國立海洋生物博物館
NATIONAL MUSEUM OF MARINE BIOLOGY & AQUARIUM
Platax 19: 01-24, 2022
doi: 10.29926/platax.202212_18.0000

Predicting the abundance of corals from simple environmental predictors with a machine-learning approach

Anderson B. Mayfield[1-2], Alexandra C. Dempsey[3], Chii-Shiarng Chen[4-6]

[1]Coral Reef Diagnostics, Miami, FL, USA

[2]International Society for Reef Studies, Tavernier, FL, USA

[3]Khaled bin Sultan Living Oceans Foundation, Annapolis, MD, USA

[4]National Museum of Marine Biology and Aquarium, Checheng, Pingtung, Taiwan

[5]National Dong-Hwa University, Checheng, Pingtung, Taiwan

[6]National Sun Yat-Sen University, Kaohsiung, Taiwan

*Corresponding author. Email: anderson@coralreefdiagnostics.com.

Abstract

As Earth's oceans continue to warm at alarming rates, scientists have ramped up efforts to learn more about arguably the planet's most thermo-sensitive ecosystems: coral reefs. However, despite covering only a small areal fraction of the ocean, well under 1% of coral reefs have been surveyed, and many have likely never even been *seen*. For this reason, the Khaled bin Sultan Living Oceans Foundation (LOF) embarked on their "*Global Reef Expedition*" (GRE) from 2012 through 2016, characterizing thousands of never-before-studied reefs in all major coral reef areas and across a biological gradient that spanned molecules to ocean basins. We sought to leverage this rich dataset herein to identify areas of high coral cover that have *not* previously been surveyed, as this capacity could 1) aid in triaging conservation efforts and 2) reduce field time spent searching for "needle-in-a-haystack" reefs with exceptionally high coral abundance. We trained over 3,000 predictive models with various combinations of common environmental parameters (e.g., temperature, type of reef, etc.) from the LOF-GRE Solomon Islands dataset as a proof-of-concept, and one neural network featuring only nine of these predictors was associated with a validation R^2 of 0.81. Although additional environmental and demographic predictors could be incorporated to attempt to more robustly predict the coral cover of unexplored reefs, this confidence is high enough to where managers and scientists could use the underlying model to predict where else in this Coral Triangle nation they are likely to find reefs with high abundance of live corals.

Key words: coral reefs, global change biology, machine-learning, marine ecology, Solomon Islands

Introduction

The rate of seawater temperature rise is now so rapid that we are unlikely to survey all extant coral reefs prior to their biology having been fundamentally altered by global-scale climate shifts stemming from humankind's immense carbon footprint (Hughes et al., 2017). This is not to say that all coral reefs will cease to exist in the coming decades, though whether their biodiversity, aesthetic appeal, and capacity to provide myriad ecosystem services to both other marine fauna and humans (e.g., coastal protection) remains to be determined (Mayfield & Gates, 2007; Reverter et al., 2022). Very little of the ocean has been studied to any great extent, and even though coral reefs occupy only a small proportion of the global ocean area (~0.1%), few coral reefs can be said to be well studied at the present time (typically limited to those abutting well-funded marine laboratories; e.g., Mayfield et al. [2012]). The world's most beautiful (Fig. 1) and high-biodiversity coral reefs are found in the "Coral Triangle" (Clifton et al., 2013), an arbitrarily defined region extending south from Taiwan into the Philippines, Indonesia, Papua New Guinea, the Solomon Islands, and, depending on the cartographer, farther beyond. With the exception of Taiwan, no Coral Triangle nation currently funds coral reef research to any great extent. To be clear, there are certainly talented, passionate marine biologists and conservationists in these countries (Al-Asif et al., 2022), though at the present time their resources (with respect to funds, equipment, vessels, etc.) are not commensurate with the scale of what needs to be accomplished to foster the resilience of their fragile coral reefs (Kleypas et al., 2021).

The SCUBA diving industry has generally been effective at identifying reefs with high tourist appeal, and in some resort areas, there is some knowledge as to where one would go to find colorful, biodiverse, coral-rich reefs. Elsewhere, the lushest, most ecologically important reefs may be known only to local fisher people. What this means is that it is almost a surety that there are undiscovered coral reefs in the Coral Triangle that may have higher abundance of coral than 1) the few existing long-term monitoring sites (Boco et al., 2020) and 2) popular tourist sites. Although it is important to note that there is no correlation between coral cover and reef resilience (Wooldridge, 2014; Mayfield et al., 2015), there may nevertheless be a desire by managers, conservation biologists, or researchers to seek out a region's most coral-abundant area(s). In sparsely populated areas with few tourists, like the Solomon Islands, most of these high coral cover reefs have likely not been previously identified, and searching for them systematically would involve hundreds of thousands of USD worth of ship time over a multi-month, or even multi-year period. If there was a way to expedite the search for high coral cover reefs, this could dramatically reduce costs and field time.

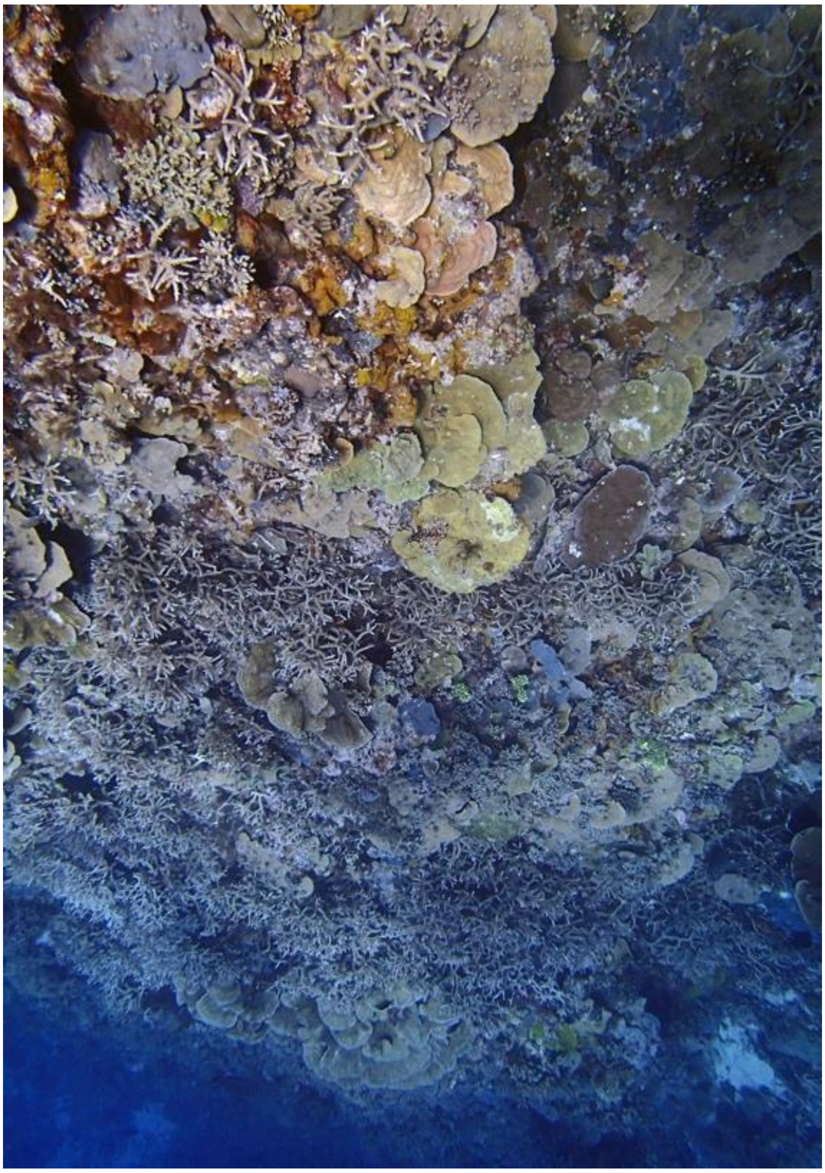


Fig. 1. A vibrant coral reef in the Solomon Islands extending from ~3 m (top of image) to ~20 m (bottom of image). This reef had not been seen by a human until 2014. Photograph by A.B.M.

