**Introduction**

As the average world temperature increases, scientists have become increasingly interested in the interaction between humanity and the corresponding changes in the natural world. Specifically, the world’s oceans are projected to play a major role in maintaining biodiversity, regulating the climate, and sustaining a healthy global economy that contributes to food security worldwide (Gattuso et al., 2018). In order to monitor the health of these key aquatic ecosystems, the NASA Data Science Team (DST) is investigating the capacity of the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite to infer the vitality of coral reefs around the world. Under DST, the Geophysical Observations Toolkit for Evaluating Coral Health (GOTECH) project seeks to use machine learning models to interpret, from CALIPSO imagery, vitality properties upon satellite pass.

The GOTECH team identified four technical areas that must be solved to provide a tool that meets the requirements. The first task is to combine multiple open-source satellite databases into a truth-source data set. Next, this data set will be time-aligned and geo-aligned with existing CALIPSO data. The third step is to then correlate imagery from the truth source dataset to describe the coral reef health in the CALIPSO dataset. Finally, statistical techniques will be applied to attain the accuracy of these coral reef predictions which will also define the success of the GOTECH team.

**Background**

The industry standard for this type of overhead classification analysis is spectral analysis of IR data (Joyce & Phinn, 2013). Spectral analysis is used often to not only identify live coral, but also the relative health of the coral and the abundance of what type of coral is present. This is done through the examination of the spectral reflection of infrared radiation (IR) collected from overhead satellites. Different bands of radiation will be present based on the proportion of living coral and the species of coral that are present. By applying deep learning principles, researchers have been able to train models to identify the health of coral reefs using this spectral data (Collin & Planes, 2012).

In the past, many researchers believed that extracting information on bleached corals using satellite imagery was infeasible or extremely difficult due to its similar spectroscopy to sand (Elvidge et al., 2004). However, Xu et al. (2015) was able to build successful models utilizing data from the MultiSpectral Instrument (MSI) of the Sentinel 2 satellite maintained by the European Space Agency. Through extensive research that relied heavily on the work done by Xu et al. (2015), the GOTECH team determined that the optimal data to train the neural models will be IR band data centered on the 532 nm wavelength with spatial resolution of 30-60 meters.

While focusing on IR data, Xu et al. (2015) identified the IR band centered on 492.4 nm as best for identifying both the location of the coral and its health there. Unfortunately, CALIPSO CALIOP does not have the same capabilities as the Sentinel 2 MSI, and the GOTECH team will use data centered on the 532 nm wavelength since it is the closest data available. In addition to CALIPSO data the GOTECH team found data from Allen Coral Atlas, Florida Bleach Watch, NOAA CoastWatch Degree Heating Week (DHW), NASA Giovanni, World Conservation Monitoring Centre (WCMC), Reefbase, and NASA CALIPSO which were suitable to combine into a feasible dataset. Lastly, the team chose the oceanic area around the state of Florida as a focused use-case in order to prove our methodology and models without being inundated by excess data. An example of the CALIPSO data is below in Figure 1 (Winker, 2021). The goal in finding data was to build a dataset which would match known coral locations with the reflectance data at different altitudes displayed in the image.

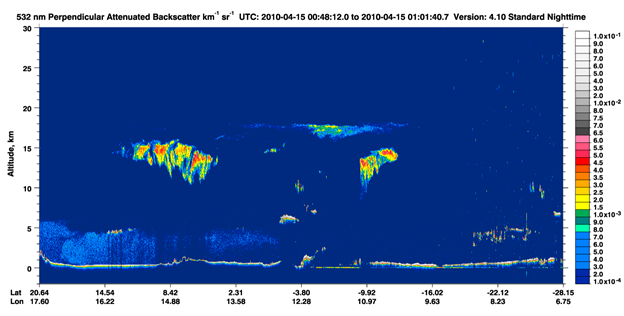


Figure 1: Example of CALIPSO CALIOP Perpendicular Attenuated Backscatter 532 nm Image

**Analysis Approach**

**CALIPSO Data**

Based off our research, we decided to use IR band data centered on the 532 nm wavelength with spatial resolution of 30-60 m from CALIPSO. We then subsetted the data based on location, time, and altitude. For each CALIPSO observation, there is a specific point defined by a latitude and longitude with measurements at different altitudes. To build the training data, we used 300 different reflectance measurements based off altitude and depth. When combined with the ground truth data, each CALIPSO measurement location had 300 reflectance measurements, a latitude and longitude marker, a class label as “Coral” or “Other”, and then the associated vitality data discussed below.

