Predicting the Extent of Harmful Algae Blooms Using Satellite Imagery and Machine Learning

Team Members: Charles Hatch, Shyam Veerasankar, Benjamin Schmitt

Introduction

Harmful algae blooms (HABs) have become an increasingly concerning issue in recent years due to their negative impacts on water quality, ecosystem health, and human activities. By leveraging ML techniques, it may be possible to detect and delineate HABs more quickly and efficiently, enabling effective management strategies to mitigate their negative impacts. One of the most common types of algae responsible for HABs is cyanobacteria. Cyanobacteria produce toxins that are poisonous to humans and other animals. Cyanotoxins can cause a wide variety of adverse human health problems including gastrointestinal distress, dermatitis, liver failure, or even death of pets and livestock when they are exposed to water with high levels of toxins. Manual water sampling is often used to assess risk from cyanobacteria. Manual water sampling is accurate, but it is time consuming and generally cost prohibitive across large areas. The use of satellite imagery and computer algorithms to detect HABs shows great promise (Clark et al. 2017 and Mishra et al. 2019). Identifying the presence and estimated extent of HABs would allow for more targeted manual sampling effort and better warnings for drinking water systems and recreational water users.

Methods have previously been developed to delineate cyanobacterial HABs using satellite data. The most notable methods are the Cyanobacterial Index (CI) and a corrected version of the CI known as CIcyano (Wynne et al. 2008, Wynne et al. 2018). These methods rely on specific sensors, like the Ocean and Land Colour Imager found on European Space Agency's Sentinel-3A and Sentinel-3B, that are only found on a handful of satellites. This limits how frequently predictions can be made over a given area.

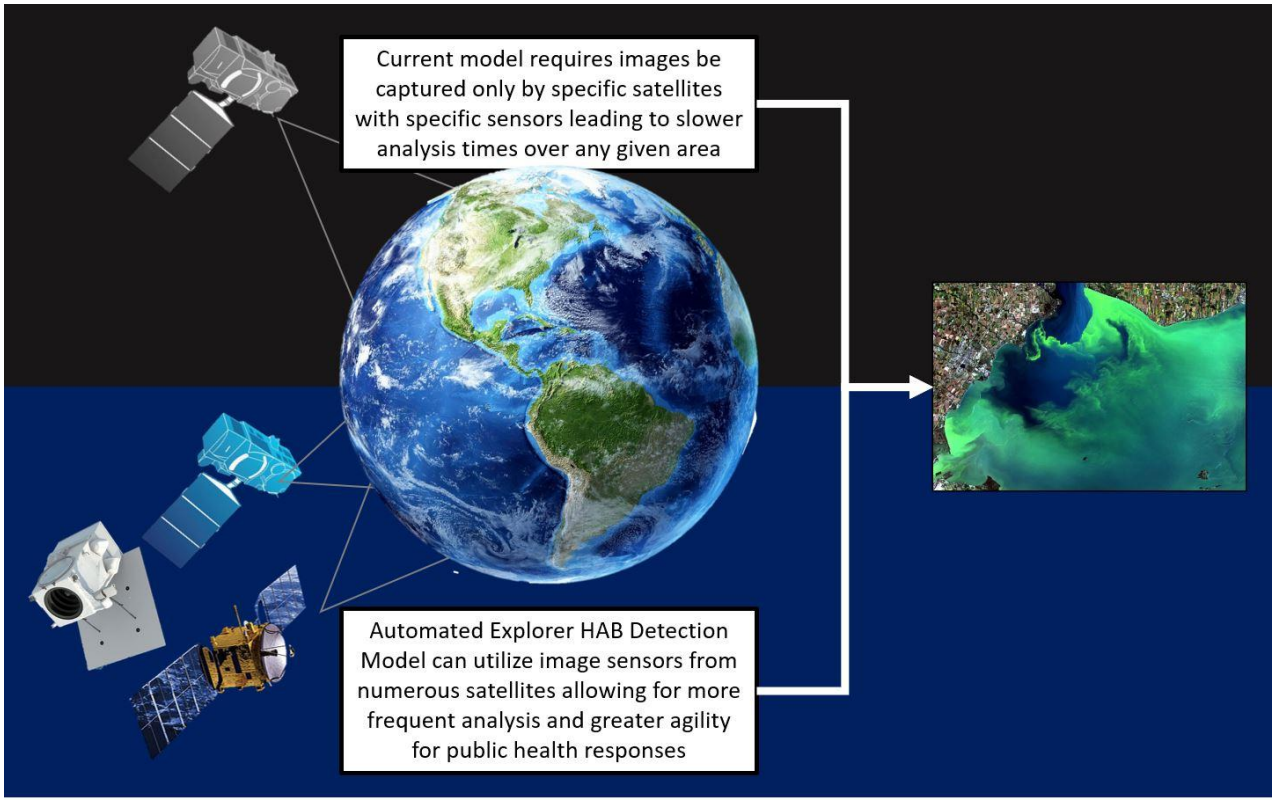


Figure 1: Comparison between current method of assessing HABs (top) vs proposed method by Team Explorer (bottom)