The only large-scale coral reef survey of the Solomon Islands was conducted in 2014 by the Khaled bin Sultan Living Oceans Foundation (LOF) as part of their "*Global Reef Expedition*" (*GRE*), the largest coral reef survey ever undertaken (Mayfield et al., 2019). Although the survey lasted a month and spanned the entire archipelago (Fig. 2), only a modest portion of the nation's coral reefs were ultimately surveyed (Bruckner, 2015). Could the resulting "molecules-to-satellites" (i.e., encompassing data from the biochemical to the 100-km scale) dataset generated by the LOF science team and locally-based scientists nevertheless be used to develop predictive models that could tell us where *other* high coral cover reefs may be found in this relatively pristine region of the Coral Triangle? Towards this end, a large number of machine-learning and other, more

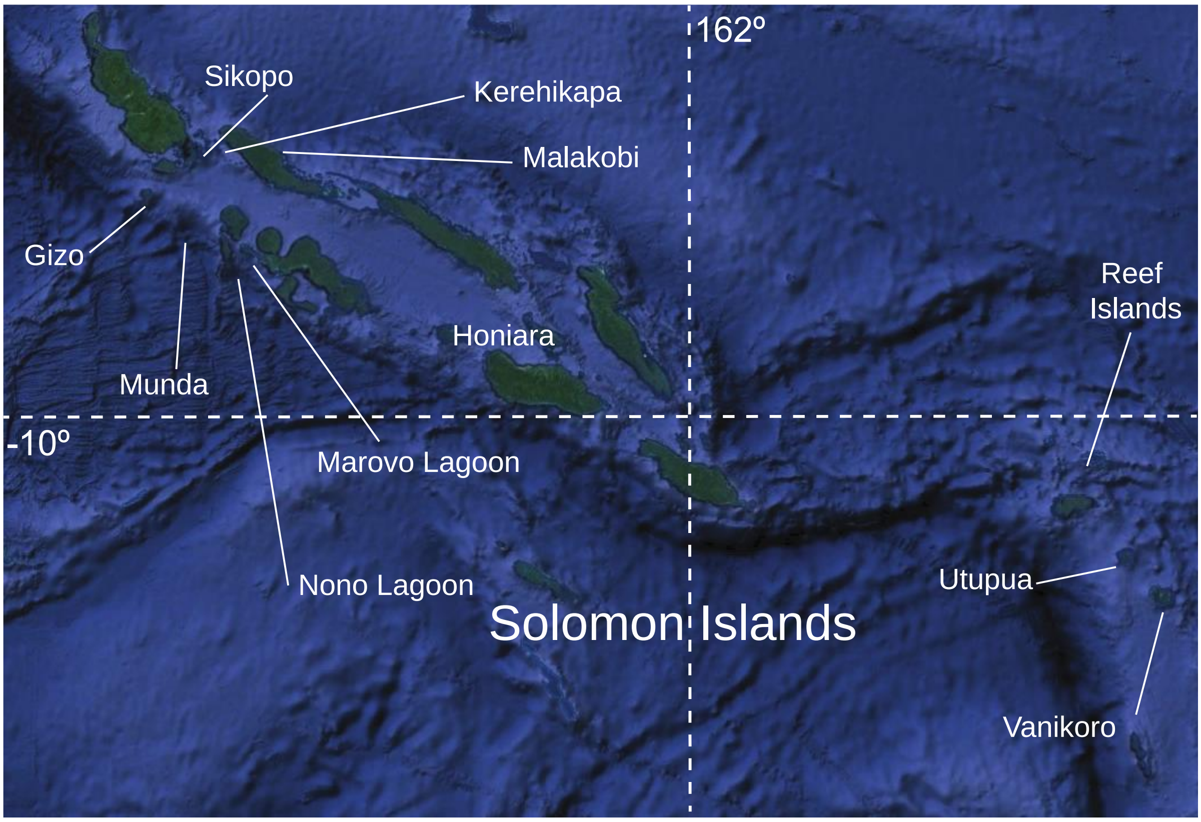


Fig. 2. Map of the Solomon Islands, with the major study regions labeled. The capitol of Honiara is depicted, as well, though no surveys were undertaken nearby. For high-resolution interactive maps and "drop-cam" videos of select reefs, please go to maps.lof.org/lof.

traditional models were built herein in JMP® Pro (NC, USA), and we hypothesized that we could use a series of simple, cheap- and easy-to-measure environmental parameters (ENV) to predict where the country's most coral-rich reefs are likely to be found.

Materials and methods

Overview of the approach. Although some general remarks on the reef survey data are made herein, we generally point readers to the field report (Bruckner, 2015) for this information. Briefly, 69 sites were visited across the Solomon Islands archipelago (Figs. 3-11)

Table 1. Statistical approaches for modeling benthic structure and live coral cover. For the multivariate analyses (multiple Y's), the ecological (ECO) data were the model Y's, with the environmental (ENV) data (n=14 unless denoted as "ENV(9)" in which only nine parameters were considered) as the putative predictors ("Model X's"). Either all 62 ECO or the factor loading scores (n=26) derived from a factor analysis were used for the benthic models ("Model Y('s)"). Gen-reg=generalized regression. MDS=multi-dimensional scaling. NA=not applicable. NN=neural network. NP-MANOVA=non-parametric multivariate ANOVA. PCA=principal components analysis. RS=response surface. Val col=validation column (training & validation samples). Val col w/test=validation column with test samples (training, validation, & test samples). *p<0.01.

To be uncovered	Model/analysis type	Model Y('s)	Model X's	Validation type	Conclusion; data location
Multivariate effects: benthic assemblage					
Relationships among transects	PCA	62 ECO	NA	NA	Fig. 13a
Similarity among transects	MDS	62 ECO	NA	NA	Fig. 13b
Effect of island on benthos	CCA	26 ECO factors	Island*	NA	Fig. 13c
Effect of reef exposure on benthos	CCA	26 ECO factors	Reef exposure*	NA	Fig. 13d
ECO dataset complexity reduction	Factor analysis	62 ECO	NA	NA	26-factor (68.7%)
ECO dataset complexity reduction	Factor analysis	50 ECO[a]	NA	NA	15-factor (83.7%)
ENV effects on benthic structure	NP-MANOVA	62 ECO	ENV^1	NA	Tab. 2
ENV effects on benthic structure	NIPALS	62 ECO	ENV^1, ENV^2, ENV^3, ENV-RS	Kfold7, val col, val w/test	Tab. 4
ENV effects on benthic structure	NIPALS	26 ECO factors	ENV^1, ENV^2, ENV^3, ENV-RS	Kfold7, val col, val w/test	Tab. 4
Univariate effects: live coral cover (%)					
ENV effects on coral cover (%)	Predictor screen	% coral cover	ENV^1	NA	Site (59%); Fig. 12
ENV effects on coral cover (%)	Model screen	% coral cover	ENV^1	Kfold5, val col, val w/test	Tab. 4
ENV effects on coral cover (%)	NN GUI-HL1	% coral cover	ENV(9)	20% holdback, val col, val w/test	R^2=0.81[b]
ENV effects on coral cover (%)	NN GUI-HL1	% coral cover	ENV^1	Kfold5, val col, val w/test	Tab. 4
ENV effects on coral cover (%)	NN GUI-HL2	% coral cover	ENV^1	Kfold5, val col, val w/test	Tab. 4
ENV effects on coral cover (%)	NIPALS	% coral cover	ENV^1, ENV^2, ENV^3, ENV-RS	Kfold7, val col, val w/test	Tab. 3
ENV effects on coral cover (%)	Gen-reg	% coral cover	ENV^1, ENV^2, ENV^3, ENV-RS	Minimum AICc	Tab. 4

[a] Subset of 50 ECO associated with the transects from which corals were sampled (i.e., 12 coral genera were not found in the vicinity of the sampled corals; see Mayfield et al., under review.). [b] Obtained when validating with the 20% holdback approach (Fig. 14): TanH(1)-Linear(1)-Gaussian(3)-Boost(17).

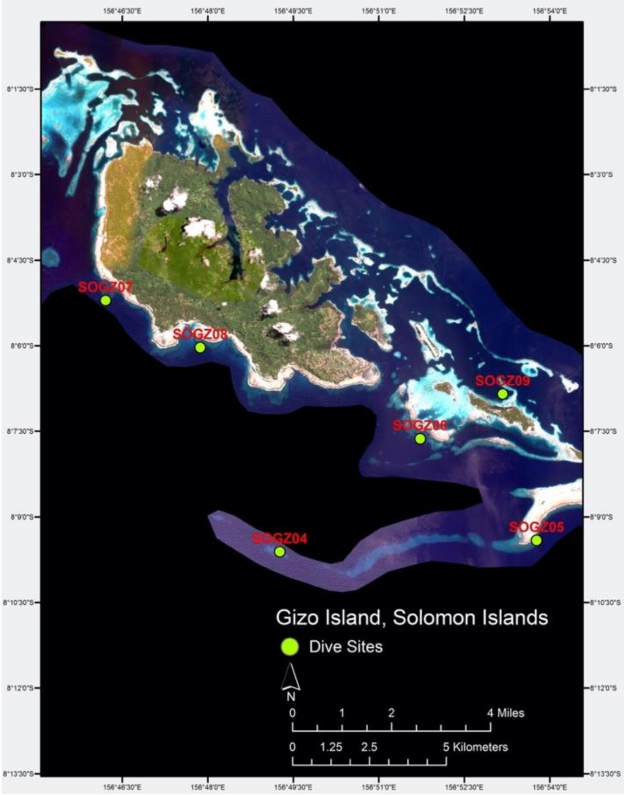


Fig. 3. WorldView-02 imagery depicting three reef sites surveyed off Munda Island (MU), Solomon Islands (2014-10-29).

