Forecasting Hawaiian Coral Reef Health: Machine Learning Models in Predicting Coral Cover from Ocean Stressors

Justin Huang

Abstract

Coral reefs, vital marine ecosystems for biodiversity and coastal protection, face increasing threats from both climate stressors and direct human impacts. Elevated sea surface temperatures, ocean acidification, and nutrient imbalances have been linked to widespread coral bleaching and declining reef resilience, alongside subsequent damage to coastal marine life which inhabit coral reefs. (Hagen 8-9) Accurate forecasting of coral reef health given these stressors is critical in informing conservation efforts. This study aims to evaluate the effectiveness of three machine learning models: linear regression, decision tree regression, and MLPRegressor (neural network) in predicting percentage coral cover based on sea surface temperature (SST), ocean pH, and nutrient levels (total dissolved phosphorous and nitrogen). Using datasets spanning 35 years (1988-2023) from the Biological & Chemical Oceanography Data Management Office (BCO-DMO) and National Oceanic and Atmospheric Administration (NOAA) hosted data from coral reef monitoring programs, these models were trained and tested on historical coral cover. Results were weak, with every model failing to effectively predict coral coverage due to limitations in data alignment and resolution. This outcome demonstrates the dependence of machine learning performance on data integrity and synchronization, emphasizing that even advanced algorithms cannot compensate for fragmented datasets. This study thus provides methodological insight for future ecological modeling, highlighting the need for temporally and spatially consistent environmental data to enable reliable coral health forecasting.

Introduction

Coral reefs are vital to the ecosystems they anchor, functioning as ecological hubs. In O’ahu alone, they provide habitat for an estimated 25% of marine species despite making up only ~0.5% of the ocean floor, sustaining fisheries that countless coastal communities use for food and livelihood (Cecchini). Their physical structure acts as a natural wavebreaker, absorbing wave energy and reducing coastal erosion (Harris). Besides their protective role, reefs also underpin Hawai’i’s largest industry: tourism (Wiener). Hawai’i’s uniquely vibrant reefs and the ecosystems that inhabit them attract divers, snorkelers, and researchers from around the world (Wiener). Perhaps most importantly, they carry deep cultural importance: in Native Hawaiian tradition, coral (*ko’a*) is regarded as the foundation of life in the sea (Gregg). This blend of ecological, economic, and cultural significance makes their protection important not only to marine biodiversity, but also to the human communities that depend on them.

Sea surface temperature (SST) is among the most immediate stressors to reefs. It reflects climate warming while also capturing acute anomalies such as marine heat waves that drive mass bleaching (Wyatt). Even small increases in baseline temperature reduce the thermal margin available to corals, amplifying the impact of short-term anomalies (Wyatt). Thermal stress destabilizes the symbiosis between corals and their zooxanthellae, the algae which provide nutrients to coral as well as some of their color, often causing bleaching, starvation, and ultimately death (Rädecker). Because such events can develop within weeks, SST serves as both a long-term climate signal and a short-term indicator of anomaly.

Ocean pH provides a parallel measure of vulnerability. Acidification reduces carbonate ion availability, impairing calcification and weakening reef structures, making them more prone to erosion (Mollica). Lower pH slows coral growth while making them more vulnerable to warming and disease, making it an important indicator of reef integrity and growth potential.

Nutrient concentrations, particularly dissolved nitrogen and phosphorus, act as proxies for eutrophication. Caused mainly by terrestrial runoff and wastewater, nutrient enrichment promotes algal blooms that outcompete corals for light and space, diminishing biodiversity (Lapointe). Because of this, nitrogen and phosphorus reflect local impacts from human activity on the shore.

To address this challenge, we evaluate three different types of machine learning algorithms, each varying in complexity and interpretability: linear regression, decision tree regression, and neural networks.