The cyanobacterial HAB detection method we developed could potentially use any color image as input. This allows for more frequent predictions over a given area because any satellite image or even drone images could be used. Providing information about HABs more quickly than was previously possible will help drinking water system managers and recreational water users make better informed decisions and reduce health risk.

Methods

Methods – Data

Image data used for this project was collected by the Copernicus Sentinel-3A and Sentinel-3B satellites in 2022 and early 2023. We focused on areas with known HAB problems in the Great Lakes region and Florida. The image data was accessed on the National Oceanographic and Atmospheric Administration (NOAA) NCCOS HAB Data Explorer webpage (NOAA 2023). These images were selected because corresponding images of predicted HAB extent, produced by NOAA using the CIcyano method, were also available.

The data was qualitatively cleaned to remove distortions, glare, poor image stitching, and images where the body of water was completely blocked out by overcast weather conditions.

Methods - Data Labeling and Class Remapping

The NOAA HAB data was processed into semantic maps utilizing the Python Image Library (PIL). In the original algal bloom identification labels, any concentration of algae was an RGB color while land and clouds were 3 channel grayscale. Data was separated into 3 classes in uint8 values (0-255) of 0 for Water, 1 for land/clouds, and 2 for algae. The original intent was to separate land and clouds, however, similar grayscale values were used for land outcroppings/clouds so they had to be combined as there was not a quantitative way to separate them.

As more than one type of analysis existed per image, the ground truth RGB images were duplicated as needed to ensure that each semantic map had a corresponding image to be matched for data loading.

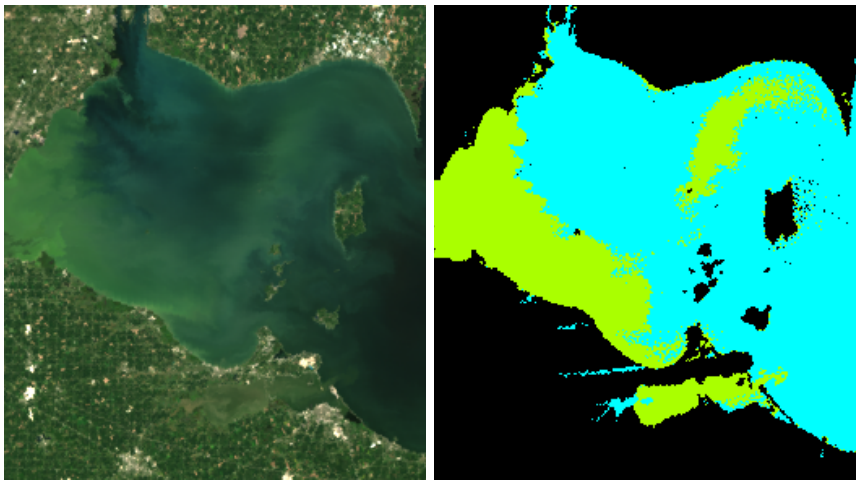


Figure 2: Image and color segmentation map pair. Land/clouds are black, cyan is water, and algae is lime green.

Methods - Pixel Analysis

To understand the distribution of pixel colors in our dataset we performed a pixel color count and channel value percentage analysis.

Methods - Model Selection and Training

We used a form of unsupervised semantic segmentation, known as Self-supervised Transformer with Energy-based Graph Optimization (STEGO), to identify clusters of related pixels in satellite images. The STEGO model was developed by Hamilton et al. (2022) and is described in a paper titled ‘Unsupervised Semantic Segmentation by Distilling Feature Correspondences’. We experimented with using the pre-trained model developed by Hamilton and with training the model on our own image data.

Source code from the author was heavily edited to reflect config parameters.

Methods - Evaluation

Using pixel classes (unique colors) generated by the STEGO model, we evaluated if any of the predicted classes corresponded to cyanobacterial blooms using quantitative and qualitative methods. To evaluate our predictions qualitatively, we plotted the location of manual samples that tested positive for cyanobacterial toxins over images of our predictions. This provided a quick visual check to see if any of the STEGO predicted classes matched the manual sample results.