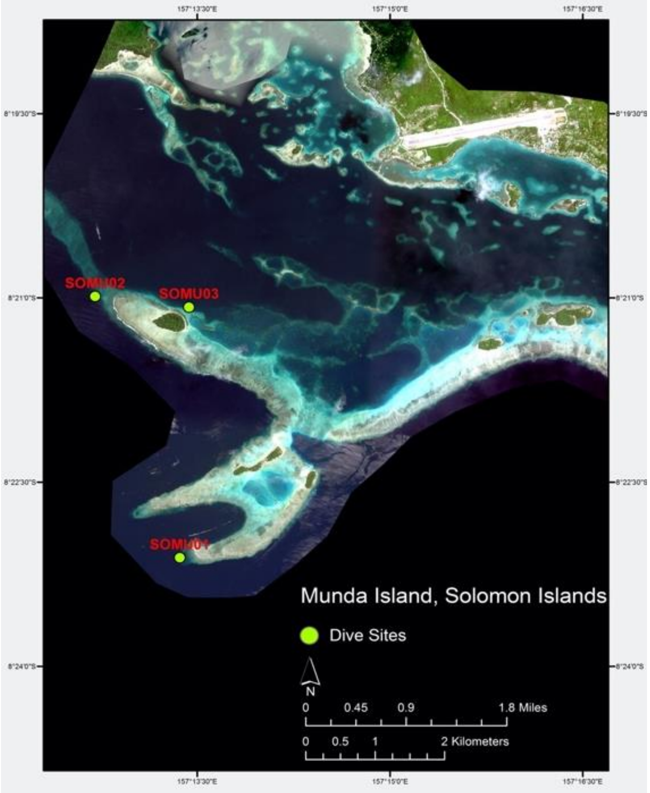


Fig. 4. WorldView-02 imagery depicting six reef sites surveyed off Gizo Island (GZ), Solomon Islands (2014-10-30 & 2014-10-31).

from October to November of 2014. Benthic surveys, seawater sampling, and reef coral sampling (Mayfield et al., 2017) were undertaken by LOF scientists (see below for details on the former.). Images of the reefs and sampled corals can be found at coralreefdiagnostics.com, while satellite and underwater video-derived habitat maps can be accessed at maps.lof.org/lof. Our goal herein was to exploit a diverse array of analytical approaches in JMP Pro 16 (*sensu* Mayfield [2020]) to understand environmental influences on coral reef ecology (Tab. 1). As a preliminary step, we sought to identify the regional environmental factors that are responsible for spatio-temporal variation in the coral reef benthic assemblages (i.e., the entire community; Rodríguez-Troncoso et al. [2019]); this analysis considered the benthos within a multivariate framework (Mayfield & Chen, 2019). To complement this analysis, we conducted a more common modeling analysis of live coral cover (%) to determine which environmental conditions are associated with the highest percent coral cover in the Solomon Islands. We then used this information to train models capable of reliably predicting coral cover of undiscovered reefs.

Data collection. Diver-based assessments of the benthos along point-intercept transects laid parallel to the shoreline at depths between 5 and 30 m were undertaken at replicate locations within each site x depth, with photo-quadrats imaged along similar depth contours in the vicinity of the transects. The benthic composition within the images was verified by eye while aboard the ship and merged with the data from the in-water diver surveys along the nearby transects. The following 14 ENV were documented and hypothesized to be potential drivers of variation in coral cover: island (n=10; see Fig. 2.), reef site, latitude, longitude, date (n=23 days), time (binned as either morning [<10:00], midday [10:00-14:00], or afternoon [>14:00]), temperature (°C), salinity (unit-less), reef type (fringing reef, barrier reef, patch reef, or other), reef exposure (protected, intermediately exposed, or exposed), reef location (fore reef or lagoon), lagoon (inside vs. outside), reef (emergent vs. submergent), and depth (m; as four bins: <8 m, 8-12 m, 12-18 m, or 18-25 m). Note that some of these factors co-vary; for instance, different sites were surveyed on different days. Also note that, although sampling time was predicted to influence coral physiology (the focus of a companion work; Mayfield et al., under review), it was not expected to have a statistically significant impact on the benthic assemblage; it was left in preliminary model-building exercises though later removed for this reason (as were four other ENV; described below). Sixty-two ECO were considered in the benthic composition analysis (all in percentages of total benthic cover; see online supplemental data file [OSDF] for complete list of all taxa, as well as abbreviations used in the manuscript’s figures & tables.): barren substrate (PB), invertebrates (PITS), six algal taxa, and 54 coral genera.

Characterizing the benthos-I: multivariate approaches. To determine the prevailing local influences on the benthic structure of the surveyed reefs (as defined by the aforementioned 62 ECO), transect data were averaged across surveyors at each depth for each reef site; this resulted in 272 site x depth groups (see depth bins above.). We did *not* pool data across depths for each site because we hypothesized that coral cover and the benthic community would differ across

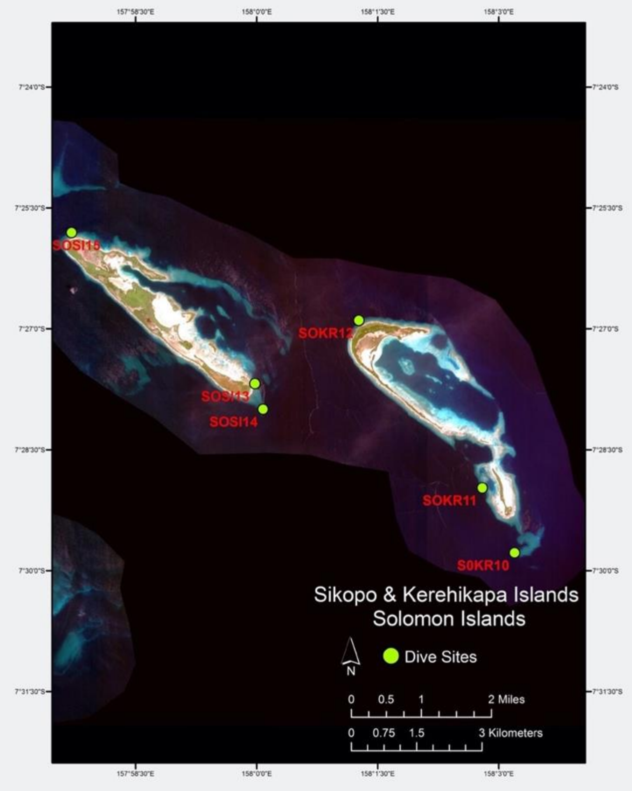


Fig. 5. WorldView-02 imagery depicting six reef sites surveyed off Sikopo (SI) and Kerehikapa (KR) Islands, Solomon Islands (2014-11-01 & 2014-11-02).

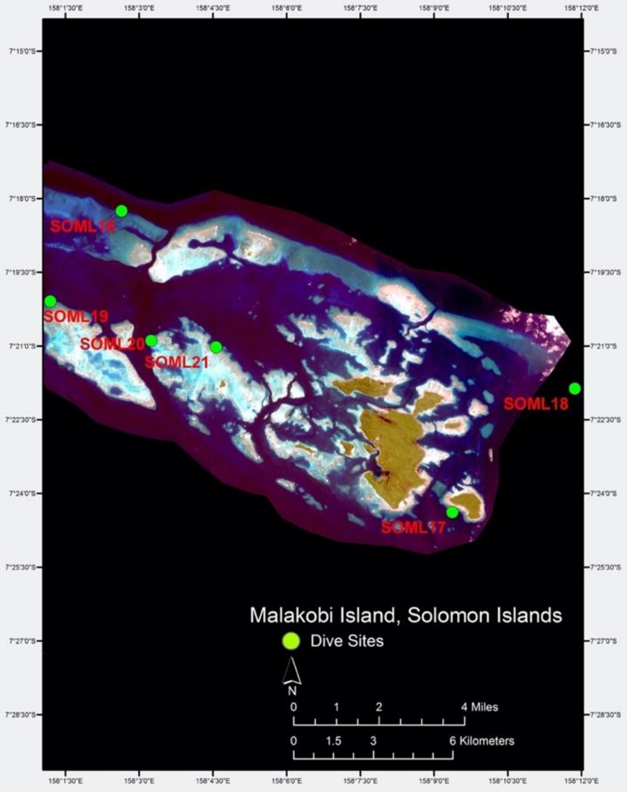


Fig. 6. WorldView-02 imagery depicting six reef sites surveyed off Malakobi Island (ML), Solomon Islands (2014-11-03 & 2014-11-04).

depths within each site. The 14 ENV were used as predictors in six multivariate analyses aimed at understanding the local factors that shaped the benthic structure (Tab. 1). First, both multi-dimensional scaling (MDS; method#1) and principal components analysis (PCA; method#2) were undertaken with the benthic dataset to depict similarity (Euclidean distance matrix of standardized data) and relationships (PCA on correlations), respectively, among the site x depth groups. The coordinates for the first six MDS dimensions were then used as model Y terms in a non-parametric multivariate ANOVA (NP-MANOVA; method#3) in which each of the 14 ENV in isolation was tested. As another means of reducing benthic dataset complexity, a factor analysis was undertaken (method#4), and 26 factors were deemed by JMP Pro 16 to represent the optimal number of dimensions for explaining the highest percentage of variation in the dataset with

doi: 10.29926/platax.202212_18.0000

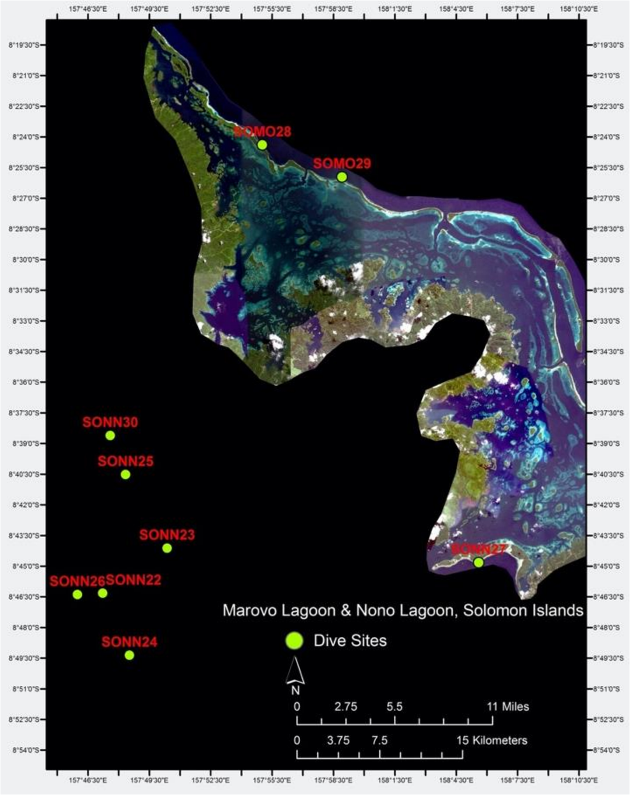


Fig. 7. WorldView-02 imagery depicting six reef sites surveyed off Marovo (MO) and Nono (NN) Lagoons, Solomon Islands (2014-11-06 & 2014-11-07).

