NASA/TM–20220010955

Geophysical Observations Toolkit For Evaluating Coral Health (GOTECH): Fall 2021 Final Report

Newton H. Campbell and Douglas M. Trent
Science Application International Corporation, Hampton, Virginia

July 2022

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/
and select the "General" help request type.

NASA/TM–20220010955

Geophysical Observations Toolkit For Evaluating Coral Health (GOTECH): Fall 2021 Final Report

Newton H. Campbell and Douglas M. Trent
Science Application International Corporation, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center

Hampton, Virginia 23681-2199

July 2022

The NASA Langley Research Center (LaRC) Data Science Team (DST), under the Office of the Chief Information Officer (OCIO), is investigating the capacity of the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite to infer the vitality of coral reefs. This report describes the Fall 2021 period of performance for the Geophysical Observations Toolkit for Evaluating Coral Health (GOTECH) project. During this effort, two student teams at Georgia Tech developed machine-learning models to predict the vitality of coral reefs in targeted geographic regions based on backscatter data from the CALIPSO satellite. To train these models, students fused data to form a common operating picture of how coral reefs have grown and decayed worldwide. This report describes the student assignment, background, and results of the semester's research.

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148

NASA Langley Research Center

Hampton, VA 23681-2199

Fax: 757-864-6500

Contents

1. Executive Summary

The world's coral reefs face many threats, from global climate change causing warming and more acidic oceans to pollution and unsustainable fishing practices.[1] The US Government has demonstrated interest in the monitoring and reconstitution of coral reefs. While many have set out to observe various properties of reefs on Earth, there is no global, real-time operating picture available to prioritize dying reefs.[2] At best, public reef databases provide, with varying certainty, a snapshot of reef health at specific locations and specific points in time.[3][4] In short, the problem is that none of the existing coral reef databases alone capture enough data to make prioritization feasible with quantifiable certainty.

Consistent with NASA's statutory responsibility to "provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof," [5] the NASA Langley Research Center (LaRC) Data Science Team (DST), under the Office of the Chief Information Officer (OCIO), is investigating the capacity of the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite to infer the vitality of coral reefs. The Geophysical Observations Toolkit for Evaluating Coral Health (GOTECH) project seeks to use machine learning models to infer reef vitality properties from CALIPSO satellite imagery. The key to GOTECH's success is using advanced data fusion algorithms to merge data from public reef databases into a dataset appropriate for machine learning.

The health of coral reefs is essential to many ecosystems. Unfortunately, reporting about bleaching and other vitality indicators typically rely on human sightings and manual data entry. This information is limited at best and inaccurate or erroneous at worst. Satellites such as GOES-R and NOAA-20 have demonstrated utility in inferring bleaching activity. While this data has proved useful, NASA seeks to further leverage its mission capability by bringing space-based LIDAR to bear on the problem. ICESat-2 is one mission that has begun to explore this capability. In GOTECH, we investigated CALIPSO as an alternative, uncharacteristic satellite for this exploration. The project is motivated by the notion that NASA can leverage unintended space instruments to address climate-related missions.

Over a 14-week semester, two teams of three students from the Georgia Institute of Technology (Georgia Tech) collaborated to develop a combined database for coral reef vitality properties and then leverage this database for machine learning. Each student team incorporated public data sources and demonstrated prediction from (solely) CALIPSO imagery on at least two machine learning models. The data fusion process was repeatable and extendable to incorporate other data sources. NASA guided the students in this effort in partnership with subject matter experts from Coral Vita. This startup company uses revolutionary methods to reconstitute dying and damaged coral reefs through terrestrial-based farms. The machine learning experts from NASA and lead scientists from Coral Vita coordinated weekly meetings with students to shape research questions and guide the student teams to their final implementations. The project concluded with two Final Reports regarding the final datasets and a description of their utility in machine learning. This Final Report captures the background of the Fall 2021 GOTECH project, preparation for the student semester, summaries of key findings, and lessons learned.

2. Project Participants

2.1. Represented Parties

NASA Langley - NASA's Langley Research Center (LaRC) has a long history of conducting Earth Science research, using a variety of satellites to image the land, oceans, and atmosphere. Data from satellite imagery, combined with data from open-source reef databases, has the potential to inform coral reef vitality in known and unexplored areas.

OCIO Data Science Team – Within Langley's Office of the Chief Information Officer (OCIO) is a Data Science team that provides Data Science consulting expertise to Principal Investigators (PI) and subject matter experts (SME) on their research projects. The Data Science team's support of PI and SME customers, primarily at Langley initially, has recently expanded to include other centers as the agency transforms to a One NASA enterprise. In the case of the GOTECH project, the SME customer is an industry partner named Coral Vita.

Coral Vita - Coral Vita was founded by ecological entrepreneurs to restore coral reefs by growing replacement coral in land-based farms. The Bahamas-based company uses advanced techniques to grow coral up to 50x faster, while boosting resiliency to warming and acidifying oceans. The hardier land-grown coral is outplanted to degraded reefs to bring them back to life. Coral Vita is one of five global winners of the first-ever Earthshot grand prize of £1 million, awarded in 2021 by Prince William, Duke of Cambridge.

Georgia Tech - The Georgia Institute of Technology is ranked in the Top Ten nationwide by US News & World Report for its Industrial Systems & Engineering (ISyE) graduate program (#1), its Statistics & Operational Research (#8), its Business, Quantitative Analysis (#6) and its college of Computer Science (#9). Georgia Tech's Masters of Science in Analytics program blends the strengths of the three colleges -- ISyE, Business and Computing – to produce graduates with the interdisciplinary skills needed to obtain deep insights into analytics problems.

The MS Analytics program includes a one-semester applied analytics Practicum as a graduation requirement. During the Practicum, small teams of students work on impactful graduate-level data science projects submitted by industry.[12] The Langley OCIO Data Science team has previously collaborated with Georgia Tech on its MS Analytics practicum and selected this program to execute the GOTECH project.

National Institute of Aerospace - The National Institute of Aerospace (NIA) was created by NASA's Langley Research Center in 2002 as a nonprofit research, graduate education, and outreach institute located in Hampton, VA. NIA collaborates with NASA, other government agencies and laboratories, universities, and industry to conduct leading-edge research and technology development in space exploration, aeronautics, and science.

2.2. Key Personnel

Project Leadership:

Dr. Newton Campbell	NASA Langley Research Center	NASA Project Lead
Mr. Douglas Trent	NASA Langley Research Center	NASA Program Manager
Dr. Katey Lesneski	Coral Vita	Director, Restoration Science

Student Research Teams:

GT Student Team 1	Tina Guo
	Josh Mattingly
	Dan Schauder
GT Student Team 2	Kareem Naguib
	Quinn Stank
	Andrew Wang

Special thanks to other contributors that made the project successful through lending expertise and coordination:

NASA	Patrick Geitner	Langley Data Science Intern
	Ed McLarney	NASA AI/ML Lead, Langley DS Lead
	Shan Zeng	Langley Earth Sciences Research
Coral Vita	Joe Oliver	Director of Restoration Operations
	Sam Teicher	Chief Reef Officer
Georgia Institute of Technology	Ann Blasick	Corporate Relations Manager
	Mariana Campili-Warren	MS Analytics Instructional Lead
	Renata Haque	Practicum Teaching Assistant
	Joel Sokol	Director, MS Analytics Program
National Institute of Aerospace	Carly Bosco	Director, Research Programs
	Shi Broadwell	Deputy Program Mgr for Langley
	Douglas Stanley	President & Executive Director

3. Background

Coral reefs are one of the most important ecosystems on the planet. Healthy coral reef ecosystems are crucial to marine life as a means of maintaining marine biodiversity and to humans as a source of food, medical advances, and tourism revenue.[6] Healthy coral reefs also "protect shorelines from storm and wave damage and form barriers that provide safe passage for shipping." [7]

Although coral reefs only occupy 1% of the Earth's surface, they are home to 25% of all marine species.[8] These incredible ecosystems are also a food source for hundreds of millions of people. Reefs power coastal economies worldwide through tourism, fishing, and recreation, and they shelter coastlines from storms and erosion. The total economic value of coral reefs' direct and indirect use exceeds an estimated $375 billion annually.[9]

Unfortunately, coral reef health is collapsing around the world. An extensive study published in 2021 shows that Earth has lost half of its coral reefs since the 1950's.[10] Reefs are under tremendous stress from seawater acidification, global warming, and a range of human activity, including tourism, overfishing, pollution, and coastal development. As reefs die, these ecological wonders and their critical benefits to humans disappear.[11]

Due to their high sensitivity to changes in the environment, coral reefs are leading indicators of adverse changes to our world. They are the "canary in the coal mine," alerting us to future threats to the well-being of our planet. Therefore, large-scale continuous monitoring of coral reef health needs to be undertaken now. At present, coral health surveys (commonly conducted by a motorboat and towed-diver) are limited in the area they can cover, and the frequency of repetition. Much more scalable and automated methods of monitoring coral health are required.

3.1. Earth Sciences Instrumentation

For NASA, remote sensing from space began in 1960 with the launch of the Television Infrared Observation Satellite, TIROS-1. Since then, NASA alone has launched approximately 75 satellites with unique instruments to observe the Earth in various ways. The NASA/USGS Landsat fleet alone extends back to 50 years of Earth observation[13]. Over that time, computer-based digital data processing from remote sensing instruments led to the development of quantitative assessments and, subsequently, quantitative indices captured in public databases. In recent years, NASA and other organizations have consolidated these databases and the access to them through APIs that increase their utility to the public. For example, Earthdata [14] from NASA Goddard Space Flight Center (GSFC) provides APIs and coding recipes for the NASA Earth Observing System Data and Information System, which contains 30 years of NASA Earth Sciences measurements. Databases such as the National Snow and Ice Data Center (NSIDC) [15] consolidate remote sensing data relevant to specific Earth Science fields of study. Databases such as these help NASA in its commitment to widespread dissemination of information concerning its activities.