Quantitative evaluation of our predictions was accomplished by comparing our results to predictions of cyanobacterial bloom extent using a corrected version of the Cyanobacterial Index (CI) known as CIcyano. The CI was first developed by Wynne et al. (2008) as the Spectral Shape around 681 nm. It was developed in Lake Erie using MERIS satellite data to detect large blooms of cyanobacteria. Development of CIcyano was driven by the observation that some other algae species blooms yielded a positive CI value even when there were no reported occurrences of cyanobacteria (Wynne et al. 2018). Each color image in our dataset has a corresponding CIcyano prediction image. Using the CI predictions we were able to calculate the intersection over union (IoU) for each of our predicted classes.

Results

Results - Pixel Analysis

Our image data was analyzed to determine the distribution of pixel colors present. The image below depicts the observed counts of the top 600 unique pixel colors in a sample of our dataset.

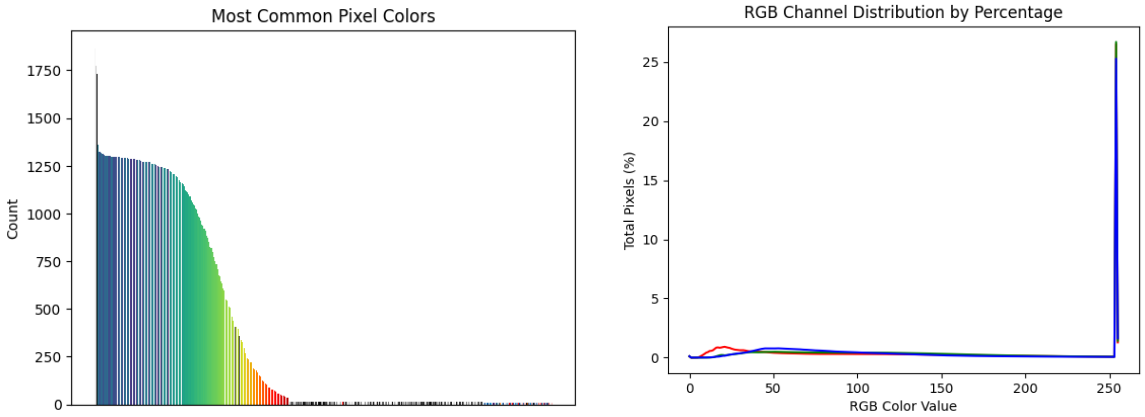


Figure 3: Color Counts and RGB Channel Distribution by percentage over the combined dataset. The presence of high color channel values on the right indicates that we have bolder, darker colors as can be seen in the color counts on the left.

Not surprisingly, since we were using images of waterbodies, our dataset contained many dark blue and green pixel colors. The most common pixel colors observed were light gray colors probably associated with clouds.

Results - Semantic Map Analysis

The semantic maps were analyzed using PIL by extracting each pixel in the image, summing them by class, and then dividing them by the total number of pixels to get the percentage of pixels by class. Color Semantic maps were also generated for visualization purposes.

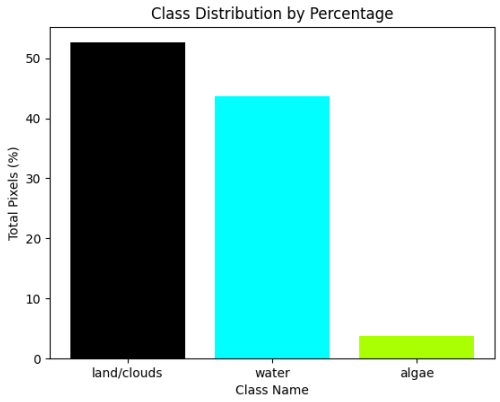


Figure 4: Class distribution by percentage, Water @ ~43.65%, land/clouds @ ~52.60% and algae @ ~3.74%. This indicates a large class imbalance for the target algae class.

Results - Model Selection and Evaluation

The STEGO model trained on images of HABs was able to achieve some promising accuracy and mean IoU results over all pixel classes.