**Ground Truth Data**

The ground truth data is based on known coral locations in addition to reported growth and decay events. The datasets from Allen Coral Atlas, Florida Bleach Watch, NOAA CoastWatch DHW, WCMC, and Reefbase contain location, timeframe, and vitality information for coral restoration and bleaching. To start, we used the CALIPSO data points to filter all the other datasets by location and timing. By doing this, we were able to take a CALIPSO measurement at a specific time and location and then match it to observations from all the other datasets at a similar time and location. What this ultimately allowed was for us to assign a class label to each measurement which could be used in modelling.

Next the team focused on matching the labeled CALIPSO data to coral vitality data. Often, coral reef restoration practitioners consider numerous environmental and physical parameters that have varying temporal effects on coral vitality (Ladd et al., 2018). Short-term factors include infrared radiation and degree heating weeks, while season factors include chlorophyll, photosynthetically available radiation, and total suspended matter. The frequency of additional parameters will be collected based on its temporal effect on coral vitality (Table 1).

| TABLE 1 Reef restoration parameters | | |
| --- | --- | --- |
| Data Source | Parameters | Temporal Frequency |
| NOAA Coastwatch | Water Temperature  Degree Heating Weeks  Chlorophyll  PAR  Bathymetry | Weekly  Weekly  Monthly  Monthly  Single Time Period |
| GIOVANNI | Infrared radiation  Total suspended matter  PAR  Particulate organic | Weekly  Monthly  Monthly  Monthly |

Typically, observations would not match exactly by either location or time. To circumvent this sparsity, the team determined a coral restoration and decay timeline dependent on the documented timeframe from the ground truth data with additional constraints from CALIPSO data. CALIPSO data became available in June 2006 which sets a lower bound on our timeline for the ground truth data. Since June 2006, there have been over 180 coral restoration projects documented with known temporal scales, most lasting 12 - 24 months. Reefbase has over 400 bleaching events reported, but only 7 events with known bleaching periods. Since bleaching may be noticeable on coral reefs anytime between 1-3 weeks after first observation, we collected data several months prior to the date of report to establish a timeline of decay. NOAA is another source of bleaching reports for coral reefs in Florida. This data contains reports for a six-month period from 2015 to 2020. Although these data provide labels for training our model and performance inferences, they are incomplete as there are missing data and key attributes that need to be collected from other sources that may improve our classification performance.

By creating these rulesets and determining coral vitality close to the same time period of each CALIPSO observation, we created a dataset that was used to determine coral vitality at predicted coral locations over different time periods. By first building our coral ground truth and then matching it to coral vitality data, the team built a database that could be used to create models that implement only CALIPSO data for predictions but are augmented by other data to inform the user of the coral health at that location and time.

**Predicting Coral Locations**

The GOTECH team applied several different models to binary classify the geographic locations as “Coral” or “Other”. Per NASA instruction, each model could only implement CALIPSO reflectance data for Spectral Analysis, and the team relied heavily on prior research to utilize the optimal spectral band and altitudes for coral reef identification. Overall, the team applied Random Forest, Logistic Regression, Feed Forward Neural Networks, and One-Dimensional Convolutional Neural Network (1D CNN) frameworks to get binary prediction labels.

Another alternative for coral health identification was to apply image segmentation to the IR dataset. Image segmentation is a subset of deep learning where photos are split into polygons of similar classes. An example would be identifying all the faces in a crowd of people or marking the individual cells of an image taken under a microscope. The backbone of the method was the U-net CNN generated by b et al. (2015). As the name implies, this CNN employs a U-shaped design where resolution is decreased to a selected parameter value and then built back up to the original resolution. The output is an image with the same resolution as the original but with the polygons of unique classes identified. Figure 2 below taken from Ronneberger et al. (2015) displays this structure.