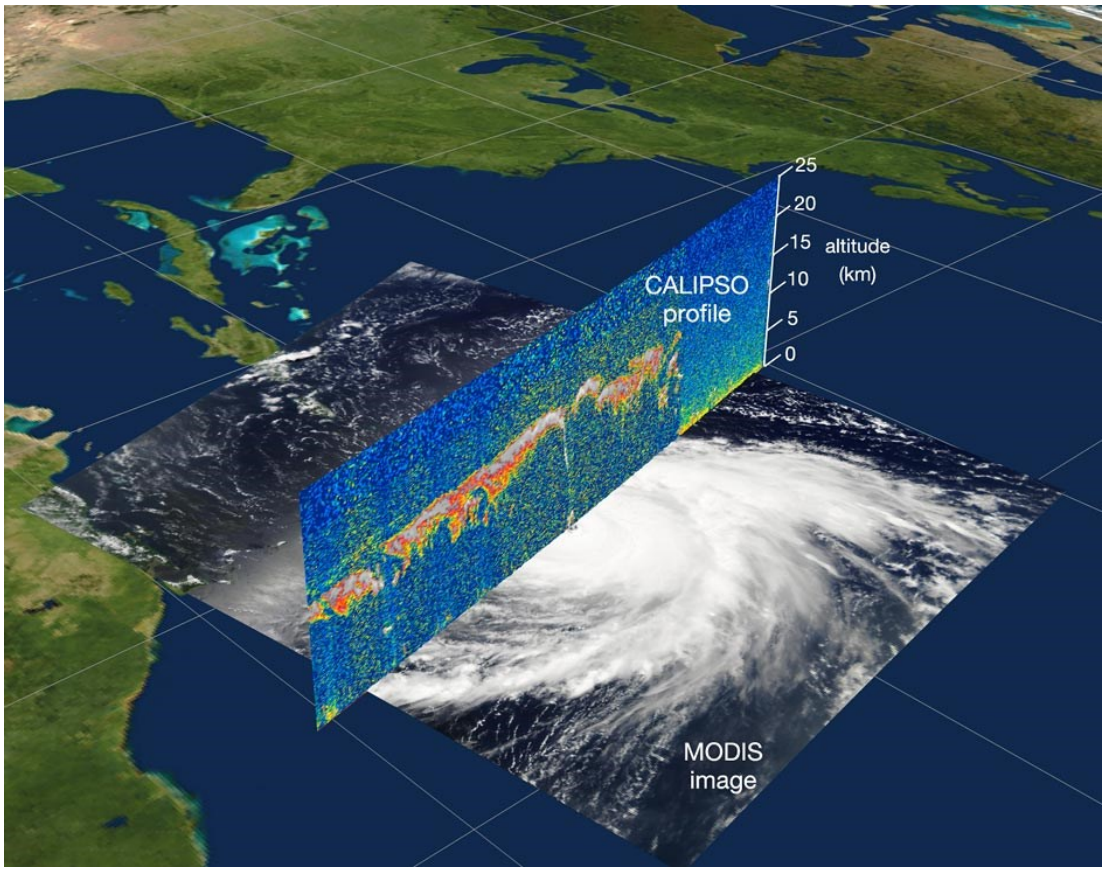

Figure 1 3D Slice of the atmosphere captured by CALIPSO

3.2. CALIPSO Satellite

The Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite is a vital mission of the NASA Langley Atmospheric Science Data Center. CALIPSO was launched in 2006 to analyze the regulating effects of clouds and aerosols on Earth's weather, climate, and air quality. CALIPSO employs active Light Detection and Ranging (LIDAR) instrument to analyze a three-dimensional slice of the atmosphere as it passes overhead (illustrated in Figure 1).[16]

CALIPSO orbits with four other satellites known as the Afternoon or "A-Train" constellation. These satellites fly in close formation and cross the equator at ever-changing longitudes at the same time each day, about 1:30 pm local time. The satellites have complementary instrumentation and their close separation, measured in tens of seconds, allows examination of the same cloud areas at nearly the same moment.[17]

Data products for CALIPSO are readily available through CALIPSO's Data Availability Site. The availability of this data enables NASA to leverage machine learning algorithms for scientific discovery. For example, in 2019, the OCIO Data Science Team (DST) collaborated with Dr. Shan Zeng of the Langley Science Directorate to apply a deep learning CNN model to CALIPSO LIDAR. Together, the team demonstrated the capacity to use CALIPSO to classify aerosols at varying altitudes.[18]

In a parallel research effort, scientists discovered that the CALIPSO data, initially designed for atmospheric study, contained backscatter that had penetrated the ocean an additional 20 meters below the surface. This data revealed a global phenomenon called the Diel Vertical Migration (DVM), where small sea creatures ascend from the ocean depths every night and consume phytoplankton near the surface before returning to the depths before sunrise. This daily event is recognized as the largest animal migration on our planet.[19]

3.3. Coral Reef Use Case

While remote sensing data from space has been robustly consolidated and made available, data specific to fields of terrestrial observations are not as consistent or readily available. One such field is the observation of coral reefs. Around the world, official surveyors, scuba divers, and boaters worldwide spot and record their observations of coral reefs and relay these observations to survey networks. Examples of survey networks include Australia's Eye on the Reef program[20], Khaled bin Sultan Living Oceans Foundation[21], and the Southeast Florida Coral Reef Initiative[22]. These networks report their results to global networks such as ReefBase and the Coral Restoration Database.

3.4. Exploratory Jam Session

On July 24, 2020, the OCIO Data Science Team (DST) hosted a Jam Session, one of its custom hackathons, to assess the suitability of the above-named databases for machine learning studies. During a typical Jam Session, the DST invites data scientists and analysts from across the NASA community to address a specific problem using advanced technologies and libraries. At this Jam Session, the team focused on analyzing coral reef databases using advanced computing resources provided by the NASA Marshall and Agency Computing Services (MACS) Google Cloud Platform (GCP) environment. A key question was whether or not the open data provided on coral reefs was sufficient to compare against other measurements for machine learning. Participants mined and summarized results from the following datasets, while performing knowledge discovery and identifying dataset inconsistencies. Participants attempted to fuse the following datasets during this exercise:

- ReefBase: A Global Information System for Coral Reefs - http://www.reefbase.org/main.aspx
- Harvard WorldFish Dataverse - https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/KUVQKY/PAMLRZ
- NASA GISS Global and Zonal Temperatures - https://data.giss.nasa.gov/gistemp/
- Australian Institute of Marine Science Coral Index - https://apps.aims.gov.au/metadata/view/7c6101f9-50a6-46fb-afe9-c16bd09334d0

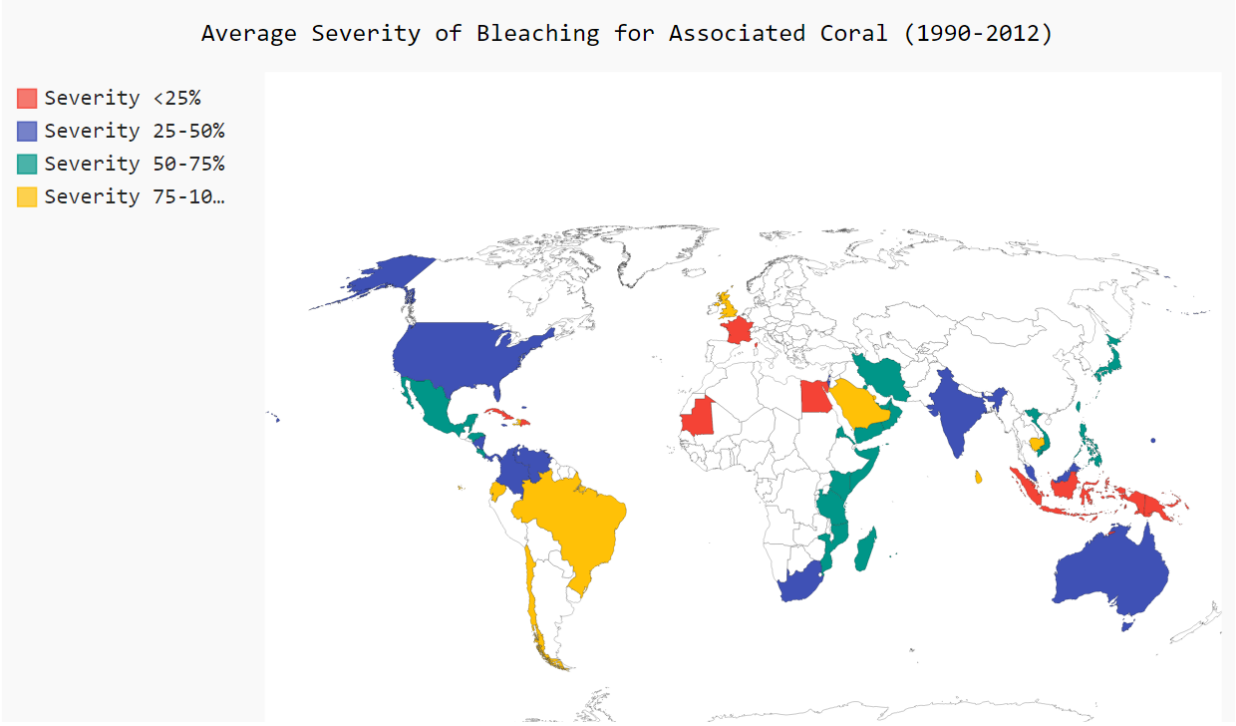

Figure 2 World Map of Coral Sightings color-coded by Bleaching Severity

Participants aligned and checked for inconsistencies across these datasets, producing results such as those shown in Figure 2 and Figure 3.