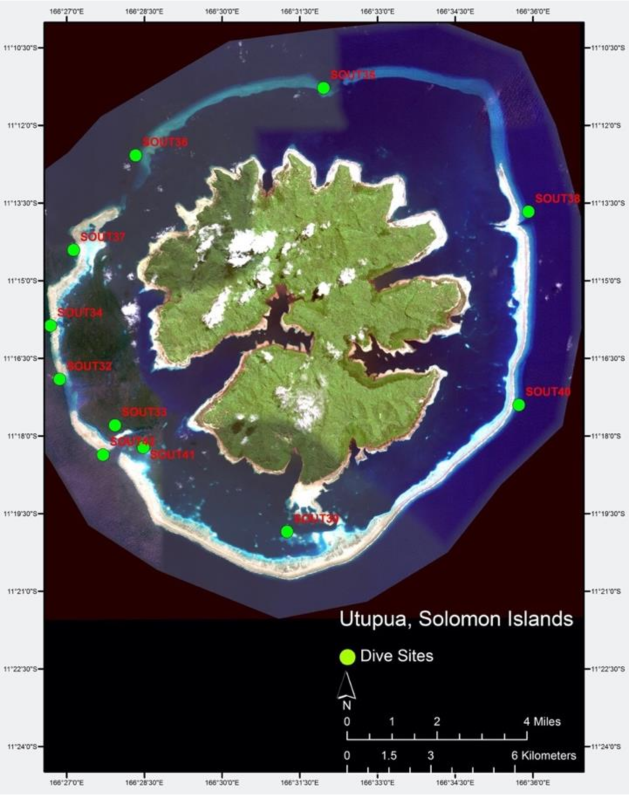


Fig. 8. WorldView-02 imagery depicting 11 reef sites surveyed off Utupua (UT), Solomon Islands (2014-11-10 to 2014-11-13).

the fewest amount of input predictors (Tab. 1). Because NP-MANOVA uncovered effects of island and reef exposure on the reef benthos (Tab. 2), a canonical correlation analysis (CCA) was undertaken (method#5), as well, using JMP Pro's "discriminant analysis" with the 26 factor loading scores as Y's.

Either the 62 ECO or their associated 26 factors were next used in partial least squares (PLS) analyses (non-iterative PLS [NIPALS]; method#6) in which various factorial combinations of the 14 ENV (first-, second-, or third-order) were the model X's (i.e., predictors); the response surface design was tested as a fourth schematic. Three validation types were tested with each grouping of X's for both the 62 ECO and the 26 factor loading scores (24 PLS models built in total): Kfold7, validation column ("Val col" in tables; transects randomly assigned as either being training [75%] or validation [25%] samples), and validation with test

Platax 19: 01-24, 2022
doi: 10.29926/platax.202212_18.0000

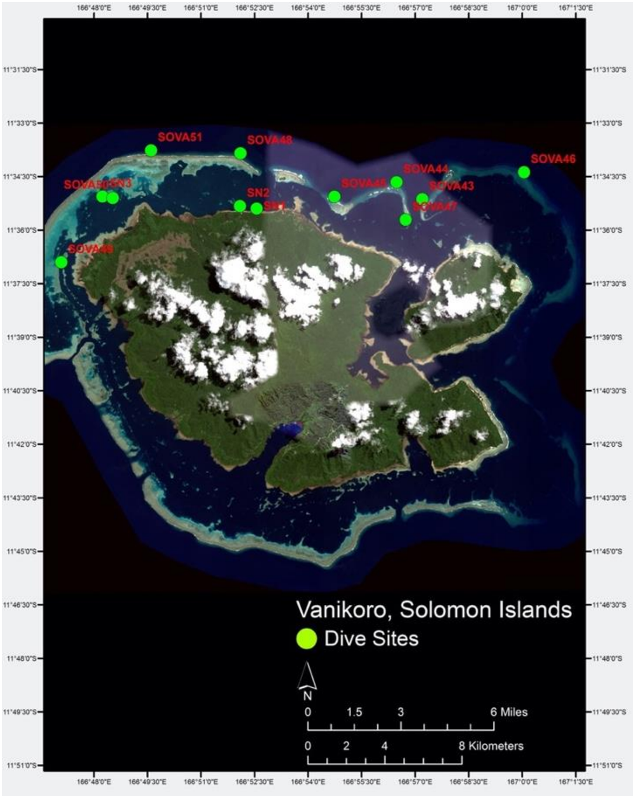


Fig. 9. WorldView-02 imagery depicting nine reef sites surveyed off Vanikoro (VA), Solomon Islands (2014-11-14 to 2014-11-16), as well as three snorkel survey sites (SN1-3).

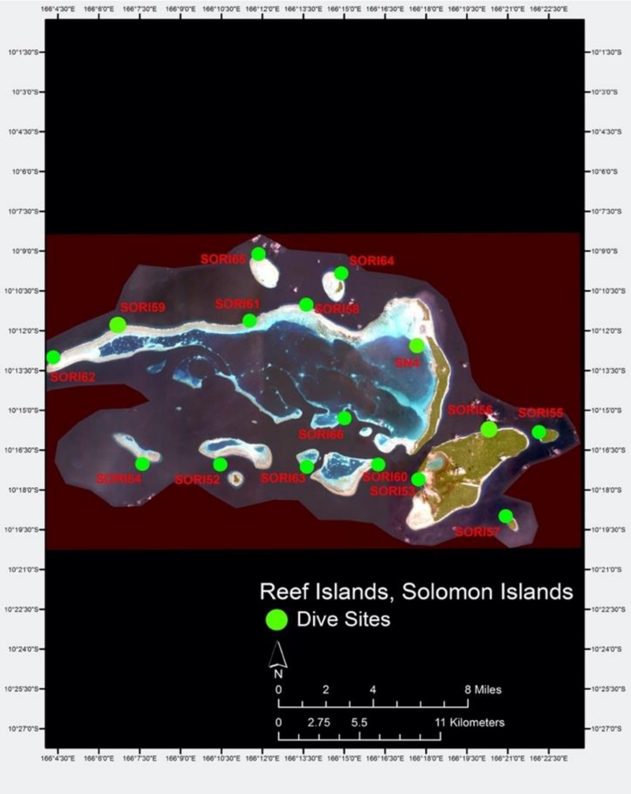


Fig. 10. WorldView-02 imagery depicting 15 reef sites surveyed off the Reef Islands (RI), Solomon Islands (2014-11-17 through 2014-11-21). One snorkel site (SN4) is also shown, though data from this site were excluded from the comprehensive analyses presented herein.

("Val w/test;" transects randomly assigned as training [75%], validation [15%], or test [10%] samples).

Characterizing the benthos-II: univariate approaches. As a simpler means of characterizing the benthic ecology, live coral cover (%) alone was considered as the lone model Y (Tab. 1). First, a predictor screen of the 14 ENV was undertaken (100 trees with a bootstrap forest algorithm) to rank the ENV with respect to their influence on coral cover. Secondly, JMP Pro's "model screen" was used to test the following 13 modeling types with first- and second-order factorials of the 14 ENV (using Kfold5 cross-validation, validation column holdback data, or both validation & test data for model validation; Tab. 4): ordinary least squares, stepwise regression, generalized regression (gen-reg), PLS (NIPALS), discriminant analysis, decision tree, bootstrap forest, boosted tree, Naïve Bayes, k-nearest neighbors, support vector machines, neural network (NN), and XGBoost. When the model with the highest validation (or test) sample R^2 was the NN, a NN model-tuning GUI developed by Diedrich Schmidt (ver. 5.0) was used to optimize the following model parameters (*sensu* Mayfield [2022]) with 20% holdback, validation column, or validation+test data column validation and the weight decay penalty method: number of hidden layers (HL; 1 vs. 2), type of activation (sigmoidal [TanH], linear, or radial [Gaussian]), number of activation nodes per hidden layer (0, 1, 2, 3, or 4), number of boosts (ensemble models; for single-layer models only: 0-20), learning rate (only for boosted models: 0.05-0.3), number of tours (1-100), covariate transformation (transformed or untransformed), and robust fit (yes or no). At least 200 models were built for each combination of HL and validation type. Furthermore, several hundred additional NN models were built (Tab. 4) in which the number of activation nodes was permitted to rise to 6, up to 30 boosts could be featured in HL1 models, and up to 200 tours were considered (both HL1 & HL2 models).