Methods

Variables

Independent variables used are SST, pH, total dissolved phosphorus, and total dissolved nitrogen from the BCO-DMO's dataset *Niskin bottle water samples and associated CTD measurements from the Hawaii Ocean Time-Series cruises from 1988-2023* (Biological & Chemical Oceanography Data Management Office, 2025). Each variable represents a known environmental stressor to coral health. The dependent variable is percent coral coverage derived from NOAA Coral Reef Assessment and Monitoring Program (CRAMP) surveys on Moloka'i, Hawai'i, conducted from 2000 to 2002. Percent coral coverage serves as a for overall coral health.

Data Preprocessing

The dependent variable BCO-DMO dataset was aligned with the NOAA survey period so that both directly overlapped, ensuring a link between predictor and response. The dependent variable percentage coral cover dataset was cleaned by removing quadrats with missing or incomplete values. Species-level percent cover values were summed to get total live coral cover per quadrant. Values were then normalized to a 0 - 100 scale in order for comparison across quadrats.

Train/test split

Because NOAA's CRAMP dataset is cross-sectional across sites rather than continuous over time, a site-level holdout method was selected for splitting training and testing data. This means that quadrats were grouped by site. Six sites were assigned to the training set, and three were reserved for testing. This prevented the model from "cheating" by seeing quadrats from the same reef in both training and testing, which would erroneously inflate performance.

Cross validation was used by repeating the split multiple times with different site combinations, which served to reduce bias from any one particular split. This design evaluates the capability of each model, trained on certain Moloka'i reefs, in generalizing to reefs it has never encountered, emulating how predictive tools might be applied to new survey locations in practice.

Modeling Approach

Three types of machine learning model were chosen to compare performance and interpretability.

Linear regression models coral cover as a weighted sum of predictors (dependent variables SST, pH, total phosphorus & nitrogen). Linear regression excels in transparency because coefficients directly represent the magnitude of each stressors impact, unlike other models where the cause of results can be hard to trace. However, this approach assumes linear, positive relationships, struggling to capture more complex interactions and subsequent trends.

Decision tree regression models split data into decision rules, essentially testing if conditions are true or false and adjusting accordingly. Decision trees are good at capturing nonlinear trends and interactions. While more flexible than linear regression, single trees are prone to overfitting data unless otherwise accounted for, in addition to their predictions being less stable.

Multilayer perceptron models (MLP models), are neural networks capable of modeling highly complex and nonlinear relationships between multiple variables. Inputs are processed through one or more layers of "neurons", which result in a model capable of detecting patterns that simpler models cannot. Their superior pattern recognition ability often makes neural networks accurate predictors. This superior predictive ability comes at the cost of interpretability. Many neural network models are described as "black box" models, in that it is extremely difficult to track the exact reasons behind predictions.

All models were trained using the quadrat-level data excluding reefs held out for testing. Model performance was evaluated using root mean squared error (RMSE), which measures average prediction error, and coefficient of determination (R^2), which measures the proportion of variance in coral cover explained by the model. In addition to numerical performance, models were qualitatively compared in terms of interpretability, ease of use, and ecological relevance.

Scripting Work

All modeling and data analysis were conducted in Google Colab, using Python (v3.10) and standard data science libraries including pandas, NumPy, matplotlib, seaborn, and scikit-learn. The script was designed to allow reproducibility through user upload of data files and automatic calculation of model performance metrics.

The input dataset was the , containing yearly-averaged environmental variables (sea surface temperature, pH, total dissolved phosphorus, and total dissolved nitrogen) from 2000–2002.

Both files were uploaded directly to the Colab.Site identifiers were standardized by assigning unique indices to each survey location.

The modeling workflow proceeded as follows. Independent variables (SST, pH, nitrogen, and phosphorus) were extracted from the merged environmental dataset, while percent coral cover served as the dependent variable. Data were split into training and testing subsets using scikit-learn's train_test_split() function. To prevent spatial leakage between sets, quadrats belonging to the same site were grouped, ensuring that no site appeared in both training and testing. This method approximated a site-level holdout, where six sites were allocated for model training and three for testing.