The July 24 Jam Session began with a virtual guest visit and introduction to the subject matter by Coral Vita, the world's first commercial land-based coral farming company for reef restoration. Coral Vita is an Earthshot Prize-winning for-profit organization focused on growing diverse and resilient corals on land up to 50x faster and planting them in degraded underwater reefs. Their reef experts provided a primer on coral reef sustainability and resilience at the Jam Session to help data scientists understand the domain prior to analysis. This Jam Session was submitted to NASA LaRC Key Activities index (KEY0066690). In addition, the Jupyter notebooks from this Jam Session were uploaded to NASA's internal Enterprise GitHub archive at developer.nasa.gov.

The session revealed significant problems with any attempts to gather reliable information from a single coral reef database. These datasets were produced mainly through human observation over widely varying time scales. For example, coral bleaching, a phenomenon caused by the breakdown of the symbiosis between corals and their symbiotic microalgae, is often identified and reported by scuba divers that observe a loss of pigments and symbionts. Many other reports come from one-time observations during a scuba expedition or deliberate observation mission. During the Jam Session, participants noted that data indicating that a patch of coral was bleaching at a specific time in one location would often be marked as normal or of low-severity by another database at the same corresponding time. In addition, each database tracks both temporal and spatial scales differently. This makes it difficult for any one of these data sources to serve as ground truth for any form of machine learning-based inference.

Two key insights were derived from this Jam Session. To provide a high-fidelity, real-time picture of the world's coral reefs, (1) more effort needs to be put into the accurate, real-time collection of

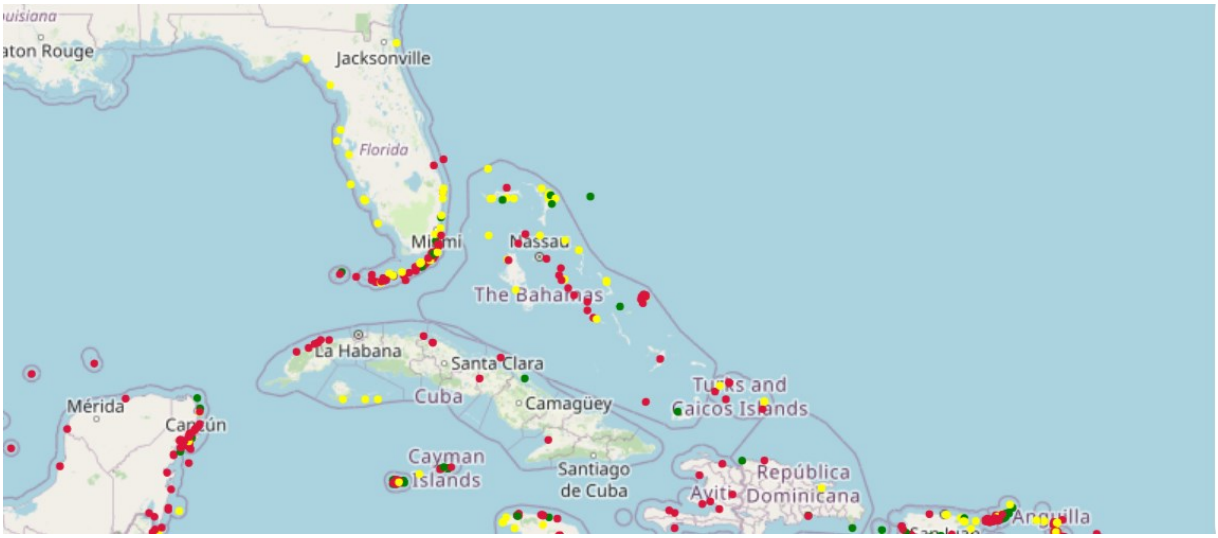

Figure 3 Map of Coral Reef Sites color-coded by level of protection: Protected (Green), Tourist (Yellow), Unprotected (Red)

global coral reef data and (2) existing datasets need to be fused and validated for accuracy through logical reasoning and subject-matter expertise.

These insights were important for organizing the GOTECH project and defining its research path.

4. Project Organization

4.1. Project Considerations

The OCIO Data Science Team (DST) organized the GOTECH project with the following considerations and constraints in mind:

- Minimal funding requirements, if any
- No hardware or software dependencies on NASA
- No security risks or exposures
- No intellectual property issues
- Student teams obliged to perform and freely share their work
- No problematic employer/employee relationships
- Total freedom of action for NASA to use the student work
- A stimulating data science assignment for students that would reinforce their positive views of NASA

4.2. NIA Collaboration Agreement

To manage within the above constraints, the duties of the multiple parties were made explicit in a Collaborative Agreement (CA) executed by the National Institute of Aerospace (NIA). The CA required some consideration, and a small amount ($2.5K) was appropriated to facilitate a possible post-project visit by the student teams to the Langley Research Center. The agreement allowed NASA leads to effectively provide mentoring capability without being involved in the creation of work and data products, which would require significant NASA approvals for use and publication.

The following were the key terms of the CA agreement:

- NASA will not directly transmit any code or data to the students. (Students were given links to NASA data that has already been published)
- No sensitive or classified information is used as part of this research.
- As only data in the public domain will be used, Georgia Tech and its students are not required to sign any Non-Disclosure Agreements (NDA's).
- Student team members need to be US Citizens or US permanent residents.
- NASA provides no NASA-owned software for the project.
- NASA provides no computers or other hardware; Georgia Tech provides all compute resources.
- At project completion, all student code, data, and reports are to be published to a public-facing website that Georgia Tech maintains.
- The school is responsible for obtaining Intellectual Property releases from the students to be able to publish the results open-source without restriction.

These terms allowed the students to successfully complete the goals of the effort without significant constraints.

The NIA CA contains the following description of the project scope and expected outcomes:

> The research will include but not be limited to the below:
>
> 1. Analyze NASA satellite LIDAR imagery and other published government data.
> 2. Detect and monitor the health of the world's Coral Reefs.
> 3. Correlate results with published data sources.
> 4. Establish a baseline for future coral reef monitoring.
>
> The principal purpose of this research is to develop a set of data fusion and machine learning tools for monitoring coral reef health and to increase understanding of the current coral reef health as a baseline for future coral reef monitoring. The project is expected to contribute to human awareness of climate change and its impact on earth biodiversity. It is important to share the results of the project as open source so they may be used immediately by the scientific community.
>
> The expected outcomes from the student team are a mid-term presentation and a final report summarizing findings. Any new code or data resulting from this research will be publicly released by Georgia Tech through publication in an open-source public facing website.

4.3. Project Objectives

The overall objective of the GOTECH project is to fuse open data sources regarding coral reefs to serve as ground truth for a machine learning model that predicts properties of the vitality of coral reefs, based solely on CALIPSO satellite imagery at the reef's location. Students worked with NASA and Coral Vita to address challenges in the following four Technical Areas:

- **Technical Area 1:** Cross-Validate the Open-Source Reef Databases
  - *Questions:* What data can be pulled from these systems to serve as ground truth from?
  - *Challenge:* For this technical area, students should identify at least four coral databases to combine into a single dataset. Students will provide scripts to download and synthesize data from each source into the single dataset. Below is an initial listing of sites. Students are welcome to identify additional databases to serve as ground truth.
    - ReefBase
    - Coral Restoration Database
    - NOAA's CoRIS
    - Giovanni
    - NOAA CoastWatch

- *Deliverable:* Students will provide one common data source with combined data from each of the original sources for training and scripts that would allow downloading and fusion.

- **Technical Area 2:** Time-Align and Geo-Align with Corresponding CALIPSO Data
  - *Questions:* What CALIPSO data do we need for training? What CALIPSO data do we need for inference?
  - *Challenge:* Develop API clients for the CALIPSO database to acquire necessary data for neural network training and inference.
  - *Deliverables:* Students will provide scripts for downloading the necessary data for training and inference, as well as all the specific training/inference data used in the report.

- **Technical Area 3:** Correlate Imagery Backscatter with Coral Vitality
  - *Question:* What neural network architectures are appropriate for this kind of inference? Can we infer growth and decay of coral?
  - *Challenge:* Choose two neural network models for this challenge. Implement methods for neural network training using the combine dataset. Use neural network inference (based on CALIPSO imagery) to demonstrate that we can predict existence of coral in an observed region. Use inference to demonstrate that we can predict properties of vitality in an observed region.
  - *Deliverables:* Students will provide scripts for inference and training, a description of the data that was used for each, a description of how each model was trained, and a saved version of reported models. These descriptions can be provided in the Final Report.

- **Technical Area 4:** Interpret Trends in a Known Global Region based on Backscatter Alone
  - *Question:* Can we infer where growth and decay are happening from CALIPSO backscatter data? How does inference work across two different models?
  - *Challenge:* Design and develop experiments to demonstrate the predictive accuracy of the developed models. Compare and contrast their performance. Characterize how changes to each model architecture impact performance.
  - *Deliverables:* Students will provide scripts for analysis (preferably in Jupyter notebooks), a description of the data that was used for analysis, a description of the response surface for each model, and future recommendations. These descriptions can be provided in the Final Report.