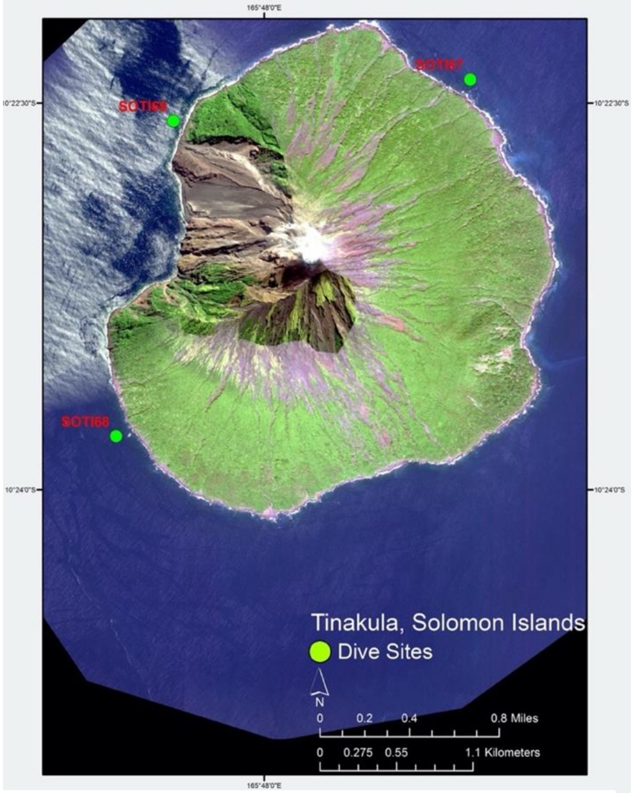
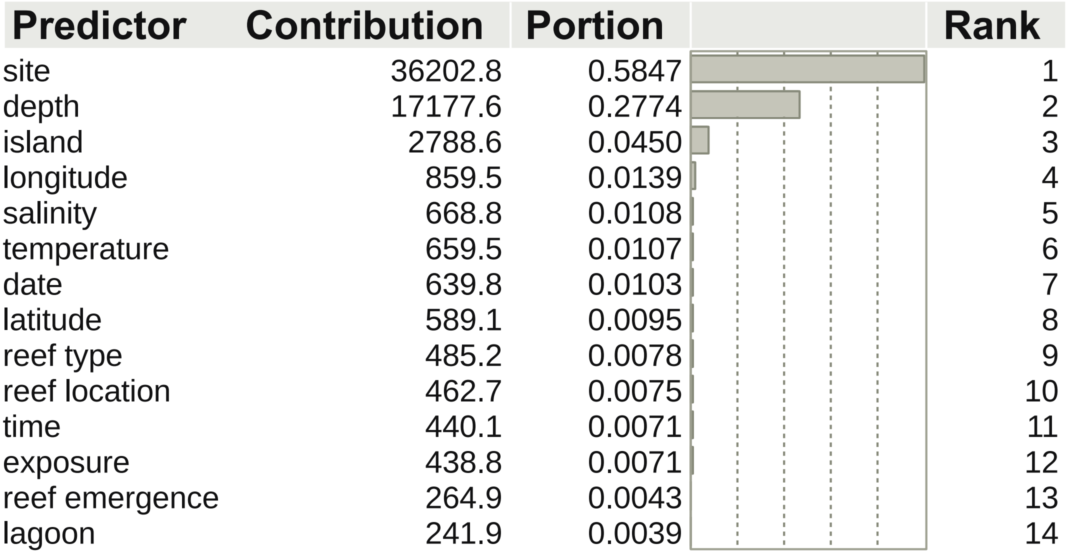


Fig. 11. WorldView-02 imagery depicting three reef sites surveyed off Tinakula Volcano (TI), Solomon Islands (2014-11-22).

The superior NN model was then used in a machine-learning "desirability analysis" in which the artificial intelligence (AI) was programmed to maximize coral cover. Note that because multiple tours were permitted, repeat runs of the same model could result in different results. For this reason, the optimal model (max. validation or test R^2) in the NN GUI was run 10 times, with the highest value obtained presented in the manuscript's tables. We set an *a priori* validation (or test) R^2 cutoff of 0.8 as being of sufficient fit to



Predictor	Contribution	Portion	Rank
site	36202.8	0.5847	1
depth	17177.6	0.2774	2
island	2788.6	0.0450	3
longitude	859.5	0.0139	4
salinity	668.8	0.0108	5
temperature	659.5	0.0107	6
date	639.8	0.0103	7
latitude	589.1	0.0095	8
reef type	485.2	0.0078	9
reef location	462.7	0.0075	10
time	440.1	0.0071	11
exposure	438.8	0.0071	12
reef emergence	264.9	0.0043	13
lagoon	241.9	0.0039	14

Fig. 12. Predictor screening. A bootstrap forest model with 100 random trees was used to estimate the contribution of each of the 14 environmental parameters to variation in live coral cover (%).

be useful for predicting live coral cover *in situ*.

Secondly, NIPALS was undertaken as described above with second- and third-order factorial combinations of the 14 ENV (as well as response surfaces) as model X's and live coral cover as the singular Y (Tab. 3). As a tertiary means of determining the ENV or combinations thereof that best explained variation in coral cover, gen-reg was used with the following algorithms (using JMP Pro 16 terminology): forward selection, pruned forward selection, "best subset," lasso (regular & adaptive), elastic net (regular & adaptive), double lasso (regular & adaptive), and ridge regression. With the exception of the latter, in which 30% holdback validation was used to evaluate model performance, the others were ranked based on their AICc (with the model with the lowest AICc deemed superior). Unlike for PLS, only first and second-order factorials were considered as model X's for gen-reg given that the computing power needed to test 2,744 putative model terms (14^3) exceeded 64 GB of RAM; cloud computing could be exploited in the future to test the efficacy and accuracy of these more complex modeling types. In general, though, gen-reg was not found to be a robust modeling approach for this dataset (Tab. 4), and the associated results have not generally been discussed at length herein.

Results and Discussion

Overview of coral cover. Live coral cover averaged 35±20% (std. dev. for this & all other error terms unless noted otherwise) across the 69 sites, reaching 84% on some reefs. A predictor screen (Fig. 12) revealed that site location and depth had the greatest influence on coral cover, and the depth effect was particularly pronounced; specifically, coral cover varied significantly across survey depth bins (one-way ANOVA p<0.0001) and averaged 33, 51, 27, and 27% at <8, 8-12, 12-18, and 18-25 m, respectively. Algal cover was oftentimes higher, averaging 46±19% across the 69 survey sites (maximum site mean=97%); however, the coral/algae ratio was slightly greater than 1 (1.1±1.2) when averaged across all sites (see OSDF.).

Tab. 2. Non-parametric multivariate ANOVA. Effects of the 14 environmental parameters (ENV) on the benthic assemblage (as 6 multi-dimensional scaling dimensions derived from Euclidean distances among 272 site x depth groups based on a similarity analysis of 62 benthic categories). Statistically significant differences (alpha=0.01) have been highlighted in bold.

ENV (predictor)	*n*	*F*	*p*	*Post-hoc* comparisons
Island	11	5.49	**<0.01**	see Fig. 13c
Site	68	2.68	**<0.01**	
Latitude	68	2.36	0.03	
Longitude	68	7.13	**<0.01**	
Survey date	23	8.42	**<0.01**	
Survey time	3	0.80	0.65	
Survey depth	4	7.29	**<0.01**	
Reef exposure	3	4.27	**<0.01**	see Fig. 13d
Reef type	4	2.94	**<0.01**	patch(a)≠fringing(bc)=barrier(c)
Lagoon	2	3.65	**<0.01**	inside lagoon≠outside lagoon
Reef emergence	2	0.94	0.47	
Reef location	2	5.25	**<0.01**	fore reef≠lagoon
Temperature	15	1.40	0.22	
Salinity	9	2.89	**<0.01**	

Benthic composition. A two-dimensional PCA (Fig. 13a) explained only 10% of the variation in the benthic assemblage (as assessed from the 62 ECO). However, upon coloring the replicate transects (depth x site) by coral cover in an MDS analysis (Fig. 13b), it was clear that the high coral cover assemblages differed from the lower ones. This is expected since reefs with higher coral cover are likely to have a more complex benthic structure except for when the benthos is dominated by a monoculture of a single dominant species (rarely observed in the Solomon Islands). A factor analysis was able to reduce the complexity of the 62-parameter benthic dataset into a 26-factor solution that encapsulated almost 70% of the variation (Tab. 1).

Both descriptive and predictive approaches were used to model the benthos across the 272-transect dataset (Tab. 1). In the former, NP-MANOVA (Tab. 2) found a number of ENV to significantly affect the benthic community structure, and both the island and reef exposure effects are depicted in two dimensions in Fig. 13c and 13d, respectively, with discriminant analysis (i.e., CCA). The island effect was mainly driven by the strange coral community at the base of the Tinakula Volcano; although corals had returned to the area since the 2007 eruption, the ecosystem was clearly still in a state of succession. Interested readers are encouraged to check out the coral reef imagery on the data servers

Tab. 3. Partial least squares (non-iterative [NIPALS]) of environmental (ENV) effects (X) on percent (%) live coral cover (Y). Val col=validation column. Val w/test=both validation and test samples used to validate model accuracy. RS=response surface.