Three regression algorithms were implemented for comparison:
Linear Regression using Scikit-Learn's LinearRegression()
Decision Tree Regression using Scikit-Learn's DecisionTreeRegressor()

Neural Network (MLP) using Scikit-Learn's MLPRegressor() with hidden layer sizes 6, 4, and 5, 10000 max iterations, and random state 1.

Each model was trained on the same predictor matrix and evaluated on the held-out sites. Predictions were compared to observed coral cover to compute three evaluation metrics: coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). These metrics were calculated using scikit-learn's r2_score(), mean_squared_error(), and mean_absolute_error() functions, with RMSE derived as the square root of the mean squared error. All metrics were stored in a pandas DataFrame for standardized display.

Model performance results were visualized through scatter plots of predicted versus observed coral cover generated with matplotlib. Each graph included a red dashed reference line ($y = x$) indicating a perfect prediction, against which model performance could be visually assessed. Annotation boxes containing each model's R^2, RMSE, and MAE were overlaid on the plot for clarity.

All code cells were executed sequentially within a single Colab notebook. Output tables, metrics, and figures were exported directly from the runtime environment. The notebook structure allows full reproducibility, provided that identical datasets are uploaded and cell order is maintained.

Results

All three models were trained using the selected environmental variables as predictors and percent coral cover as the response. Model performance was evaluated using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The neural network produced the highest R^2 value (0.301), while the decision tree and linear regression performed slightly lower (0.284 and 0.169, respectively). Error values were similar across models, with RMSE ranging from 23.881 to 25.198, and MAE from 20.670 to 21.247.

Table 1: Performance metrics for each machine learning model

Model	R^2	RMSE	MAE
Decision Tree	0.284	25.032	21.127
Linear Regression	0.169	23.881	20.670
Neural Network	0.301	25.198	21.247

True % Coral Cover vs. Predicted % for Decision Tree Regression, Linear Regression, and MLP Regression