Each student team was required to address the challenges of every technical area.

4.4. Project Logistics and Operation

The GOTECH project launched on August 25, 2021. Students, who were enrolled both in the online and on-campus GT Data Analytics Program, participated in the effort. GOTECH began with a project kickoff presentation. Leads from NASA, Coral Vita, and Georgia Tech described the size and scope of the problem, the technical details students would need to begin research, and the required project deliverables.

Georgia Tech required explicit specification of the minimum data that students would need to complete the project by August 23, 2021 (the beginning of the semester). In addition, for semester requirements, Georgia Tech requested the deliverables and requirement in Figure 4.

On-campus and Online students:

- Midterm Report is due on October 6 at 11 PM PT (October 7 at 2 AM ET).
- Final Report is due on December 1 at 11 PM PT, December 2 @ 2 AM ET.
- Company/organization submits performance evaluation to GT by December 8 at 11 PM PT, December 9 at 2 AM ET.

On-campus: Students present their final results/recommendations to the company/organization by December 8 at 11 PM PT, December 9 at 2 AM ET.

Team formation: This should be finalized within the first two weeks after classes start (by Sept 6).

Figure 4 Georgia Tech Project Requirements for Students

These deliverables requirements were consistent with the requirements established by NASA leads, shown in Figure 5. The overlap between the specifications of Figure 4 and Figure 5 allowed students to submit deliverables that met both requirements without significant extra effort. NASA leads sent the students a very specific set of instructions for starting their research, in the form of a Project Launch Document.

Preferred Method of Communication: The team will use Slack as the preferred method of communication for team collaboration and coordination. Please get in touch with the points of contact for any access issues.

Frequency of Meetings: The team will have weekly video calls to discuss status, findings, and project issues throughout the semester. At the project launch meeting, the team will determine a regular day/time for these meetings.

Software Configuration Management: The Fall 2021 team will use a Git repository to upload all software for the project. The student team will be free to configure software however they choose. However, the data fusion algorithms, machine learning models, and data must be accessible to NASA and Coral Vita at the scheduled deliverable due dates.

Document Management and Software Submission: All student deliverables should be deposited to one repository for NASA and Coral Vita personnel to download. Students will provide a link to a Georgia Tech endpoint for NASA and Coral Vita to download their work. Students should consult with Georgia Tech to gain access to this endpoint.

Fall 2021 Semester Schedule

Due Date	Deliverable
Aug 25, 2021	Attend Practicum Project Launch
Sep 15, 2021, 11:59 PM	Abstract of Semester-Long Approach for the development of data fusion algorithms and machine learning models, as well as selection of data sources
October 13, 2021	First demonstration of data fusion algorithms and combined data source; first implementation of training/inference pipeline
Nov 24, 2021	Final Demonstration of data fusion algorithms, combined data source, and machine learning performance
Nov 29, 2021, 11:59 PM	Final Project Write-Up Due

Figure 5 NASA Project Requirements for Students

5. Key Outcomes

Most of NASA's interaction with the Georgia Tech teams was through weekly status meetings. Every week, an individual student would present a deep-dive research presentation to the group. These meetings served as an opportunity for NASA and Coral Vita team members to provide regular feedback that would benefit both teams. It was also a method for the students to regularly convey critical findings to the group and provide visuals for stakeholders to relay back to subject-matter experts in each group. This section discusses those findings. Detailed results are in Appendices A and B.

5.1. Student Team Technical Approaches

Each student team was tasked at the outset to define a technical approach for addressing the challenges of the four Technical Areas (Section 5.3). The task forced teams to scope their work and focus on a narrow set of technical problems throughout the project. A summary of each team's initial approach is offered here to describe some of their initial assumptions.

Team 1 developed a structured approach to downloading data from each coral reef data source, aligning it to discrete points in time, binning Earth latitude/longitude points (into polygons), and labeling areas based on the existence of coral and bleaching. The team analyzed data from four different sources:

Dataset	**Utility in Study**
CALIPSO	Spatial/Temporal LIDAR observations used for training data and sole dataset for inference
NASA Earth Observations Chlorophyll Concentration Map	8-day aggregations of chlorophyll concentration in the ocean, with a granularity of ~.69 square miles.
Allen Coral Atlas	Coral/Algae recorded polygons aggregated to centroids by the student team. Potentially coincides with ~2.4 million CALIPSO records
Australian Institute of Marine Science	High-accuracy coral observations concentrated on the Great Barrier Reef using Manta Tow observations

Table 1 GOTECH Team 1 Data Sources

Coral Vita and the students identified a high correlation between chlorophyll concentration and coral/algae sightings. NASA Earth Observations Chlorophyll Concentration maps were identified as a critical data source. Students wrote a bot to scrape this data and use it to confirm, with higher confidence, coral sightings in the Allen Coral Atlas. In addition, Team 1 vetted the UNEP Global Distribution of Coral Reefs, NOAA Coral Reef Watch, NOAA Ocean Acidification datasets, and ReefBase These were removed from the team's final results, not because they were not useful to the concept of operations, but because they did not fit the timeline and scope of the Team's approach.

Team 1 later considered using spatial databases for storing data, based on recommendations from the NASA mentors. Spatial databases are ideal for data fusion with this kind of data. As the team quickly saw, making sense of these data and fusing them required an initial filter and references to

location. Spatial queries can simplify some of the efforts in building these filters and relations [23][24][25].

For machine learning, Team 1 proposed using K-Means clustering to manually investigate geo-aligned features across multiple datasets to determine response variables for given geographical areas. Then, they used the resulting clustering to select specific features for machine learning. These responses helped define the training data labels for a feed-forward artificial neural network (ANN) and convolutional neural network (CNN) that make predictions based on CALIPSO backscatter data. Finally, Team 1 researched methods for verifying their results throughout the semester and thought about ethical risks such as incidentally pinpointing objects outside of the realm of marine biology and ways of conveying confidence in the prediction.

Team 2's approach gave significant consideration to the temporal aspects of this problem. They viewed the data fusion problem inherent to Technical Areas 1 and 2 as putting together a timeline. In order to create classifiers that could infer the existence of coral, Team 2 proposed the fusion of the following data sources:

Dataset	**Utility in Study**
CALIPSO	Spatial/Temporal LIDAR observations used for training data and sole dataset for inference
UNEP World Conservation Monitoring Center (WCMC)	Worldwide coral observations specified using geojson polygons
Allen Coral Atlas	Coral/Algae recorded polygons aggregated to centroids by the student team. Potentially coincides with ~2.4 million CALIPSO records
ReefBase	Coral location and bleaching data from over 120 countries and territories

Table 2 GOTECH Team 2 Data Sources for Coral Observations

In addition, Team 2 vetted a separate set of data sources to understand issues of coral vitality:

Dataset	**Utility in Study**
CALIPSO	Spatial/Temporal LIDAR observations used for training data and sole dataset for inference
Florida Bleach Watch Report	Comprehensive surveys on coral health provided by volunteer field observations
NOAA Coastwatch Coral Reef Watch Bleaching Report	Satellite coral bleaching heat stress monitoring with 5 km coverage areas
NASA Giovanni	Contains properties for inference such as chlorophyll, organic/inorganic particulates, and PAR over 5 km coverage areas

Table 3 GOTECH Team 2 Data Sources for Coral Vitality

Team 2 scoped the area of experimentation for their study to the Bahamas and Florida. The students proposed two models for correlating data: UNETs [26] and Deep Neural Networks using

linear SVMs[27]. They proposed to evaluate the ability of both models to make predictions about the growth or decay of coral in CALIPSO polygons over time.

5.2. Key Findings and Research Issues

As students met weekly with the program leads, they reported the issues they encountered and how they addressed each technical challenge. Mitigation of these issues came by having subject-matter experts in marine biology, satellite data processing, and data science answer their questions and give tutorials as needed.

An early task for the students was to understand enough about the marine biology domain to gather requirements for establishing a coral health dataset. To help students move up the learning curve, Dr. Katey Lesneski of Coral Vita began the semester by giving students a full hour-long tutorial on coral reefs and vitality issues. This set the stage for students to identify critical parts of the data significant to experimentation.

A common problem with understanding these requirements was the unstandardized nature of the datasets themselves. The table below describes the key issues that were ascribed to each dataset. More details regarding these issues can be found in the final reports given in Appendix A and Appendix B.

Dataset	**Issues that Compromise Study**
Allen Coral Atlas	• The presence of coral and the presence of algae are highly correlated; this made it challenging to separate observations
Australian Institute of Marine Science	• Very small number of highly accurate observations, due to the data collection method and size/shape of the Great Barrier reef
CALIPSO	• Sometimes challenging to query specific areas; requires large data pulls • Temporal gaps due to flyovers from the satellite make it challenging to align with other datasets • Spatial gaps due to satellite trajectory miss a significant number of potential data alignments
Florida Bleach Watch	• Smaller number of observations based on data collection methods • Sometimes unreliable observations based on data collection methods
NASA Earth Observations Chlorophyll Concentration Map	• Significant temporal and spatial misalignment with CALIPSO
NASA Giovanni	• 4km resolution was insufficient for model prediction confidence • Data acquisition scales exponentially

NOAA CoastWatch (DHW)	• Bleaching reports are from modeling, not real-world observations • Reported bleaching is based on more extensive environmental conditions and not necessarily those that impact coral
ReefBase	• Significant gaps in temporal, spatial, and characteristic data features
UNEP World Conservation Monitoring Centre (WCMC)	• 85% of database observations are older (1999-2002) • ~43% of the database is unvalidated

Ingesting and fusing each data source proved challenging for each student team. In addition, storage and computation of their data also provided challenges. The nature of the data required the capture, storage, and joining of millions of data points at a time. The teams used Georgia Tech's PACE-ICE environment for their development work, with a storage quota of 500GB. The use of an HPC platform was critical to each team's successful exploration of fusion and modeling techniques.