Model X's	Validation	#factors (% variation explained)
ENV[1]	Kfold7	2 factors (61%)
ENV[1]	Val col	2 factors (69%)
ENV[1]	Val w/ test	15 factors (68%)
ENV[2]	Kfold7	5 factors (89%)
ENV[2]	Val col	4 factors (84%)
ENV[2]	Val w/ test	6 factors (93.6%)
ENV[3]	Kfold7	4 factors (82%)
ENV[3]	Val col	4 factors (84%)
ENV[3]	Val w/ test	6 factors (94%)
ENV-RS	Kfold7	4 factors (82%)
ENV-RS	Val col	4 factors (84%)
ENV-RS	Val w/ test	6 factors (93.5%)

mentioned above. With respect to the strong reef exposure effect, which was expected given the well-studied influence of wave energy on reef ecosystems (Ferrario et al., 2014), it is interesting to note that the exposed and protected reefs were not most distinct from one another in terms of the benthic composition. Instead, the exposed reefs' benthic community fell between those of the intermediately exposed and protected reefs.

Predicting coral cover. To instead attempt to make predictions of benthic composition, various factorial combinations of the 14 ENV were used as model X's, with either all 62 benthic bins or the 26 factors derived from the aforementioned factor analysis as Y's, in a series of PLS analyses (Tab. 4). In general, however, the percent variation explained by even the more complex PLS model types were <10%, and for this reason we have instead focused on attempting to predict a singular ecological response metric: live coral cover. The underlying models of percent live coral cover (Tabs. 3 & 4) were characterized by higher R^2 values. Even when considering the 14 ENV alone (first-order), R^2 values ranged from 0.6 to 0.7 (depending on validation type; Tab. 3). When incorporating squared or cubed factorial combinations of the environmental parameters, values rose to 0.8-0.9. However, upon validating the PLS model with the highest R^2 in Tab. 3 (0.94) with holdback data, the R^2 dropped to a value that was below our *a priori* cutoff of 0.8 (0.74). Furthermore, it is important to note that five of the underlying ENV-site, date, time of day, and, to a lesser extent, salinity and temperature- would *not* be of use to those looking to use these models to predict where *other* high coral cover reefs would be found in the Solomon Islands. Date, in particular should not be included since future survey will inherently occur on different days. We have left island in the models, though an argument could also be made to only consider truly general properties of the reefs, not to mention the fact that latitude and longitude will inherently covary with island. Upon removing these five terms, the model screening analysis was re-run, and the R^2 of one neural network was actually slightly higher: 0.81 (Tab. 1 & Fig. 14a). It is worth noting that this is far higher than R^2 values obtained in prior meta-analyses that sought to predict coral cover from environmental data (e.g., 0.3-0.4 [Hochberg & Gierach, 2021]).

Unlike more traditional modeling types like standard least squares, the inclusion of fewer predictors does not necessarily result in a weaker fit in neural networks, especially when boosting is utilized. In a dependent resampled inputs analysis with this superior model (Fig.

14a), depth (total effect=0.43) was the most influential predictor. This is due in part to the significantly higher coral cover between 8 and 12 m mentioned above. In a desirability analysis of this neural network (Fig. 14b), a theoretical maximum coral cover of 75% could be reached at 8-12 m on an intermediately exposed fringing fore reef, and this is where surveyors looking to find high coral-cover reefs should go first in the all-too-likely event that field time may be limited. Note that this is actually a lower coral cover percentage than documented at some of the field sites (up to 84%; OSDF), meaning there are reefs in the country whose coral cover is actually higher than that predicted by the AI; this discrepancy should certainly be explored or addressed in future works.

Even the superior neural network model could not accommodate ~20% of the variation in coral cover; additional environmental variables that were not measured herein, such as nutrient levels (Huang et al., 2020), are clearly important in determining live coral cover in the Solomon Islands and should be included in future analyses. We also plan to later include demographic data as potential predictors, namely population of the nearest human settlements to the study reefs, education level/literacy, carbon footprint, seawater pollution levels, and other such variables that could be hypothesized to affect the marine environment, and specifically coral cover. During surveys, we also collected a wealth of fish biodiversity and biomass data; these data should also be incorporated into predictive model building given the importance of herbivores in particular in maintaining reef health and function (Cramer et al., 2017). Upon factoring in additional environmental, ecological, seawater quality, and demographic data, it is not unreasonable to expect that predictive models of coral cover surpassing R^2 of 0.9 could be generated. This would greatly aid managers in triaging survey or conservation efforts in instances in which a research team could not realistically survey all reefs present during a field season or research expedition (normally the case) by allowing them to more rapidly find areas with high coral cover; although not necessarily of higher resilience, there may nevertheless be a desire to prioritize such areas of high coral abundance for conservation.

Conclusions

Although we were not able to use a series of 14 ENV to robustly model the complex, multivariate nature of the shallow (0-30 m) coral reef communities of the Solomon Islands, the prediction of

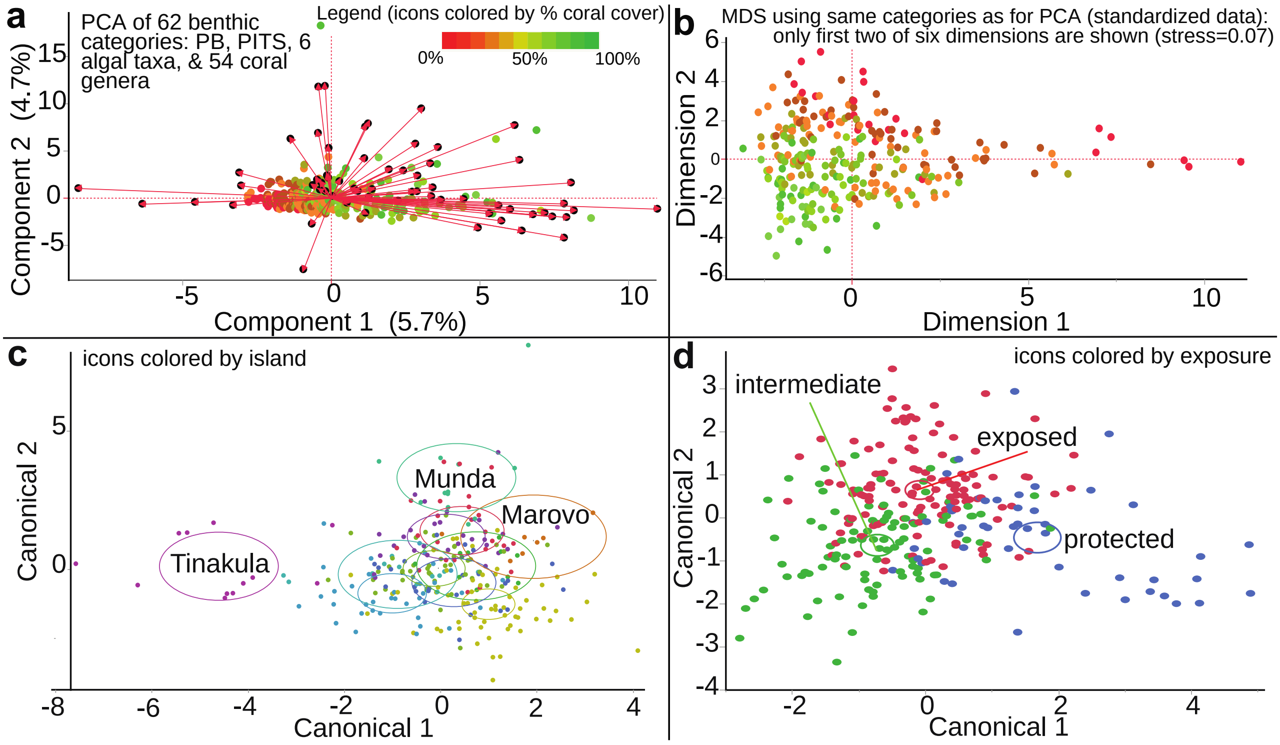


Fig. 13. Multivariate analysis of the reef benthos. Principal components analysis (PCA; on correlations; a), multi-dimensional scaling (MDS; on standardized data; b), and canonical correlation analysis (CCA; c-d) were undertaken with the 62-category benthic dataset (n=272 transects). For the biplot rays in panel a, please see the online supplemental data file. The legend in panel a extends to panel b. PB=percent barren substrate. PITS=percent invertebrate cover. Ellipses in the island (c) and reef exposure (d) CCA plots represent 95% confidence, and transects have been coded by island and reef exposure, respectively. Three island groups have been listed in panel c. Please see Tab. 2 for NP-MANOVA results.

Tab. 4. All predictive models (n>3,000) of environmental (ENV) effects on the benthic structure (ECO) or coral cover (%). Partial least squares (PLS; non-iterative [NIPALS]) of coral cover are instead found in Tab. 3. Either all 14 ENV (“ENV(14)”) or a subset of 9 (“ENV(9)”) hypothesized to be of greater utility to managers and future researchers were included (excluding date, site, time, temperature, & salinity). Parenthetical sample sizes in the right-most column (“n=”) correspond to number of terms in the superior model. Gen-reg= generalized regression. NN=neural network. RS=response surface. Val col=validation column. Val w/test=both validation and test data columns.

