotic regions with high slope and rugosity values, high to moderate perennial current flow, and their occurrence peaks around 100 m (Supplementary material, Table A3, Figs. A6 – A10). Montipora corals peak in occurrence around 60 m and colonize regions less exposed to high energy winter swells (Supplementary material, Table A4, Figs. A11 – A12). Predicted presence probability values (θ) averaged 0.051 for Leptoseris and 0.040 for Montipora models in the validation data (Figs. 5 – 6). These values agree well with the actual presence frequencies in that data (0.052, 0.042). To better interpret these realistically low theta values, we chose a theta threshold to transform the probability estimates to presence/absence values. This is standard practice when examining the results of a rare events logistic regression, but less common when performing OLR (Liu et al. 2005, Bai et al. 2011).

Objective selection of a theta threshold on a per–model basis is more scientifically sound than, for example, an arbitrary assignment of 0.5 (Cramer 2003). The transformed model is valid if a threshold value yields a high percentage of correctly classified observations and a low number of FP and FN observations (Gobin et al. 2001). We selected an appropriate threshold for each model (Table 4) in order to maximize $SS_{\text{max}}$ (Liu et al. 2005).

Our final hotspot maps show the results of our analysis for Leptoseris, Montipora, and both genera combined across all islands (Figs. 7 – 9). We show hotspots of habitat suitability for both coral genera in red for areas of highest suitability and blue for areas of lowest suitability.
We identify a cell as a hotspot when the sum of its value and the values of its nearest neighbors is much higher or lower than the mean over all cells. When the local sum of a cluster is very different from the expected value, a statistically significant hotspot is identified ($G_i^*$ statistic $\geq 1.96$ or $G_i^*$ statistic $\leq -1.96$). Neither genus clearly dominated the summed probabilities hotspot identification across any of the islands. Large *Leptoseris* hotspots were identified in southwest Moloka‘i, northeast O‘ahu, west Hawai‘i, and the central ‘Au‘au Channel. *Montipora* hotspots were identified in east Ni‘ihau, southwest Kaua‘i, west and south O‘ahu, west Hawai‘i, and the central ‘Au‘au Channel.

**Discussion**

In this study, we used logistic regression with rare events corrections to predict the habitat preferences of two dominant scleractinian coral genera across the entire mesophotic zone surrounding the main Hawaiian Islands. The habitat preferences of *Montipora* in the mesophotic zone appear distinct from those of *Leptoseris*. *Montipora* prefers the middle mesophotic zone (50–80 m) of reefs less exposed to high-energy winter swells. *Leptoseris* prefers steep, rugose slopes and the lower mesophotic zone (> 80 m) in regions of high year-round current flow.

**Important environmental covariates**

Predicted *Montipora* presence peaks at about 60 meters (median occurrence probability = 7.5%); *Leptoseris* presence peaks at about 100 meters (median occurrence probability = 7.5%). These predictions are consistent with the inferences of Rooney et al. (2010), which separates...
mesophotic reefs into three distinct depth sections: upper (30–50 m), branching/plate dominated (50–80 m), and *Leptoseris* dominated (>80 m). The depth at which suitability peaks for *Leptoseris* occurs at a range where steep ridges and drop-offs are plentiful in our study region, and therefore the mean preferred depth may be prone to slight overestimation.

In addition to depth, four environmental covariates appeared to influence the distribution of *Leptoseris*: rugosity, slope, summer mean current velocity (northward), and winter mean current velocity (eastward). Scleractinians easily colonize environments that are relatively calm and rugose due to the larger amount of available surface area, and this positive correlation was reflected in our model. *Leptoseris* habitat preference was also positively associated with slope, which was not observed for *Montipora*. Corals that inhabit the upper mesophotic zone may be more susceptible to damage from debris displaced by high wave energy, and are therefore less likely to colonize steep slopes (e.g., Harmelin-Vivien and Laboute 1986, Bridge and Guinotte 2013). The deeper distribution of *Leptoseris* may protect it from damage related to wave intensity, allowing it to colonize slopes (e.g., White et al. 2013). Another possibility is that the model is picking up drop-offs from masses accreted during the last glacial maximum. These steep drop-offs are present between 90–120 m in the *Leptoseris*-dominated lower mesophotic zone (Yokoyama et al. 2001, Webster et al. 2004).