Team 1 leveraged a PostGIS database to conserve disk space and optimize query time. Of the 582 CALIPSO backscatter signal features, Team 1 included bins 382-582 for data fusion and model prediction. Team 2 used PostgreSQL and included features 282-582. The teams suggested improving model accuracy in the future by using a more robust feature selection. This can be done by removing features with negative importance and algorithmic transformations such as PCA.

In the end, both teams' model prediction results were compelling. Team 1 made predictions for Florida's coral reefs with 80.1% accuracy and the Australian Great Barrier Reef with 74% accuracy. Team 2 made predictions for Florida's coral reefs with an accuracy of 77.67%. The lack of separability between coral and algae and temporal and spatial sparsity among the datasets were key barriers that might be overcome by introducing additional datasets. The students' models were foundational yet elementary, suggesting that further work in feature selection, tuning, and fusion of additional data sources (including additional satellites) will be necessary to provide actionable results for coral reef restoration groups.

6. Conclusions

The Geophysical Observations Toolkit for Evaluating Coral Health (GOTECH) project is one of many case studies demonstrating the value of combining NASA satellite instruments with data fusion and machine learning. In this case study, two teams of students from the Georgia Institute of Technology (Georgia Tech) demonstrated the ability to fuse disparate coral reef datasets into confident training data for machine learning models. They then aligned this fused dataset with observations from the NASA CALIPSO satellite to train machine learning models that predict the existence and vitality of coral reefs on satellite pass. Their results were captured in the Final Reports, included as Appendices A and B.

Experimentation showed that the models could predict coral and vitality with an academically reasonable confidence level. Should more advanced machine learning models and data fusion techniques be used, GOTECH would allow coral reef restorers to save a massive amount of time and resources that are currently spent on identifying reefs that need to be restored. By continuously orbiting Earth, NASA assets have the unique capability of answering questions beyond the scope of their original mission. As new climate-related issues arise on Earth, machine learning and data fusion permit NASA space assets to have an indefinite use case.

The two student teams proved that data fusion across these kinds of data sets was possible. They developed novel approaches for scraping and geo-alignment of coral reef data, and explored new analysis and machine learning techniques. With many possible algorithms to be applied, expanding this project to be executed among different schools would allow for a thorough exploration of tools that could be used for this mission. Expanding the project would also reveal shortcomings and incorrect assumptions that are currently made about bleaching predictions. Today's university student teams are highly motivated due to the nature of the project. As catastrophic events due to climate change become more prevalent, young students have become extremely passionate about using their skills to address climate-related issues. For that reason, GOTECH and similar projects will have no shortage of students that would select such projects if made available to them through Practicums similar to Georgia Tech's.

Within NASA, GOTECH demonstrates the need for data scientists to have a more significant footprint across the Administration. NASA space assets have use cases well beyond their initial intended missions. Experienced data scientists saw the potential for fusing coral reef data sources to achieve the GOTECH concept, based on their understanding of open-source data and modern machine learning algorithms. All around NASA, projects that yield global continuous monitoring data have potential use cases that data scientists with experience can see. NASA would benefit from an enterprise-level entity specializing in providing data science expertise to identify, propose, implement, and lead projects based on these use cases. As NASA proceeds further into the 21st century, such an entity is required to achieve another facet of its statutory responsibility[1].

The GOTECH project required establishing a format and instruction sets for graduate students to follow, such that they would produce high-quality research. NASA personnel have documented this format and preserved the instruction sets such that similar projects can easily be hosted at other schools. At the time of this writing, the leads from the NASA LaRC OCIO Data Science Team intend to leverage the format and instruction sets to repeat the exercise with two University programs. The instructions will continue to be improved and versioned to adapt to various

[1] To "seek and encourage, to the maximum extent possible, the fullest commercial use of space" (51 U.S.C. § 20112)

University programs. The key benefit is that this structure permits crowd-sourcing of ideas for fusing and leveraging NASA assets for Earth missions by deeply passionatestudents that have a cutting-edge understanding of the latest data science techniques.

Finally, during the Fall 2021 effort, student teams were provided with dedicated experts in data science and coral reefs who discussed project status and results with them every week. The students could have benefitted from a dedicated expert from the CALIPSO satellite team. While the CALIPSO data is well-structured, having experts for the satellite instrument would allow students to pose direct questions that could have benefitted their research easily. In addition, CALIPSO experts have enough familiarity with instrument data to identify insights that the teams may have missed. While the GOTECH leads consulted with some of CALIPSO's researchers to address technical issues, future efforts will ensure that a satellite instrument expert is part of the routine project activities.

7. References

[1] NASA. (2022, March 7). *The causes of climate change*. NASA. Retrieved May 13, 2022, from https://climate.nasa.gov/causes/

[2] *NASA mission a game changer for Coral Reef Science*. CSIRO. (2020, March 11). Retrieved May 13, 2022, from https://bit.ly/3xWRTbB

[3] NOAA Coral Reef Watch. (2021). *Coral Reef Watch Home*. NOAA Coral Reef Watch Data Resources. Retrieved May 13, 2022, from https://coralreefwatch.noaa.gov/crtr/data_resources.php

[4] *A global information system for coral reefs*. ReefBase. (2021). Retrieved May 13, 2022, from http://www.reefbase.org/main.aspx

[5] *51 U.S.C. 20112 - Functions of the Administration*. Govinfo. (2021). Retrieved May 13, 2022, from https://www.govinfo.gov/app/details/USCODE-2012-title51/USCODE-2012-title51-subtitleII-chap201-subchapII-sec20112/summary

[6] Greicius, T. (2016, June 2). *Five things about coral and Coral*. NASA. Retrieved May 13, 2022, from https://www.nasa.gov/feature/jpl/five-things-about-coral-and-coral

[7] Perez, M. (2016, March 21). *A reef scientist talks about NASA's Coral Campaign*. NASA. Retrieved May 13, 2022, from https://www.nasa.gov/feature/jpl/a-reef-scientist-talks-about-nasas-coral-campaign

[8] Burke, L., D. Bryant, J. McManus, and M. Spalding. (2008). *Reefs at Risk*. World Resources Institute (WRI): 56 p.

[9] Costanza, R. and C. Folke. (1997). *Valuing ecosystem services with efficiency, fairness and sustainability as goals*. In: Daily, G. (Ed.), Nature's Services: Societal Dependence on Natural Ecosystems. Island Press, Washington, DC, pp. 49-70.

[10] Eddy, T. (2021, September 17). *Global decline in capacity of coral reefs to provide ecosystem services*. One Earth. Retrieved May 13, 2022, from https://www.cell.com/one-earth/fulltext/S2590-3322(21)00474-7

[11] *Why reefs?* Why Reefs? (2021). Retrieved May 13, 2022, from https://www.coralvita.co/why-reefs

[12] *Georgia Institute of Technology Master of Science in Analytics: Georgia Institute of Technology: Atlanta, GA*. Applied Analytics Practicum Sponsorship – Call for Proposals | Master of Science in Analytics | Georgia Institute of Technology | Atlanta, GA. (2021). Retrieved May 13, 2022, from https://www.analytics.gatech.edu/career-services/practicum-call-proposals

[13] NASA. (2021). *Landsat science*. NASA. Retrieved May 13, 2022, from https://landsat.gsfc.nasa.gov/

[14] NASA. (2021). *Earthdata - Open Data for Open Science*. NASA. Retrieved May 13, 2022, from https://earthdata.nasa.gov/

[15] *National Snow and Ice Data Center*. National Snow and Ice Data Center |. (2022, March 22). Retrieved May 13, 2022, from https://nsidc.org/

[16] Winker, D. (2021). *Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations*. NASA. Retrieved May 13, 2022, from https://www-calipso.larc.nasa.gov/

[17] Winker, D. (2021). *Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations*. NASA. Retrieved May 13, 2022, from https://www-calipso.larc.nasa.gov/about/atrain.php

[18] Zeng, S., et al. (2020). *Identifying Aerosol Subtypes from CALIPSO Lidar Profiles Using Deep Machine Learning. Atmosphere*, 12(1), 10. https://doi.org/10.3390/atmos12010010

[19] Northon, K. (2019, November 27). *NASA, French space laser measures massive migration of Ocean Animals*. NASA. Retrieved May 13, 2022, from https://www.nasa.gov/press-release/nasa-french-space-laser-measures-massive-migration-of-ocean-animals

[20] *Eye on the reef*. GBRMPA. (2021). Retrieved May 13, 2022, from https://www.gbrmpa.gov.au/our-work/eye-on-the-reef

[21] *Scientific surveys on coral reefs by the Living Oceans Foundation*. Living Oceans Foundation. (2019, September 12). Retrieved May 13, 2022, from https://www.livingoceansfoundation.org/science/scientific-surveys/

[22] *SEFCRI project reports and products*. Florida Department of Environmental Protection. (2021). Retrieved May 13, 2022, from https://floridadep.gov/rcp/coral/content/sefcri-project-reports-and-products

[23] *Spatial databases - build your spatial data empire*. GIS Geography. (2022, January 27). Retrieved May 13, 2022, from https://gisgeography.com/spatial-databases/

[24] Mitchell, T. (2021, April 23). *Geospatial Basics, Spatial Databases & nosql examples*. The Couchbase Blog. Retrieved May 13, 2022, from https://blog.couchbase.com/geospatial-basics-spatial-databases-and-nosql-examples/

[25] Sami, R. (2019, September 25). *How to work with Big Geospatial Data?* Medium. Retrieved May 13, 2022, from https://towardsdatascience.com/how-to-work-with-big-geospatial-data-4ba919a8ffc2

[26] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). *U-net: Convolutional networks for biomedical image segmentation*. In International Conference on Medical image

computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
[27] Tang, Y. (2013). *Deep learning using linear support vector machines*. arXiv preprint arXiv:1306.0239.