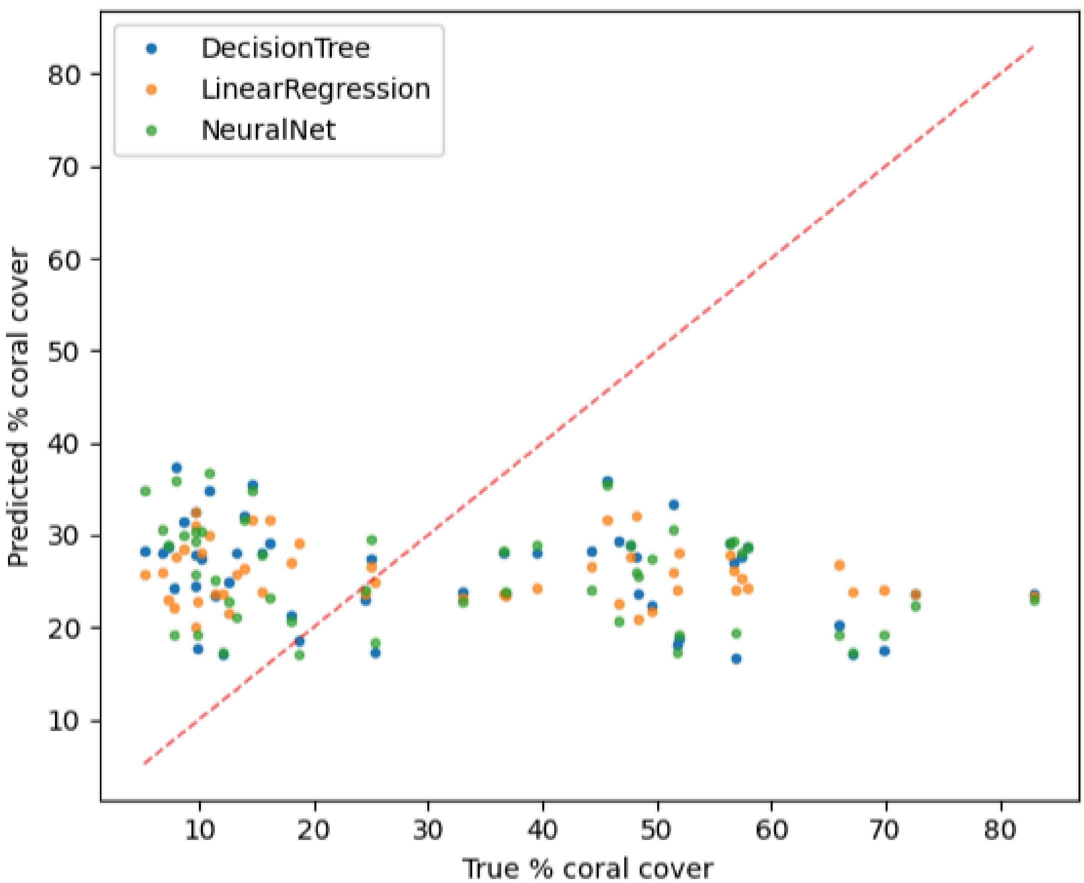

Fig 1: Graph of performance of three machine learning models. Red dotted line is hypothetical perfect result

Discussion

Results indicate that none of the three models were able to capture a strong predictive relationship between the selected environmental variables and coral cover. R^2 values remained low across all models, with the neural network's best performance being only 0.301. Error values were consistently high, suggesting that the models produced predictions only marginally better than chance.

These poor results are largely attributable to issues in data rather than the algorithms themselves. The datasets used were misaligned in both temporal and spatial resolution, meaning that independent variables (such as SST, pH, and nutrient levels) did not correspond correctly with the quadrat-level coral cover measurements. The models failed to establish a connection due to there being little to none existing in the first place.

Although these outcomes fall short of demonstrating predictive power, they embody the importance of careful dataset selection and preprocessing in modeling. Model performance is fundamentally constrained by data quality and alignment. Future studies should prioritize sourcing datasets collected within the same timeframes and locations. By addressing these limitations, it is possible to fully utilize the predictive potential that the tested algorithms can provide.

In an ideal context, the models tested in this study would be capable of producing meaningful predictions. Each algorithm possesses the theoretical capacity to model complex nonlinear relationships between environmental stressors and coral health. However, this potential can only be realized when the underlying data are both comprehensive and concurrent. A robust modeling framework would require that all predictor variables (sea surface temperature, pH, total dissolved nitrogen, and total dissolved phosphorus) be collected at the same locations and times as the dependent variable, percent coral cover. Such synchronization would eliminate temporal and spatial mismatches, minimize confounding effects from unmeasured local factors, and allow the models to fit to the true relationship rather than broad correlations. Under these conditions, it is likely that one or more of the evaluated models could achieve strong predictive performance, demonstrating that the primary limitation lies not in the algorithms themselves, but in the fragmented nature of current public data.

Limitations and Challenges

This study faced several limitations that constrained the usability of the results. The most significant challenge was datasets. The independent environmental variables and dependent coral coverage were not collected on matching spatial or temporal scales. Without alignment points, the predictors and response were not synchronized, greatly reducing the ability of the models to capture any meaningful relationships.

Another limitation was data coverage. While the environmental datasets spanned a broad temporal range, coral cover observations were limited to specific locations and far fewer instances. This is because coral coverage is far more complex to measure than metrics like pH, temperature, or nutrient levels, requiring more complex procedures, equipment, and more time. Ultimately, this resulted in a large disconnect between the data collected for use in predicting, and that used in validating response, which pre-processing could not fix.

All of the previous issues amplified the weaknesses of machine learning models. The small number and lack of detailed data on coral coverage dates led to overfitting and an incredibly low initial predicted range, only becoming slightly useful with heavy processing to the inputted datasets in order to try and mend the lack of overlap in independent and dependent data.

Conclusion

This study evaluated the potential of three machine learning algorithms (linear regression, decision tree regression, and a multilayer perceptron neural network) in predicting coral cover from key environmental stressors. While none of the models achieved strong predictive accuracy, their performance reveals a central constraint of data-driven ecological forecasting: even robust algorithms cannot compensate for fragmented or misaligned data. Results emphasize the importance of temporal and spatial coherence in environmental datasets.