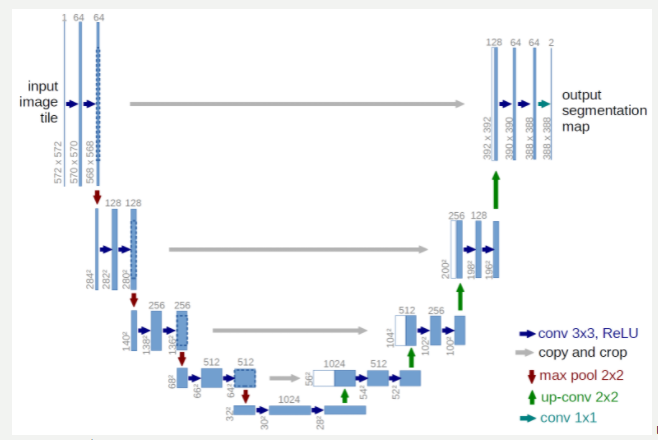
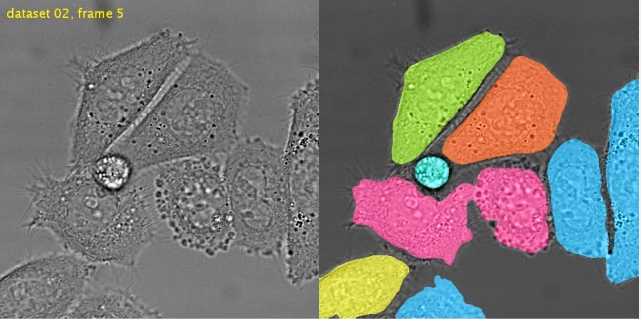
 

Figure 2: Visual Depiction of the U-net CNN (Ronneberger et al., 2015) on the left. Example Image Segmentation Classification (Ronneberger et al ,2015) on the right

**Interpreting Global Trends**

Once the best performing model was identified, the GOTECH team created a visualization to allow for ease of interpretation. The visualization is an interactive, time-series chart that shows every CALIPSO value in the Florida use-case collected from 2006 to now and the associated label the model assigned to that location. With this visualization, the NASA team can now focus on areas of interest and quickly sift through informative imagery to identify trends.

**Results**

The team was able to generate a comprehensive fused dataset of eight well-known coral repositories. The fused dataset contains 337 variables for over 43k locations which includes CALIPSO reflectance measurements and coral vitality data that are within 5km of confirmed class labels. This dataset was utilized to train several different models on the features including a Feed Forward Neural Network, 1D CNN, Logistic Regression, and Random Forest.

Per the NASA prompt, the models must rely only on CALIPSO reflectance features, but the team was able to explore different distance thresholds for the proximity of features to the ground truth labels. In addition to the distance thresholds, the team also performed feature selection and identified 23 (of the 300 total) CALIPSO features that appeared to have the best chance of improving model performance. Image 2 shows a plot of the first 100 CALIPSO features with circles on locations that appear to have a strong statistical separation between classes. Using this methodology, the team selected CALIPSO features 0-10, 205-210 and 220-225.



Image 2: No Statistical Difference in Reflectance Values for Each Class

The results for each model and distance threshold can be found in Appendix A, but the best performing model was the 1D CNN with all 300 CALIPSO features, distance threshold of 1000m, and an accuracy of 75.84%. The parameters for this model include: dropout rate of 0.3, adam optimizer, loss function of binary cross entropy, batch size of 16, 100 epochs, a validation set of 20%, and kernel size of 5. Lastly, the smaller CALIPSO feature set actually decreased model performance by an average of 6.3% across all 4 models.

Contrary to the prompt, the team next performed analysis to determine if there were any features in the Giovanni dataset that could improve model performance. To accomplish this goal, the team selected the variables Photosynthetically Available Radiation (PAR), Chlorophyll a, Inorganic Particulate, and Organic Particulate to augment the models. Using similar CALIPSO feature and distance threshold measurements, the team repeated the modeling. This time, the Feed Forward Neural Network with all 300 CALIPSO features, additional Giovanni parameters, and a distance threshold of 1000m performed best with an accuracy of 77.67%. Once again, the results for all these models are in Appendix A.

At this point, the team was unable to train a model that is more accurate than the 77.67% represented by Feed Forward Neural Network. However, the team was able to successfully predict coral health from the coral vitality data. Specifically, data from the NOAA CoastWatch Report provided comprehensive information on coral stressors which the team used to report on coral status. Implementing a time-series animation, the results were visualized in a way that allows a user to explore trends in the data.

Lastly, the team was unable to implement the U-net model due to limitations in the dataset. The CALIPSO data is more sparse than originally expected and does not provide the necessary fidelity to create the imagery used as a label in this model framework. Image 1 below displays how the CALIPSO data is bound by the unique ground track of the satellite in orbit, and for a given orbit there are not enough measurements to apply the model. By looking at the image, it is clear to see the linear ground track and also observe the large amount of missing data. At higher resolutions, the problem appears worse.