Appendix A – Team 1 Final Report and Presentation

Project GOTECH:
Predicting Coral Presence Using Satellite-Based LiDAR

Josh Mattingly, Tina Guo, Dan Schauder
jmattingly31@gatech.edu, yguo96@gatech.edu, dschauder3@gatech.edu

***Abstract*—** While approximately 30% of the world's marine species depend on coral reefs, "all the coral reefs in the world could be gone by 2070 if global heating continues on its current path" (Morrison et al, 2019). Despite the urgent and dramatic threat to our planet's coral ecosystems, existing monitoring methods for evaluating coral presence are fragmented and limited in scope (Foo & Asner, 2019). This paper describes the Geophysical Observations Toolkit for Evaluating Coral Health (GOTECH) project, in which graduate students at the Georgia Institute of Technology combined open-source coral data with satellite-borne LiDAR (Light Detection and Ranging) data to build machine learning models capable of detecting coral presence, offering a novel avenue for scalable and sustainable coral tracking.

1 INTRODUCTION

The GOTECH project employed a nearest-neighbor algorithm to merge disparate open-source coral datasets and establish a set of canonical binary labels denoting whether coral is present in a given coastal region. This baseline dataset was then merged with backscatter data from a satellite-borne LiDAR sensor aboard NASA's Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite. Lastly, a set of machine learning classification models were trained to predict coral presence in two regions where coral populations have been studied closely (Florida and the Great Barrier Reef).

2 BACKGROUND

Foo & Asner asserted in 2019 that remote sensing technology such as LiDAR offers a promising complementary avenue for monitoring coral populations on a global scale. Bathymetric LiDAR has been successfully used to map shallow warm water coral, utilizing camera-equipped drones to establish ground truth (Collin et al, 2018). The findings of Collin et al, though, are limited by their access to an airplane equipped with the proper sensors which can scan any given area. Considering the urgency of action related to coral bleaching, the ability to regularly monitor areas with an orbital satellite could provide an opportunity to deliver results more quickly, reliably, and consistently as compared to sea or airbased solutions.

Satellite-based LiDAR methodologies have been explored (Parrish et al, 2019) using the ATLAS ICESat-2 sensor. While designed specifically for monitoring ice, as its name implies, its bathymetric uses proved promising. Because their model, as with Collin et al, requires drone-based ground truth, Parrish et al state further research should be performed on sites outside of their study. Leveraging the Allen Coral Atlas (see Section 3.2) as our ground truth, we were able to train, test and validate against a variety of coral regions.

Readings from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) sensor aboard CALIPSO have been used specifically to monitor trends in marine species. Behrenfeld et al described a method for inferring surface phytoplankton concentrations from CALIOP backscatter data (Behrenfeld et al, 2016) and later expanded upon this method to track the biomass of marine life engaging in a migratory cycle known as DVM (Diel Vertical Migration) in which animals ascend to the ocean surface to feed on plankton at night and recede to the ocean depths during the day to avoid predators (Behrenfeld et al, 2019). Our project further extended Behrenfeld et al's methods to explore the hypothesis that CALIOP's remote detections of particulate backscatter density at the ocean surface are correlated with coastal coral presence.

Coral reef presence and health can be impacted by a variety of anthropogenic factors and natural disturbances. Contaminants, such as chlorophyll, cause declines in hard corals and phototrophic octocorals. In a study done to measure the relationship between four biotic groups and water quality, -45% variation between each group was identified with 18 – 46% of the variation due to water quality effects (De'ath & Fabricius, 2010). Thus, chlorophyll metrics were used to help better determine the presence of coral in various geographic locations and assess the effectiveness of LiDAR-based methods.

3 DATA SOURCES

3.1 CALIPSO LiDAR Backscatter

The primary data source for this project, the CALIPSO satellite, orbits the Earth in a sun-synchronous orbit, with the track repeating every 16 days (see Figure 1). Given a specified latitude-longitude boundary, LIDAR readings along the orbit path were downloaded, processed, and stored in a POSTGIS database. Along with the backscatter values, the date and time of the reading, position of the satellite, and the land/water mask categorization were also loaded into the database.

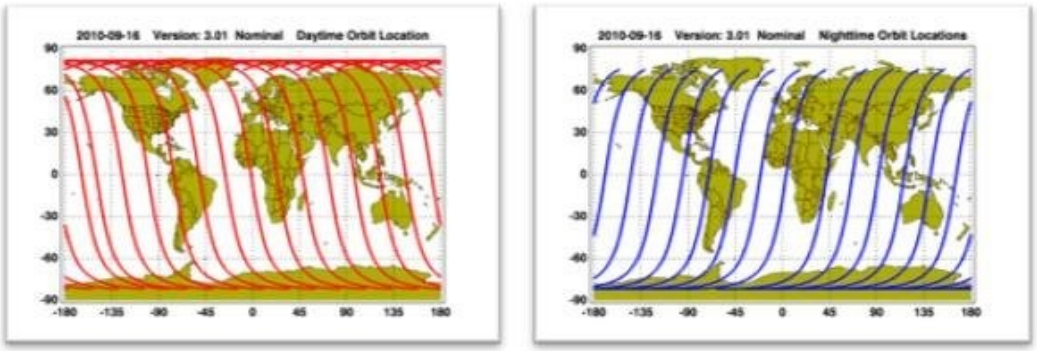

Figure 1 – The 16-day orbital pattern of the CALIPSO satellite.

3.2 Allen Coral Atlas

Benthic classifications were taken from the Allen Coral Atlas website for the Florida region. The benthic data was then aggregated to a Boolean response variable, with 1 for Coral/Algae and 0 for anything else.

While prediction accuracy proved promising, the inability to segregate coral from algae and the lack of a time-series component challenges the viability of the Allen Coral Atlas dataset as a short-

window monitoring solution for marine biologists to quickly identify and respond to changes in coral reef health.

Figure 2 **–** Research locations (Left to Right): Grand Bahama, Florida, and the Great Barrier Reef as taken from the Allen Coral Atlas. Areas classified as "Coral/Algae" are marked in red.

To conserve storage space and processing time, the latitude and longitude bounding boxes established via the Allen Coral Atlas were used to limit CALIPSO queries to only regions selected for further research.

3.3 NASA Earth Observatory Chlorophyll

Chlorophyll data from NASA Earth Observations (NEO) is provided to 0.01 degrees of longitude and latitude and can be accessed at either monthly or eightday aggregates. A bot was created to automatically download the eight-day files for years 2016 through 2020.

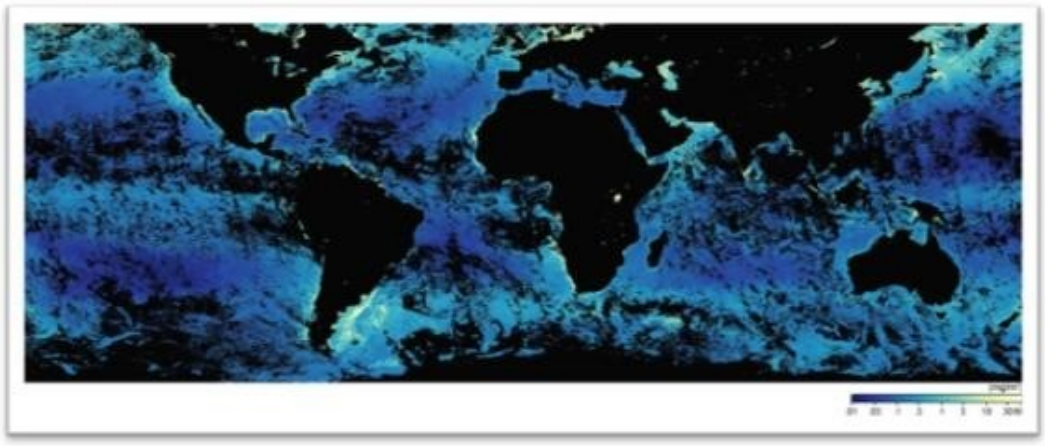

Figure 3 **–** Eight-day aggregate of chlorophyll data.

3.4 Australian Institute of Marine Science (AIM) Survey

The Australian Institute of Marine Science provides access to raw Manta Tow
Survey data conducted at locations across the Great Barrier Reef. For the purposes of our experiment, the data collected, which includes a count of coral in various states of health, as well as the presence of other marine life including coral predators, was truncated to only include the presence of live coral. A threshold was chosen to create a Boolean response variable which would then be used within our predictive framework.

While the AIM survey data solves the stated concerns of the Allen Coral Atlas set (survey data is provided as a time-series and separates out coral from marine life,) sparseness (~300 points of survey data over the course of six years) severely reduces prediction accuracy.