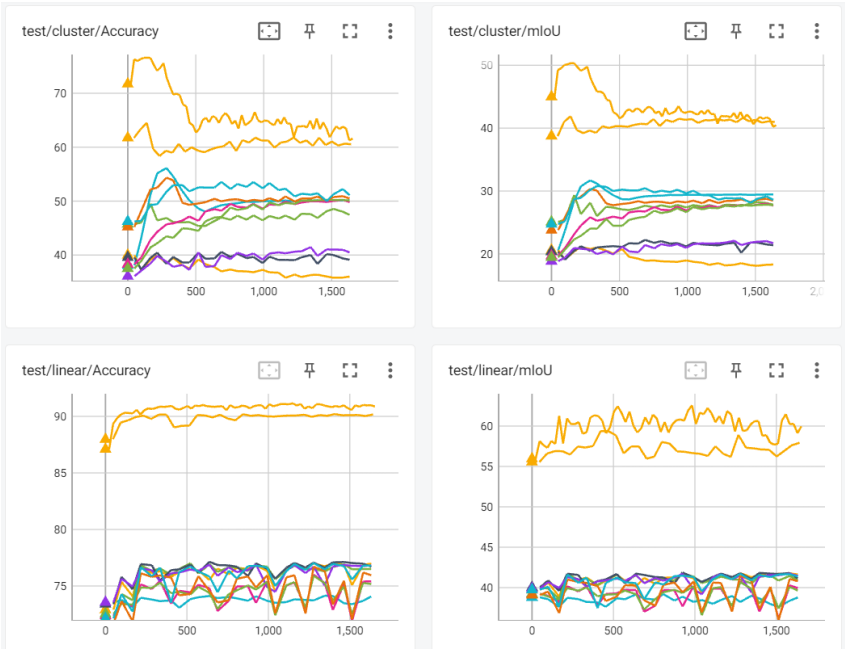


Figure 5: Model performance metrics taken from tensorboard trained with STEGO on Great Lakes and Florida HAB data. The top orange plot line was trained only on Great Lakes data, while the second from the top orange line was trained on combined Great Lakes and Florida data.Cluster represents the KNN output while linear represents post CRF denoising.

Both the pre-trained STEGO models and the STEGO models we trained on HAB images did not achieve predictions comparable to the CIcyano method when considering only the algae class. However, the STEGO models did seem to group some pixels associated with HABs into the same class. Limiting the input images to include only those with few clouds improved model performance.

The chart below depicts a comparison of model performance, measured as mean IoU, on a subset of our image data. The results depicted below are only for the algae related pixels. The images used for prediction were restricted to weeks with positive cyanobacterial detections and few clouds. Both of these conditions increased the correct identification of HAB related pixels.