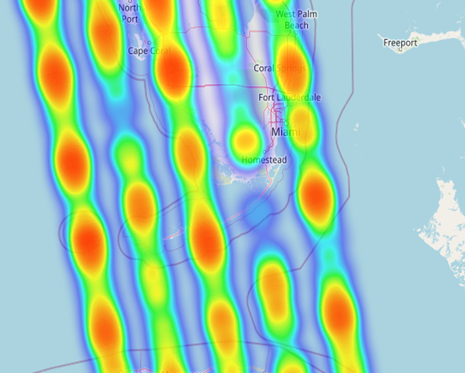


Image 1: Heat Map Depiction of Sparse CALIPSO data

**Future Work**

For future work, the GOTECH team explored applying a more robust version of IR data for U-net modeling. Both the European Space Agency (ESA) Sentinel and NASA MODIS-Aqua satellites appear to provide more comprehensive reflectance data that could be combined with our ground truth polygons to train modeling. With imagery, convolutional neural networks (especially the U-net) have proven to be accurate and successful. The baseline set by this team will hopefully set the groundwork for future work with this model and a different dataset.

Next, the team did not implement any type of data processing prior to extracting CALIPSO data. Reflectance data can be affected by sun angle, cloud cover, and even humidity. While the NASA team provides their own version of processing, the ESA provides a tool named “ACOLITE” that could be applied to future datasets to hopefully improve results. In a research paper by Xu et al. (2015), the authors discussed applying dark spectrum fitting (DSF) from the ACOLITE model by Vanhellemont & Ruddick (2014, 2015, 2016). They also utilized several ground control points (GCPs) from Google Earth to perform geometric corrections and georeferencing in the images. In addition, brightness in a near IR band could be utilized to deglint the visible wavelength bands based on the linear relationships between near IR and visible bands (Hedley et al., 2005). Finally, pixels containing boats, whitecaps (sea foam), clouds and their shadows, and land could be masked in the imagery (Gapper et al., 2019). In general, future work implementing these techniques may see an increase in model accuracy.

**Conclusion**

Overall, the GOTECH team was unable to train a model with accuracy greater than 75.84% following the NASA requirements of utilizing CALIPSO IR data alone. The team assembled a fused dataset with 337 variables and over 43k observations that can reliably report coral vitality data but fell short on applying that data to predicting coral locations. While several models were tested, the One-Dimensional Convolutional Neural Network utilizing data with a distance threshold of 1000m and parameters of a dropout rate of 0.3, adam optimizer, loss function of binary cross entropy, batch size of 16, 100 epochs, a validation set of 20%, and kernel size of 5 performed best at 75.84% accuracy. The team also visualized the results from the models’ predictions and provided information for coral location and vitality that can be used to understand trends over time in the data. While not a complete success, this study provided a good framework which will hopefully inform future success in subsequent research.

**Appendix A**

| Model | Full Model Accuracy | 23 Features Accuracy | Full + Giovanni | 23 Features + Giovanni |
| --- | --- | --- | --- | --- |
| Feed Forward Neural Network | 68.63 | 65.51 | 76.16 | N/A |
| 1D CNN | 75.31 | 66.78 | N/A | N/A |
| Logistic Regression | 53.65 | 49.72 | 66.27 | 56.27 |
| Random Forest | 60.75 | 58.55 | 69.17 | 66.68 |

Table 2: Model Performance (in %) for 10m Distance Filter

| Model | Full Model Accuracy | 23 Features Accuracy | Full + Giovanni | 23 Features + Giovanni |
| --- | --- | --- | --- | --- |
| Feed Forward Neural Network | 72.85 | 66.40 | 74.82 | N/A |
| 1D CNN | 75.38 | 69.49 | N/A | N/A |
| Logistic Regression | 51.11 | 51.11 | 64.61 | 53.60 |
| Random Forest | 59.88 | 58.07 | 69.41 | 68.94 |

Table 3: Model Performance (in %) for 50m Distance Filter

| Model | Full Model Accuracy | 23 Features Accuracy | Full + Giovanni | 23 Features + Giovanni |
| --- | --- | --- | --- | --- |
| Feed Forward Neural Network | 73.14 | 68.65 | 77.67 | N/A |
| 1D CNN | 75.84 | 71.98 | N/A | N/A |
| Logistic Regression | 48.68 | 48.68 | 63.96 | 63.96 |
| Random Forest | 61.74 | 51.97 | 68.88 | 65.22 |

Table 4: Model Performance (in %) for 50m Distance Filter

**Works Cited**

Collin, A., & Planes, S. (2012, October 23). *Enhancing coral Health detection using Spectral Diversity indices from WORLDVIEW-2 imagery and machine learners*. MDPI. Retrieved from <https://www.mdpi.com/2072-4292/4/10/3244/htm>.