Future work should prioritize the coordinated collection of biological and oceanographic data at shared sites and timescales. By integrating high-resolution, concurrent measurements of sea surface temperature, pH, and nutrient levels with coral cover observations, it will become possible to evaluate these algorithms under conditions that truly reflect their predictive capacity. With improved data alignment and scale, the methods tested here could provide valuable tools for monitoring and forecasting coral reef health in a changing climate.

References

[1] Hagen, Rachel. *Underwater and Underrated: Coral Reefs and Climate Change*. American Security Project, 2018. *JSTOR*, http://www.jstor.org/stable/resrep19816. Accessed 14 Aug. 2025.

[2] McClanahan, Timothy R., et al. "Effects of Climate and Seawater Temperature Variation on Coral Bleaching and Mortality." *Ecological Monographs*, vol. 77, no. 4, 2007, pp. 503–25. *JSTOR*, http://www.jstor.org/stable/27646104. Accessed 14 Aug. 2025.

[3] Venn, Alexander A., et al. "Impact of Seawater Acidification on pH at the Tissue–Skeleton Interface and Calcification in Reef Corals." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 5, 2013, pp. 1634–39. *JSTOR*, http://www.jstor.org/stable/41992092. Accessed 14 Aug. 2025.

[4] LESSER, MICHAEL P. "Eutrophication on Coral Reefs: What Is the Evidence for Phase Shifts, Nutrient Limitation and Coral Bleaching." *BioScience*, vol. 71, no. 12, 2021, pp. 1216–33. *JSTOR*, https://www.jstor.org/stable/27229436. Accessed 14 Aug. 2025.

[5] Cecchini, Pansa et al. "Saving coral reefs: significance and biotechnological approaches for coral conservation." Advanced biotechnology vol. 2,4 42. 22 Nov. 2024, doi:10.1007/s44307-024-00049-1

[6] Harris, Daniel L., and Ana Vila-Concejo. "Wave Transformation on a Coral Reef Rubble Platform." *Journal of Coastal Research*, 2013, pp. 506–10. *JSTOR*, http://www.jstor.org/stable/26482020. Accessed 9 Sept. 2025.

[7] Wiener, Carlie S., Mark D. Needham, and Paul F. Wilkinson. "Hawaii's Real Life Marine Park: Interpretation and Impacts of Commercial Marine Tourism in the Hawaiian Islands." *Current Issues in Tourism*, vol. 12, no. 5–6, 2009, pp. 457–474. *JSTOR*, doi:10.1080/13683500903042880.

[8] Gregg, T. M. "Puka Mai He Koʻa: The Significance of Corals in Hawaiian Culture." *Ethnobiology Letters*, vol. 7, no. 1, 2016, pp. 19–25. *JSTOR*, doi:10.14237/ebl.7.1.2016.698.
Wyatt, A.S.J., Leichter, J.J., Washburn, L. et al. Hidden heatwaves and severe coral bleaching linked to mesoscale eddies and thermocline dynamics. Nat Commun 14, 25 (2023). https://doi.org/10.1038/s41467-022-35550-5

[9] Rädecker, Nils, et al. "Heat Stress Destabilizes Symbiotic Nutrient Cycling in Corals." Proceedings of the National Academy of Sciences, vol. 118, no. 19, 2021, e2022653118.
Mollica, N. R., et al. "Ocean Acidification Affects Coral Growth by Reducing the Availability of Carbonate Ions." Proceedings of the National Academy of Sciences, vol. 115, no. 8, 2018, pp. E1734–E1742.

[10] Lapointe, Bruce E., et al. "Sewage Pollution in Negril, Jamaica: Effects on Nutrition and Ecology of Coral Reef Macroalgae." Chinese Journal of Oceanology and Limnology, vol. 29, no.

2, 2011, pp. 289–302.