4 METHODS

4.1 Spatial-Only Fusion

Because the Coral/Algae classification taken from the Allen Coral Atlas dataset lacked a time-series component, the response variable was only matched on latitude and longitude using a nearest-neighbors (NN) approach. To account for CALIPSO's orbit path, as well as the surface area covered by each signal response, a backscatter pattern was only considered a match to a coral/algae record if it was within 5-kilometers (~0.5 degrees).

As each NEO chlorophyll dataset contained readings for the entire planet, a straightforward match of the nearest point via Euclidean distance was possible.

4.2 Spatial and Temporal Fusion

In contrast to the Allen Coral Atlas dataset, the AIMS manta tow survey results provided the date surveys were conducted along with the latitude and longitude coordinates.

As well as the nearest-neighbor approach used in the spatial-only fusion process, a CALIPSO reading was only considered a valid match if the satellite reading was recorded within 14 days before or after the tow survey. This timeframe was decided upon after consultation with a subject matter expert at Coral Vita.

4.3 Data Loading Process

To conserve disk space and optimize query time, each CALIPSO and NEO fusion process was performed on a per-file basis, with only those feature records with a corresponding match in the response datasets being loaded into the POSTGIS database.

4.4 Predictive Models

To expedite the testing of fused datasets, a model competition framework was created. Using a fused dataset, pared down to a series of features and a response variable, a model competition object would be created. Once created, a series of classification algorithms (Random Forest, XGBoost Classification, ANN, and SVC) would be trained, and predictions stored based on a test set.

To prevent oversampling of a specific category (coral/algae or no coral/algae), a subsample was created from the initializing dataset. This subset was then used to construct the training and test datasets, with a default 70/30 split.

Establishing the structure of the model inputs, as well as the object-oriented structure of the framework, allowed for additional models to be added as our research progressed, and gave us the ability to quickly expand our project scope to include the AIMS manta tow survey data.

5 RESULTS

5.1 Florida

5.1.1 Baseline (NEO features only)

Our original machine learning experiments, which included spatial and temporal alignment of the five lowest altitude CALIPSO backscatter signals with NEO chlorophyll datasets (with a coral/algae response variable) returned results close to 85%. Upon further investigation, it was discovered that the high correlation between the presence of coral and/or algae and elevated levels of chlorophyll were skewing our prediction results. When taking the feature sets separately, NEO-only accuracy increased to over 88%, while CALIPSO-only accuracy was no better than chance.

For the remainder of our research, we established the NEO-only model as our baseline for comparison.

5.1.2 CALIPSO (200 backscatter features + Land/Water Mask)

In addition to excluding the CALIPSO data from the baseline model, we also expanded the number of backscatter readings to 200, also adding in the Land/Water Mask feature. The original selection of the lowest five features was due to storage and compute limitations at the start of the research process coupled with assumptions on the way the backscatter data was aggregated. While the expanded CALIPSO data was not included in the initial chlorophyll model, the high correlation of the NEO dataset made us comfortable maintaining chlorophyll-only as the overall baseline.

5.1.3 Model Accuracy Scores

Running each fused dataset through the model competition framework showed different model performing better for each fusion method, with the Random Forest classifier returning an accuracy score 16 percentage points higher than the competing models, while the Artificial Neural Network saw the best performance in the competition when using the main research dataset. While XGBoost "won" the Great Barrier Reef competition with 74% accuracy, the combination of low overall accuracy and the sparsity of the AIMS data when compared to the other competitions makes it difficult to determine whether this fusion methodology would be beneficial for coral-only or manta tow survey-based data.

Model	Baseline	Florida Dataset	Great Barrier Reef
Random Forest	**90.7%**	79.2%	73.6%
XGBoost Classifier	74.7%	78.9%	**74.0%**
SVC	71.6%	74.8%	71.7%
ANN	72.3%	**80.1%**	68.9%

Table 1 – Model competition accuracy scores. Bold scores were the strongest predictors for a given dataset.

5.1.3 Random Forest Feature Importance Results

A view of feature importance output from the Random Forest classifier showed the Land/Water Mask feature to be by far the most important predictor of the presence of coral/algae, with the next most important feature (backscatter bin 563) providing a 6% increase in overall model accuracy.

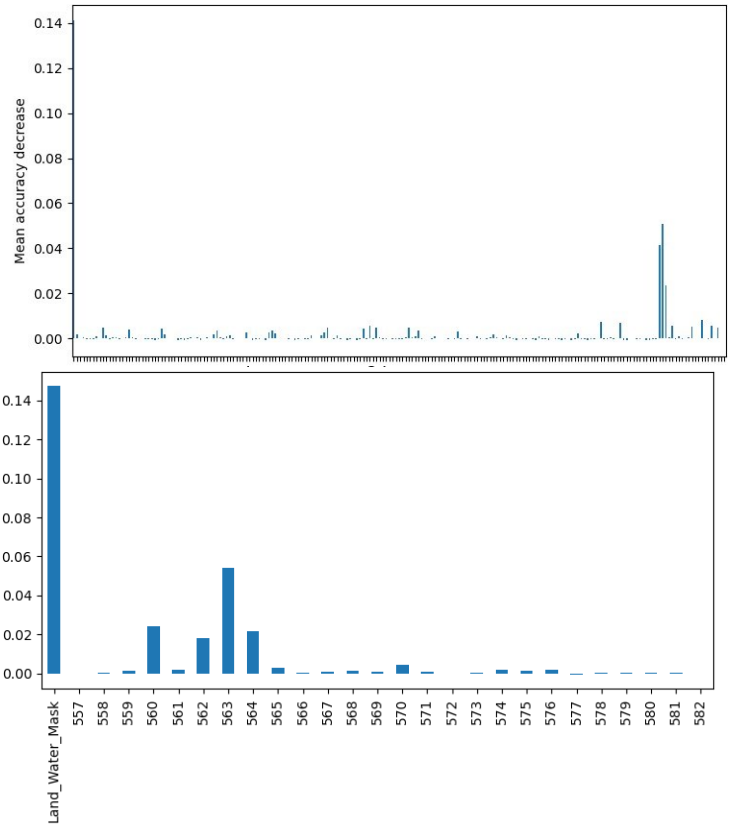

Figure 4 – Top: Feature importance for all 201 CALIPSO features (Land/Water Mask + 200 backscatter features)

Bottom: Close-up of view showing top five features (Land/Water and signals 560, 562, 563 and 564)

6 FUTURE WORK

The machine learning and artificial intelligence models used for this experiment were kept at their default settings. Modifying the competition framework to include model optimization, such as grid search, finding the optimal spatial/temporal distance (outside of 5-kilometers and 14 days,) and a more accurate representation of the 5-kilometer x 500-meter CALIPSO coverage area all hold the potential to increase overall model accuracy. In each case, a balance will need to be struck between overall accuracy and computational resources.

The original challenge set out for the GOTECH teams was to use satellite-based LiDAR, specifically CALIPSO, to make predictions on coral reef health. Further research into additional data sources, both open and proprietary, would be the most reasonable next step to strengthening classification accuracy.

Because it was brought into the project scope later in the process, the chlorophyll (NEO) fusion experiment was not conducted on the AIMS manta tow survey. Further analysis of these datasets, which contain both spatial and time-series components, holds potential for further discovery.

While the AIMS data was a smaller sample size, the data recorded by the survey team includes details on coral health (live, dead, bleached, etc) as well as marine life, including the presence of coral predators. A regression-based testing framework, as opposed to the Boolean approach used for this experiment, could provide a more useful long-term monitoring solution. Also, because reef surveys and CALIPSO flyovers were independent of each other, the odds of matching the two datasets were no better than chance. Coordinating reef surveys with LiDAR scans, satellite-based or otherwise, would allow for the building of a more accurate picture of backscatter in relation to coral reefs.

Finally, conducting the experiment using other forms of LiDAR, such as airplane, ship, or drone-mounted sensors, as well as disaggregated readings, offers the potential for more accurate reef classification.

7 REFERENCES

1. Allen Coral Atlas maps, bathymetry and map statistics are © 2018-2021 Allen Coral Atlas Partnership and Vulcan, Inc. and licensed CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)
2. Australian Institute of Marine Science (AIMS). (2015). AIMS Long-term Monitoring Program: Crown-of-thorns starfish and benthos Manta Tow Data (Great Barrier Reef), https://doi.org/10.25845/5c09b0abf315a, accessed 10Nov-2021.
3. Behrenfeld, M. J., Hu, Y., O'Malley, R. T., Boss, E. S., Hostetler, C. A., Siegel, D. A., Sarmiento, J. L., Schulien, J., Hair, J. W., Lu, X., Rodier, S., & Scarino, A. J. (2016). Annual boom–bust cycles of polar phytoplankton biomass revealed by space-based lidar. In Nature Geoscience, 10(2), 118–122. https://doi.org/10.1038/ngeo2861
4. CALIPSO data was obtained from the NASA Langley Research Center Atmospheric Science Data Center from https://subset.larc.nasa.gov/calipso/.
5. Chlorophyll Concentration: 8-DAY AQUA/MODIS. (2018-2020). Retrieved Oct 25, 2021, from https://neo.gsfc.nasa.gov/view.php?datasetId=MY1DMW_CHLORA
6. Collin, A., Ramambason, C., Pastol, Y., Casella, E., Rovere, A., Thiault, L., Espiau, B., Siu, G., Lerouvreur, F., Nakamura, N., Hench, J., Schmitt, R., Holbrook, S.J., Troyer M., Davies, N. (2018). Very high resolution mapping of coral reef state using airborne bathymetric LiDAR surface-intensity and drone imagery. In International Journal of Remote Sensing, 39:17, 5676-5688, DOI: 10.1080/01431161.2018.1500072
7. De'ath, G., & Fabricius, K. (2010). Water quality as a regional driver of coral biodiversity and macroalgae on the great barrier reef. In Ecological Applications, 20(3), 840–850. https://doi.org/10.1890/08-2023.1