Analysis	Model Y’s	Model X‘s	Validation	Conclusion (model details)
ENV effects on benthic structure				
NIPALS	62 ECO	ENV(14)[1]	Kfold7	4-factor (8.0%)
NIPALS	62 ECO	ENV(14)[1]	Val col	2-factor (4.4%)
NIPALS	62 ECO	ENV(14)[1]	Val w/test	4-factor (8.7%)
NIPALS	62 ECO	ENV(14)[2]	Kfold7	1-factor (2.5%)
NIPALS	62 ECO	ENV(14)[2]	Val col	2-factor (5.6%)
NIPALS	62 ECO	ENV(14)[2]	Val w/test	3-factor (8.0%)
NIPALS	62 ECO	ENV(14)[3]	Kfold7	1-factor (3.0%)
NIPALS	62 ECO	ENV(14)[3]	Val col	1-factor (3.3%)
NIPALS	62 ECO	ENV(14)[3]	Val w/test	3-factor (9.0%)
NIPALS	62 ECO	ENV(14)-RS	Kfold7	3-factor (7.4%)
NIPALS	62 ECO	ENV(14)-RS	Val col	2-factor (5.6%)
NIPALS	62 ECO	ENV(14)-RS	Val w/test	3-factor (7.9%)
NIPALS	26 ECO factors	ENV(14)[1]	Kfold7	4-factor (7.7%)
NIPALS	26 ECO factors	ENV(14)[1]	Val col	2-factor (4.5%)
NIPALS	26 ECO factors	ENV(14)[1]	Val w/test	3-factor (6.9%)
NIPALS	26 ECO factors	ENV(14)[2]	Kfold7	1-factor (2.2%)
NIPALS	26 ECO factors	ENV(14)[2]	Val col	2-factor (6.0%)
NIPALS	26 ECO factors	ENV(14)[2]	Val w/test	2-factor (5.3%)
NIPALS	26 ECO factors	ENV(14)[3]	Kfold7	1-factor (2.6%)
NIPALS	26 ECO factors	ENV(14)[3]	Val col	2-factor (7.4%)
NIPALS	26 ECO factors	ENV(14)[3]	Val w/test	2-factor (6.5%)
NIPALS	26 ECO factors	ENV(14)-RS	Kfold7	1-factor (2.2%)
NIPALS	26 ECO factors	ENV(14)-RS	Val col	2-factor (6.0%)
NIPALS	26 ECO factors	ENV(14)-RS	Val w/test	2-factor (5.3%)

Analysis	Model Y's	Model X's	Validation	Conclusion (model details)
ENV effects on coral cover (%)[a]				
Model screen	% coral cover	ENV(9)^1	Kfold5	NN (R^2=0.52)
Model screen	% coral cover	ENV(9)^1	Val col	NN (R^2=0.44)
Model screen	% coral cover	ENV(9)^1	Val w/test	Stepwise regression (R^2=0.43)
NN GUI-HL1	% coral cover	ENV(9)^1	20% holdback	R^2=0.81[b] (300 models run)
NN GUI-HL1	% coral cover	ENV(9)^1	20% holdback	R^2=0.83 (30 models run) TanH(2)-Linear(1)-Gaussian(3)-Boost(19)
NN GUI-HL2	% coral cover	ENV(9)^1	20% holdback	R^2=0.73 (300 models run)
NN GUI-HL1	% coral cover	ENV(9)^1	Val col	R^2=0.73 (300 models run)
NN GUI-HL2	% coral cover	ENV(9)^1	Val col	R^2=0.63 (300 models run)
NN GUI-HL1	% coral cover	ENV(9)^1	Val w/test	R^2=0.59 (300 models run)
NN GUI-HL2	% coral cover	ENV(9)^1	Val w/test	R^2=0.59 (300 models run)
Model screen	% coral cover	ENV(14)^1	Kfold5	NN (R^2=0.67)
Model screen	% coral cover	ENV(14)^1	Val col	All model R^2<0.4
Model screen	% coral cover	ENV(14)^1	Val w/test	Bootstrap forest (R^2=0.60)
NN GUI-HL1	% coral cover	ENV(14)^1	Val col	R^2=0.62 (200 models run)
NN GUI-HL2	% coral cover	ENV(14)^1	Val col	R^2=0.37 (200 models run)
NN GUI-HL1	% coral cover	ENV(14)^1	Val w/test	R^2=0.73 (200 models run)
NN GUI-HL2	% coral cover	ENV(14)^1	Val w/test	R^2=0.70 (200 models run)
Gen-reg	% coral cover	ENV(14)^2	Min AICc	Adaptive double lasso (n=41; R^2=0.58)
Gen-reg	% coral cover	ENV(14)^2	Min AICc-val col	Forward selection (n=14; R^2=0.46)
Gen-reg	% coral cover	ENV(14)-RS	Min AICc	Adaptive double lasso (n=41; R^2=0.58)
Gen-reg	% coral cover	ENV(14)-RS	Min AICc-val col	Forward selection (n=5; R^2=0.20)

[a]SinH-ArcsinH-transformed data. [b]See Fig. 14.

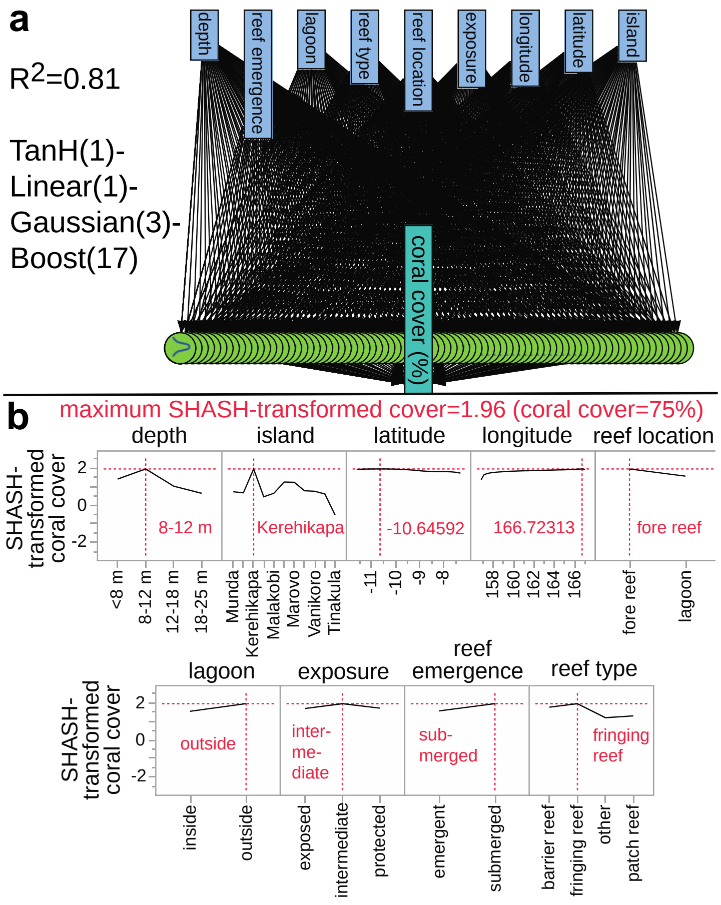
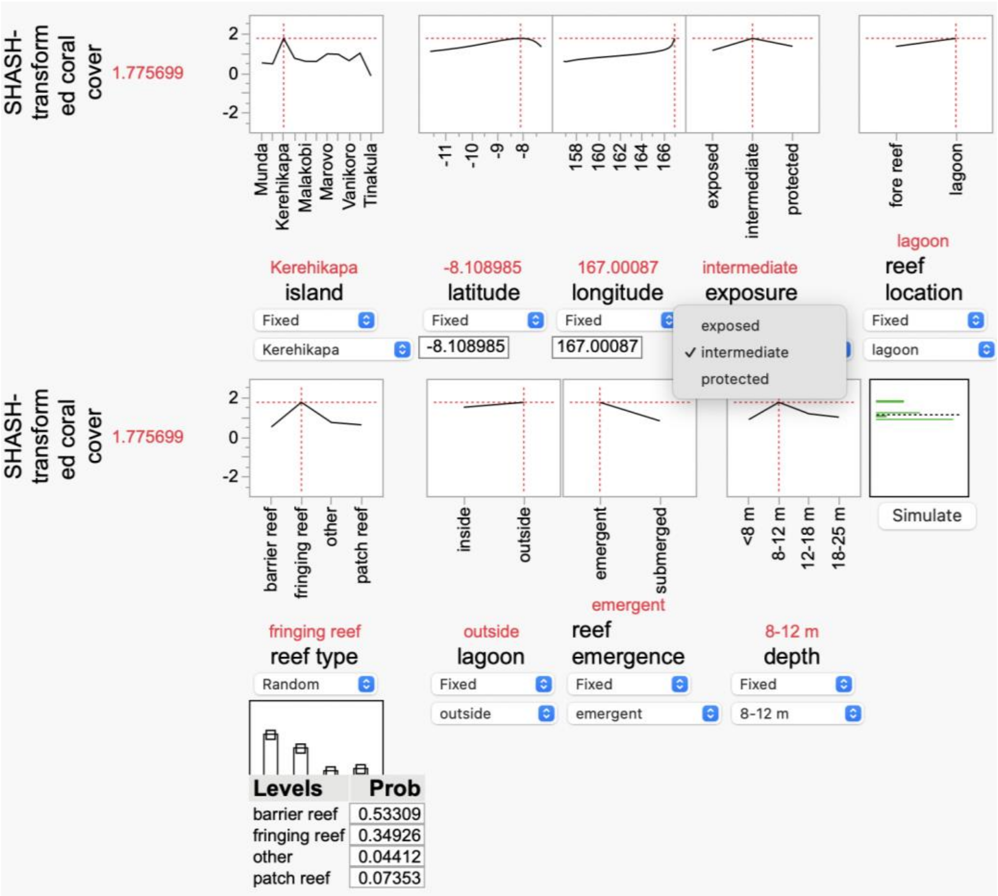