*Leptoseris* also favors well-flushed areas exposed to year-round moderate current flow (i.e., up to 0.3 m/s). The plate-like morphology of *Leptoseris* corals effectively boosts sunlight capture by its symbiotic zooxanthellae and zooplankton capture by the corals themselves, but it also makes the coral vulnerable to smothering by sediment accumulation (Bak et al. 2005, Bongaerts et al. 2010, Marcellino et al. 2013). The success of *Leptoseris* corals in areas of moderate current flow may be related to the improbability of sediment settlement and
accumulation. While the model did not capture the same effect of current flow on *Montipora* distribution, we recognize that the morphology of some *Montipora* species is extremely similar to that of *Leptoseris*. We do not expect either genus to readily colonize highly turbid regions, especially given that certain species of heterotrophic *Montipora* are thought to exploit strong currents to meet their energy requirements (Grottoli et al. 2006, Rooney et al. 2010).

Substrate hardness, a variable known to influence coral colonization, was notably absent from each model. Substrate hardness values were derived from acoustic backscatter imagery readings. The base resolution of these readings (50 m x 50 m) was not sufficiently detailed for purposes of this analysis. We noted plentiful coral colonization along larger surfaces like lava fingers, the hardness of which would be detectable by backscatter surveys, as well as across small rock fragments strewn across a sand flat, which would be obscured by the softness of the surrounding benthos. We can conclude that measurements of benthic hardness are not detailed enough for predictive modeling purposes at a 25 m x 25 m resolution.

We emphasize that the purpose of this study was to build a pan–Hawai‘i predictive habitat map for two dominant coral genera *within the mesophotic zone*. Because the scope of this study included all main Hawaiian Islands, we were constrained by the coarseness of available full–coverage environmental data. As we build on this analysis, we plan to use our maps to identify areas of interest for further study at higher resolution and to include additional variables currently only available in certain regions, such as light intensity and temperature at depth. For example, our predictive and hotspot maps identify Penguin Bank (southwest Moloka‘i) as particularly suitable for *Leptoseris* colonization, which has not been verified by video or photo records. High resolution backscatter data (1 m x 1 m) exist for this region, and incorporation of these data into new analyses of subsets of our study area may refine our conclusions.
Error sources and model reliability

We examined two types of error (false negatives and false positives) and analyzed our models without giving preference to either one. This approach is widely accepted as the best method of overall error minimization (e.g., Liu et al. 2005, Fielding and Bell 1997). Rare events corrected models for both *Leptoseris* and *Montipora* achieved levels of specificity and sensitivity well above the null, indicating good predictive power. Additionally, both models attained about 74% overall prediction success. We assumed coral detectability was constant across the study region and that we can therefore consider the true absence observations to be reliable indicators of a potentially unsuitable habitat for corals. For each genus, the model tended to slightly overpredict presence observations; large numbers of false positives lowered sensitivity values. This is inevitable in the analysis of severely imbalanced or sparse binary data; the ongoing addition of presence observations to the dataset will improve overall model classification accuracy.

While the consistent identification of southern coastal areas as suitable is reliable, the comparatively infrequent selection of northern coasts is likely due to the source of the model building observations. The vast majority of mesophotic exploration has been along southern coastlines, which is often where waters are calmest in Hawai`i. It is speculated that because mesophotic corals are more shielded from winter long-period wave energy than their shallow water counterparts, they are able to flourish at depth along northern coastlines (Grigg 1998, Rooney et al. 2010). The addition of data sourced from northern expeditions would likely improve predictive power of the model across north-facing coastlines (Alin 2010).

We acknowledge that the original data were not collected in a standardized fashion (e.g.,...
variation in vessel traveling speed or differences in data collection vessel and/or quality). Our careful exclusion of overlapping observation points within each 25 x 25 m rectified this sampling design flaw as much as possible and eliminated pseudoreplication.

Distinctions between coral genera

Our *Montipora* model was simpler than the *Leptoseris* model in that the only variable included other than depth was winter significant wave height. Though uncertainty was highest at lower values of significant wave height, *Montipora* demonstrated a preference in colonizing habitats that experience lower significant wave height during winter. This preference contrasts with *Montipora* species in shallow waters that were more likely to be observed in higher wave height environments (Franklin et al. 2013). This likely influenced the inability of the model to identify any suitable habitat around Ni’ihau, where the average winter significant wave height equaled 1.78 meters, almost double the mean significant wave height of our model training data (0.91 m). Though mesophotic corals are generally thought to be exempt from the growth limitations faced by shallow water corals in regions of high wave energy, prolonged wave intensity has been shown to negatively affect the colonization of upper mesophotic scleractinians, especially in sloping areas prone to debris avalanches (Bridge and Guinotte 2013, Kahng et al. 2014). Continuation of this work might include a more in-depth examination of the relationship of this coral genus with the combined effects of slope of available substrate and exposure to wave energy.