8. Foo, S. A., Asner, G. P. (2019). Scaling Up Coral Reef Restoration Using Remote Sensing Technology. In Frontiers in Marine Science, 6. https://doi.org/10.3389/fmars.2019.00079
9. McLarney, E., Gawdiak, Y., Oza, N., Mattman, C., Garcia, M., Maskey, M., Tashakkor, S., Meza, D., Hestnes, P., Wolfe, P., Illingworth, J., Shyam, V., Rydeen, P., Prokop, L., Powell, L., Brown, T., Miller, W., Little, C. (2021). NASA Framework for the Ethical Use of Artificial Intelligence (AI). https://libguides.umgc.edu/c.php?g=1003870&p=7270670
10. Morrison, T. H., Hughes, T. P., Adger, W. N., Brown, K., Barnett, J., & Lemos, M. C. (2019). Save reefs to rescue all ecosystems. In Nature, 573(7774), 333–336. https://doi.org/10.1038/d41586-019-02737-8
11. Parrish, C.E.; Magruder, L.A.; Neuenschwander, A.L.; Forfinski-Sarkozi, N.; Alonzo, M.; Jasinski, M. Validation of ICESat-2 ATLAS Bathymetry and Analysis of ATLAS's Bathymetric Mapping Performance. In Remote Sens. 2019, 11, 1634. https://doi.org/10.3390/rs11141634

8 APPENDIX: ETHICAL CONSIDERATIONS

Ethical considerations for this project were evaluated according to the guidelines set forth in NASA's Framework for the Ethical Use of Artificial Intelligence (2021). The framework consists of six key principles.

8.1 Fair

> "AI systems must include considerations regarding how to treat people, including refining solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion" (McLarney et al., 2021).

The team actively searched for bias in our data which might negatively influence outcomes or disproportionately impact minority populations. Due to time and resource constraints, the project was limited to certain geographic regions. However, the methodology described in this project may be readily applied to study additional geographic regions and serve as a fruitful avenue for future exploration.

8.2 Explainable and Transparent

> "Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented" (McLarney et al., 2021).

The machine learning techniques explored in this project aim to advance human understanding, and the project does not involve automated decision making. All data sources employed are open-source and publicly available, and the team has formally documented its work in this paper, presentations, and in a publicly accessible code repository.

8.3 Accountable

> "Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight. AI developers should consider potential misuse or misinterpretation of AI-derived results (intentional or otherwise) and take steps to mitigate negative impact" (McLarney et al., 2021).

The development of this project was overseen by Dr. Newton Campbell at NASA and Dr. Katey Lesneski at Coral Vita. In weekly status meetings, the research team presented our progress and findings. To mitigate the potential for misuse or misinterpretation of our results, our findings and code were published on GitHub along with documentation of our processes.

8.4 Secure and Safe

> "AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use" (McLarney et al., 2021).

No personal or sensitive data was gathered or used in the process of this undertaking. All automation was under the direct supervision of human authors, and no automated decision-making was included in this work.

8.5 Human-centric and Societally Beneficial

> "AI systems must obey human legal systems and must provide benefits to society. At the current state of AI, humans must remain in charge, though future advancements may cause reconsideration of this requirement" (McLarney et al., 2021).

The algorithms described in this work due not violate laws, and the primary motivation of this effort is to preserve and protect the environment humans currently depend on for survival.

8.6 Scientifically and Technically Robust

> "AI systems must adhere to the scientific method NASA applies to all problems, be informed by scientific theory and data, robustly tested in implementation, well-documented, and peer reviewed in the scientific community" (McLarney et al., 2021).

The team conducted critical assessments of the data sources used in this work, and aspired to transparency in describing the provenance and quality of the data. Additionally, methodology and results were regularly reviewed by fellow students as well as SMEs at NASA and Coral Vita.

This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Appendix B – Team 2 Final Report and Presentation

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.

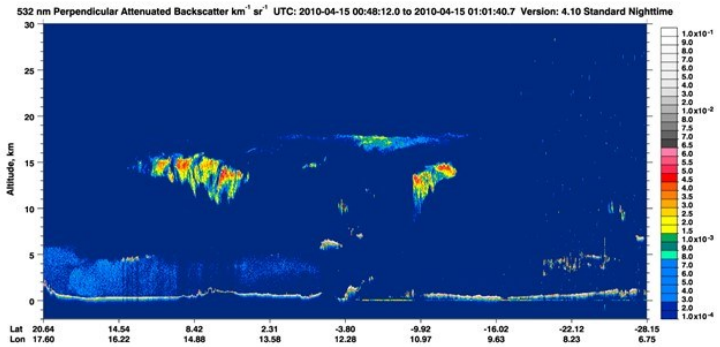

Figure 1: Example of CALIPSO CALIOP Perpendicular Attenuated Backscatter 532 nm Image

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	Weekly
	Degree Heating Weeks	Weekly
	Chlorophyll	Monthly
	PAR	Monthly
	Bathymetry	Single Time Period
GIOVANNI	Infrared radiation	Weekly
	Total suspended matter	Monthly
	PAR	Monthly
	Particulate organic	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.

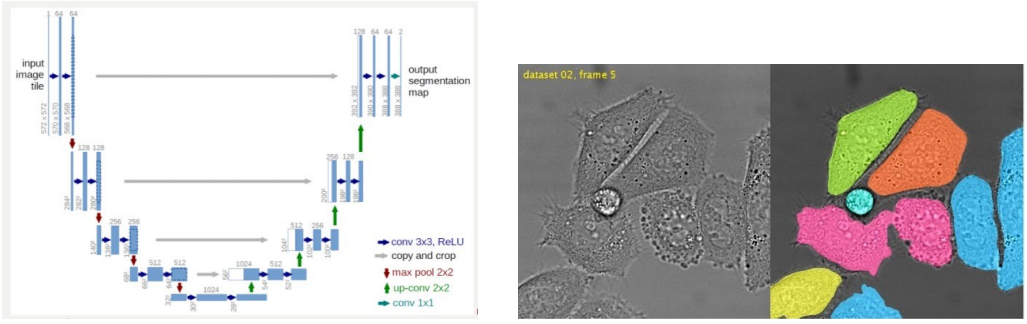

Figure 2: Visual Depiction of the U-net CNN (Ronneberger et al., 2015) on the left. Example Image Segmentation Classification (Ronneberger et al ,2015) on the right

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 220225.

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.

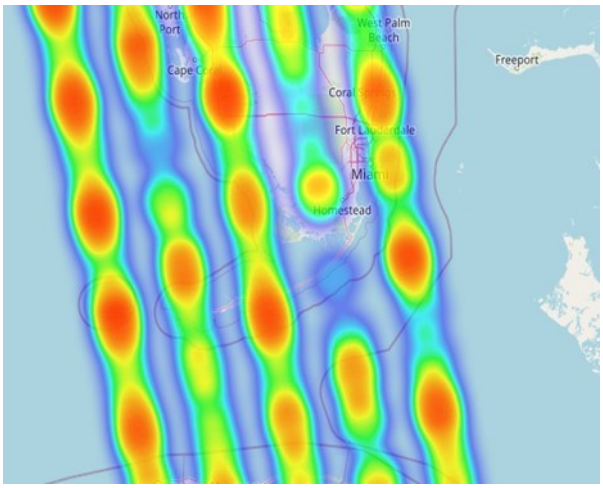

Image 1: Heat Map Depiction of Sparse CALIPSO data

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-coralcover/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://wwwcalipso.larc.nasa.gov/resources/calipso_users_guide/browse/index.php.

Appendix C – Published Project Deliverables

As one of its obligations under the NIA contract, Georgia Tech has posted the student presentations and final reports on an open-source public-facing website.

- https://sites.gatech.edu/gotech/teams/team-1/
- https://sites.gatech.edu/gotech/teams/team-2/

Appendix D – Consumed Public Datasets

The following datasets were used by the student teams in the conduct of their research for the GOTECH project:

- **Allen Coral Atlas maps, bathymetry and map statistics:** © 2018-2021 Allen Coral Atlas Partnership and Vulcan, Inc. and licensed CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

- **CALIPSO data:** Obtained from the NASA Langley Research Center Atmospheric Science Data Center from https://subset.larc.nasa.gov/calipso/.

- **Chlorophyll Concentration:** 8-DAY AQUA/MODIS. (2018-2020). Retrieved Oct 25, 2021, from https://neo.gsfc.nasa.gov/view.php?datasetId=MY1DMW_CHLORA

- **Florida Bleach Watch Report:** Direct download from site at https://floridadep.gov/rcp/coral/content/bleachwatch

- **NASA Giovanni**: Data Collection via API described at https://giovanni.gsfc.nasa.gov/giovanni/

- **NOAA CoastWatch Coral Bleaching Monitoring Products**: Data Collection via API described at https://coastwatch.pfeg.noaa.gov/erddap/griddap/NOAA_DHW.html

- **Reefbase - A Global Information System for Coral Reefs:** Direct download from site at http://www.reefbase.org/main.aspx

- **UN Environment Programme World Conservation Monitoring Centre:** Direct download from website at https://www.unep-wcmc.org/resources-and-data