Fig. 14. Machine-learning model-building for maximizing coral cover. A neural network (a) for predicting coral cover (SHASH-transformed) from nine environmental (ENV) parameters and a machine-learning-based "desirability analysis" (b) based on the associated model. The conditions/levels resulting in a hypothetical maximum coral cover (~75%) are displayed in the center of each plot in panel b.

coral cover was more successful (R^2>0.80), even when including only nine easy-to-measure predictors (e.g., reef type, which can be deduced simply by consulting a map). The neural networks specifically suggest that those interested in uncovering high coral cover reefs should target intermediately exposed fringing reefs. We encourage interested scientists and managers alike to consult the application ("app") derived from the model (Fig. 15); this GUI allows one to toggle different



Kerehikapa
island
Fixed
Kerehikapa

-8.108985
latitude
Fixed
-8.108985

167.00087
longitude
Fixed
167.00087

intermediate
exposure
exposed
✓ intermediate
protected

lagoon
reef location
Fixed
lagoon

fringing reef
reef type
Random

outside
lagoon
Fixed
outside

emergent
reef emergence
Fixed
emergent

8-12 m
depth
Fixed
8-12 m

Levels	Prob
barrier reef	0.53309
fringing reef	0.34926
other	0.04412
patch reef	0.07353

Fig. 15. An interactive simulator built in JMP® Pro 16 that allows users to predict live coral cover (%) based on modifying 14 common environmental predictors. The underlying model is that depicted in Fig. 14, and the desirability analysis' theoretical maximum value is shown next to the y-axis (corresponding to a live coral cover of 75-80%); note that there are differences between the plots and those of Fig. 14 because of the large number of tours incorporated (described above). In this example, the reef "exposure" box was clicked to show that the optimal level of "intermediate" could be changed to "exposed" or "protected" were one interested in seeing how this would affect the predicted live coral cover. "Reef type" was set to "random;" one could then click "Simulate" to the right to generate simulated data from different proportions of the four reef types found in the country. This application is currently hosted on coralreefdiagnostics.com.

environmental parameters or input new data to see how the theoretical coral cover changes and could be useful for those who are already in possession of basic reef feature data but have not yet embarked on surveys. Although we believe this application/GUI will be useful for those working on coral reefs of the Solomon Islands, it is important to note some limitations and caveats. During model validation, survey data from the *GRE* were held back from the models. Then, the AI treated the held back data as “test” samples to validate the models. What this means is that the predictions were not actually ground-truthed; we did not ask our local colleagues to survey random intermediately exposed fringing reefs to see if their coral cover was higher than, for instance, protected lagoonal back reefs. This is surely the critical next step for using these, or similar, machine-learning models for making accurate predictions of the locations of the most coral-abundant regions. Whether this pattern is the same elsewhere in the Coral Triangle (or even further abroad) remains to be determined but will soon be addressed by tapping into similar datasets obtained on the *GRE* from elsewhere in the Indo-Pacific.

Acknowledgements

This work was funded by the Khaled bin Sultan Living Oceans Foundation (A.C.D. & A.B.M). Additional laboratory analyses were made possible by awards from the United States Fulbright Scholar Program and the MacArthur Foundation (A.B.M.). Funds from the Friendly Bear Editorial Service (A.B.M.) were used for purchasing of some laboratory reagents/chemicals. A.B.M. would like to thank Diedrich Schmidt for designing the machine-learning GUI that permitted the massively parallel assessment of thousands of neural networks on a personal laptop.

References

Al-Asif, A.-, A.H.M. Kamal, H. Hamli, M.H. Idris, G.J. Gerusu, J. Ismail, M.K.A. Bhuiyan, M.H. Abualreesh, N. Musa, M.E.A. Wahid & M. Mishra. 2022. Status, biodiversity, and ecosystem services of seagrass habitats within the Coral Triangle in the Western Pacific Ocean. Ocean Science Journal 57: 147–173.

Boco, S.R., J.B.P. Cabansag, E.A. Jamodiong & V.S. Ticzon. 2020. Size-frequency distributions of scleractinian coral (Porites spp.) colonies inside and outside a marine reserve in Leyte Gulf, central Philippines. Regional Studies in Marine Science 35: 101147.

Bruckner, A.W. 2015. Global Reef Expedition: Solomon Islands. Khaled bin Sultan Living Oceans Foundation, Landover MD.

Clifton, J., R. Unsworth & D. Smith. 2013. Marine Research and Conservation in the Coral Triangle. Nova Science Publishers, United Kingdom.

Cramer, K.L., A. O’Dea, T.R. Clark, J. Zhao & R.D. Norris. 2017. Prehistorical and historical declines in Caribbean coral reef accretion rates driven by loss of parrotfish. Nature Communications 8:

14160.

Ferrario, F., M.W. Beck, C.D. Storlazzi, F. Micheli, C.C. Shepard & L. Airoldi. 2014. The effectiveness of coral reefs for coastal hazard risk reduction and adaptation. Nature Communications 5: 3794.

Hochberg, E.J. & M.M. Gierach. 2021. Missing the reef for the corals: unexpected trends between coral reef condition and the environment at the ecosystem scale. Frontiers in Marine Science 8: 727038.

Huang, Y.L., A.B. Mayfield & T.Y. Fan. 2020. Effects of feeding on the physiological performance of the stony coral *Pocillopora acuta*. Scientific Reports 10: 19888.

Hughes, T.P., J.T. Kerry, M. Álvarez-Noriega, J.G. Álvarez-Romero, K.D. Anderson, A.H. Baird, R.C. Babcock, M. Beger, D.R. Bellwood, R. Berkelmans, T.C. Bridge, I.R. Butler, M. Byrne, N.E. Cantin, S. Comeau, S.R. Connolly, G.S. Cumming, S.J. Dalton, G. Diaz-Pulido, C.M. Eakin, W.F. Figueira, J.P. Gilmour, H.B. Harrison, S.F. Heron, A.S. Hoey, J.-P.A. Hobbs, M.O. Hoogenboom, E.V. Kennedy, C.-Y. Kuo, J.M. Lough, R.J. Lowe, G. Liu, M.T. McCulloch, H.A. Malcolm, M.J. McWilliam, J.M. Pandolfi, R.J. Pears, M.S. Pratchett, V. Schoepf, T. Simpson, W.J. Skirving, B. Sommer, G. Torda, D.R. Wachenfeld, B.L. Willis & S.K. Wilson. 2017. Global warming and recurrent mass bleaching of corals. Nature 543: 373–377.

Kleypas, J., D. Allemand, K. Anthony, A.C. Baker, M.W. Beck, L.Z. Hale, N. Hilmi, O. Hoegh-Guldberg, T. Hughes, L. Kaufman, H. Kayanne, A.K. Magnan, E. Mcleod, P. Mumby, S. Palumbi, R.H. Richmond, B. Rinkevich, R.S. Steneck, C.R. Voolstra, D. Wachenfeld & J.-P. Gattuso. 2021. Designing a blueprint for coral reef survival. Biological Conservation 257: 109107.

Mayfield, A.B. 2020. Exploiting the power of multivariate statistics for probing the cellular biology of thermally challenged reef corals. Platax 17: 27–52.

Mayfield, A.B. 2022. Machine-learning-based proteomic predictive modeling with thermally-challenged Caribbean reef corals. Diversity 14: 33.

Mayfield, A.B., A.W. Bruckner, C.H. Chen & C.S. Chen. 2015. A survey of pocilloporids and their endosymbiotic dinoflagellate communities in the Austral and Cook Islands of the South Pacific. Platax 12: 1–17.

Mayfield, A.B., P.H. Chan, H.M. Putnam, C.S. Chen & T.Y. Fan. 2012. The effects of a variable temperature regime on the physiology of the reef-building coral *Seriatopora hystrix*: results from a laboratory-based reciprocal transplant. Journal of Experimental Biology 215: 4183–4195.

Mayfield, A.B. & C.S. Chen. 2019. Enabling coral reef triage via molecular biotechnology and artificial intelligence. Platax 16: 23–47.

Mayfield, A.B., C.S. Chen & A.C. Dempsey. 2017. Identifying corals displaying aberrant behavior in Fiji’s Lau Archipelago. PLoS ONE 12: e0177267.

Mayfield, A.B., A.C. Dempsey & C.-S. Chen. 2019. Modeling environmentally-mediated variation in reef coral physiology. Journal of Sea Research 145: 44–54.

Mayfield, A.B. & R.D. Gates. 2007. Osmoregulation in anthozoan-dinoflagellate symbiosis. Comparative Biochemistry and Physiology A: Molecular and Integrative Physiology 147: 1–10.

Reverter, M., S.B. Helber, S. Rohde, J.M. de Goeij, & P.J. Schupp. 2022. Coral reef benthic community changes in the Anthropocene: Biogeographic heterogeneity, overlooked configurations, and methodology. Global Change Biology 28: 1956–1971.

Rodríguez-Troncoso, A.P., F.A. Rodríguez-

Zaragoza, A.B. Mayfield, & A.L. Cupul-Magaña. 2019. Temporal variation in invertebrate recruitment on an Eastern Pacific coral reef. Journal of Sea Research 145: 8–15.

Wooldridge, S.A. 2014. Assessing coral health and resilience in a warming ocean: Why looks can be deceptive. BioEssays 36: 1041–1049.