We found no records of *Montipora* presence when processing our O‘ahu dataset, which probably contributed to the very low predicted mean probability of *Montipora* occurrence there.
We believe this is due in part to the sampling pattern across south Oʻahu; we recorded 62.3% of all observations processed for this region at a depth of 75 m or greater. *Montipora* prevalence is greater in the upper to middle mesophotic zone, and the relative deepness of the Oʻahu dives likely influenced their nonappearance in this portion of the dataset. We emphasize that the dearth of *Montipora* observations around Oʻahu is an artifact of the dataset we used to construct our model; *Montipora* corals have been observed in mesophotic depths across Oʻahu (e.g., Fig. 4b, Rooney et al. 2010). The mean significant wave height across the mesophotic zone was lower across the southern and western coasts (1.50 m) than that observed across the northern and eastern coasts (2.37 m) of the island. As at Niʻihau, we assume that this high northern and eastern average height, coupled with the absence of *Montipora* presences in Oʻahu in the training dataset, greatly impacted our model’s ability to detect areas of suitable habitat around the island. The results of our Getis*–*Ord *Gi* * Hotspot Analysis corroborate the findings of Costa et al. (2015), who used Maximum Entropy software to predict the highest occurrence probability of *Leptoseris* and *Montipora* in the middle and mid-coastal ʻAuʻau Channel, respectively (Costa et al. 2015).

The factors influencing the distribution of coral species in shallow and mesophotic habitats differ. One of the fundamental drivers of the occurrence and abundance of coral species on shallow reefs in Hawaiian waters is wave stress (Dollar 1982, Grigg 1983, Franklin et al. 2013). Given the depth range of MCEs, wave stress is unlikely to serve as a direct influence on coral occurrence but may provide secondary effects as wave events lead to debris reaching MCEs (Kahng 2014). Furthermore, the decoupled effects of environmental drivers on shallow and mesophotic zones extend between the islands. In shallow reef communities *Montipora* species become relatively more dominant from Hawaii Island to Niʻihau (Franklin et al. 2013), but appear to peak in occurrence in the mesophotic zone of Maui Nui. While strong
environmental drivers influence the distributions of shallow corals, the occurrence patterns of mesophotic corals may reflect a more stable environment with an increased influence of biotic factors such as interspecific competition in a habitat zone with limited light and space resources available.

Conclusions

We implemented a rare events corrected logistic regression to determine the most influential environmental predictors of Montipora and Leptoseris colonization in the mesophotic zone. Habitat preference differences between these genera appear distinct and multi-faceted. Montipora thrives in the middle mesophotic zone in areas sheltered from high intensity winter swells, while Leptoseris tends to colonize steep, rugose, well-flushed areas in the lower mesophotic zone. Improved understanding of the distribution of mesophotic corals will enable resource managers to propose the construction of seafloor power cables and other offshore infrastructure in areas less likely to contain coral communities. Results will likewise facilitate efforts to protect these communities by supplementing scientific dive planning and strategies for conservation, such as marine spatial planning.

Acknowledgments

Authors thank the many staff members of HURL, PIFSC, and SOEST who helped acquire and process our data. L.V. thanks Dr. David Hondula (Researcher at Arizona State University) for his dedicated GIS instruction. We dedicate this manuscript to the memory of our coauthor, Dr. John

References


26


DOI:10.1016/j.marpol.2008.03.012.


Menard, S. 2000. Coefficients for Determination in Multiple Logistic Regression Analysis. *The*


White, K.N., Ohara, T., Fujii, T., Kawamura, I., Mizuyama, M., Montenegro, J., Shikiba, H.,
Naruse, T., McClelland, T.Y., Denis, V. and Reimer, J.D. 2013. Typhoon damage on a shallow

Wright, D.J., Pendleton, M., Boulware, J., Walbridge, S., Gerlt, B., Eslinger, D., Sampson, D.
and Huntley, E. 2012. ArcGIS Benthic Terrain Modeler (BTM), Environmental Systems
Research Institute, NOAA Coastal Services Center, Massachusetts Office of Coastal Zone
Management.

Yokoyama, Y., Purcell, A., Lambeck, K., and Johnston, P. 2001. Shoreline reconstruction
around Australia during the last glacial maximum and late glacial stage. *Quaternary
International* 83: 9–18.