Douglas, A. E. (2003, March 20). *Coral bleaching––how and why?* Marine Pollution Bulletin. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0025326X03000377>.

Flood, N., Watson, F., & Collett, L. (2019, June 28). *Using a U-NET convolutional neural network to Map woody VEGETATION extent from high resolution satellite imagery across Queensland, Australia*. International Journal of Applied Earth Observation and Geoinformation. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0303243419302041>.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018, December 7). *AI4People - an ethical framework for a good AI Society: Opportunities, Risks, principles, and recommendations*. SSRN. Retrieved from <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3284141>.

Gattuso, J.-P., Magnan, A. K., Bopp, L., Cheung, W. W. L., Duarte, C. M., Hinkel, J., Mcleod, E., Micheli, F., Oschlies, A., Williamson, P., Billé, R., Chalastani, V. I., Gates, R. D., Irisson, J.-O., Middelburg, J. J., Pörtner, H.-O., & Rau, G. H. (2018, October 4). *Ocean solutions to address climate change and its effects on marine ecosystems*. Frontiers. Retrieved from <https://www.frontiersin.org/articles/10.3389/fmars.2018.00337/full>.

Joyce, K. E., & Phinn, S. R. (2013, February 5). *Spectral Index Development for mapping live coral cover*. SPIE Digital Library. Retrieved September 13, 2021, from <https://www.spiedigitallibrary.org/journals/journal-of-applied-remote-sensing/volume-7/issue-01/073590/Spectral-index-development-for-mapping-live-coral-cover/10.1117/1.JRS.7.073590.full?SSO=1>.

Ladd, M. C., Miller, M. W., Hunt, J. H., Sharp, W. C., & Burkepile, D. E. (2018, April 4). *Harnessing ecological processes to facilitate coral restoration*. The Ecological Society of America. Retrieved September 13, 2021, from <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/fee.1792>.

Roelfsema, C. M., & Phinn, S. R. (2017). Spectral reflectance library of healthy and bleached corals in the Keppel Islands, Great Barrier Reef

*Remote sensing of bottom reflectance and water attenuation parameters in shallow water using aircraft and Landsat data*. Taylor & Francis. (n.d.). Retrieved from <https://www.tandfonline.com/doi/abs/10.1080/01431168108948342>.

Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). *U-Net: Convolutional networks for biomedical image segmentation*. arXiv.org. Retrieved from <https://arxiv.org/abs/1505.04597>.

Tang, Y. (2015, February 21). *Deep learning using linear support vector machines*. arXiv.org. Retrieved from <https://arxiv.org/abs/1306.0239>.

Vanhellemont, Q., & Ruddick, K. (2014, February 25). *Turbid wakes associated with offshore wind turbines observed with landsat 8*. Remote Sensing of Environment. Retrieved September 13, 2021, from <https://www.sciencedirect.com/science/article/pii/S0034425714000224?via%3Dihub>.

Vanhellemont, Q., & Ruddick, K. (2015, March 4). *Advantages of high quality swir bands for Ocean Colour Processing: Examples from landsat-8*. Remote Sensing of Environment. Retrieved September 13, 2021, from <https://www.sciencedirect.com/science/article/pii/S0034425715000577?via%3Dihub>.

Vanhellemont, Q. (2019, March 13). *Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the Landsat and Sentinel-2 Archives*. Remote Sensing of Environment. Retrieved September 13, 2021, from <https://www.sciencedirect.com/science/article/pii/S0034425719301014>.

Xu, J., Zhao, J., Wang, F., Chen, Y., & Lee, Z. (1AD, January 1). *Detection of coral reef bleaching based on sentinel-2 multi-temporal imagery: Simulation and case study*. Frontiers. Retrieved from <https://www.frontiersin.org/articles/10.3389/fmars.2021.584263/full>.

Winker, D. (2021, June 10). *Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations*. NASA. Retrieved September 13, 2021, from <https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/browse/index.php>.