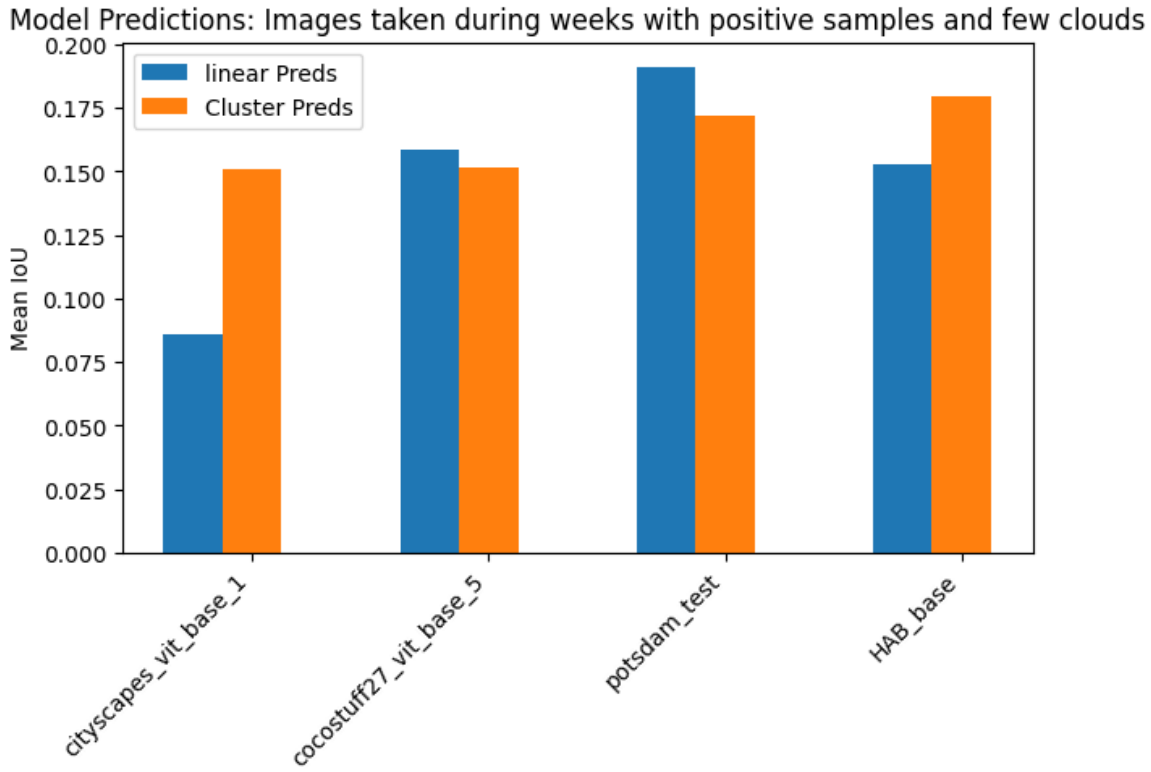


Figure 6: Evaluation of Great Lakes and Florida HAB data on pre-trained models compared to the current model. The potsdam dataset contains aerial imagery that may be similar to our HAB dataset

As shown in the figure above, the STEGO model pretrained on the potsdam dataset obtained the highest mean IoU when considering only the algae class.

The Images below show an example of model output compared to original image and the predicted HAB extent using the CIcyano method. These images were collected on August 31, 2022 and depict western Lake Erie. As you can see the STEGO model seems to group pixels associated with HABs into the same two classes. Possibly even grouping pixels with similar

algal concentrations into the same class. However, the predicted classes do not match very well with the HAB extent predicted using the CIcyano method.

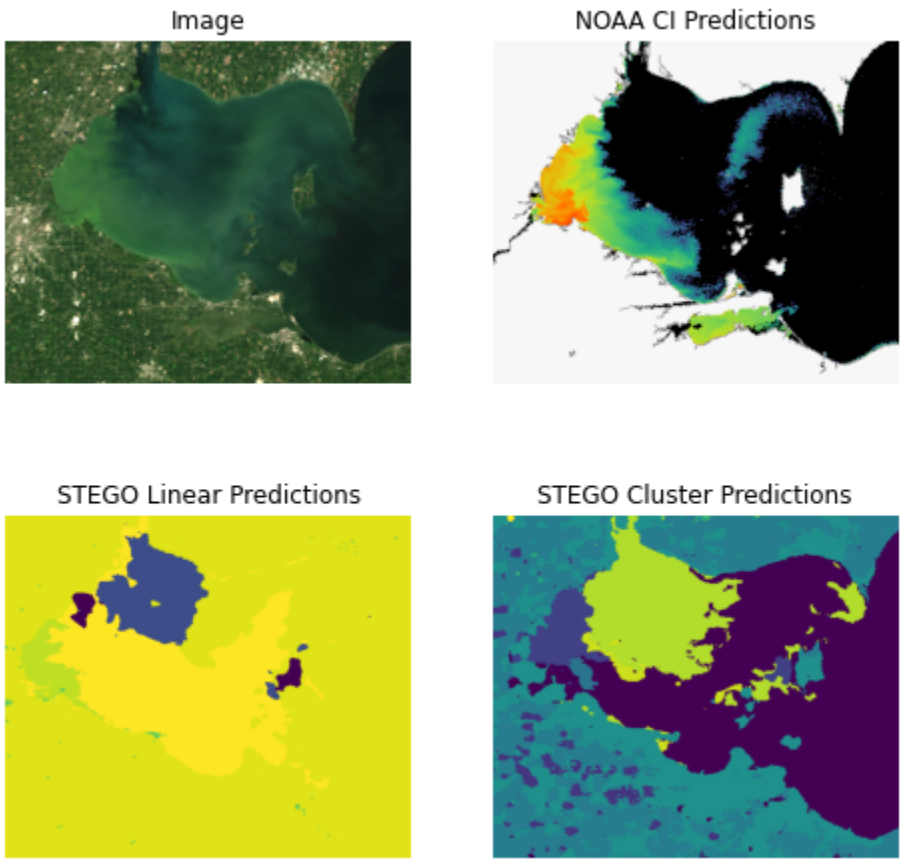


Figure 7: HAB image (top left) with CI-Cyano label (top right) with STEGO CRF prediction (bot. left) and STEGO knn cluster prediction (bot right). The linear prediction captures some of the highest algal bloom concentrations, but not the lesser.

The image below is a model prediction with the location of manual samples that tested positive for cyanobacterial toxins depicted using red dots. This is a prediction made using the pretrained model from the same August 31, 2022 image in western Lake Erie. This image shows that the STEGO model predicts one class that seems associated with the HAB but is not capturing the entire HAB in the same predicted class.

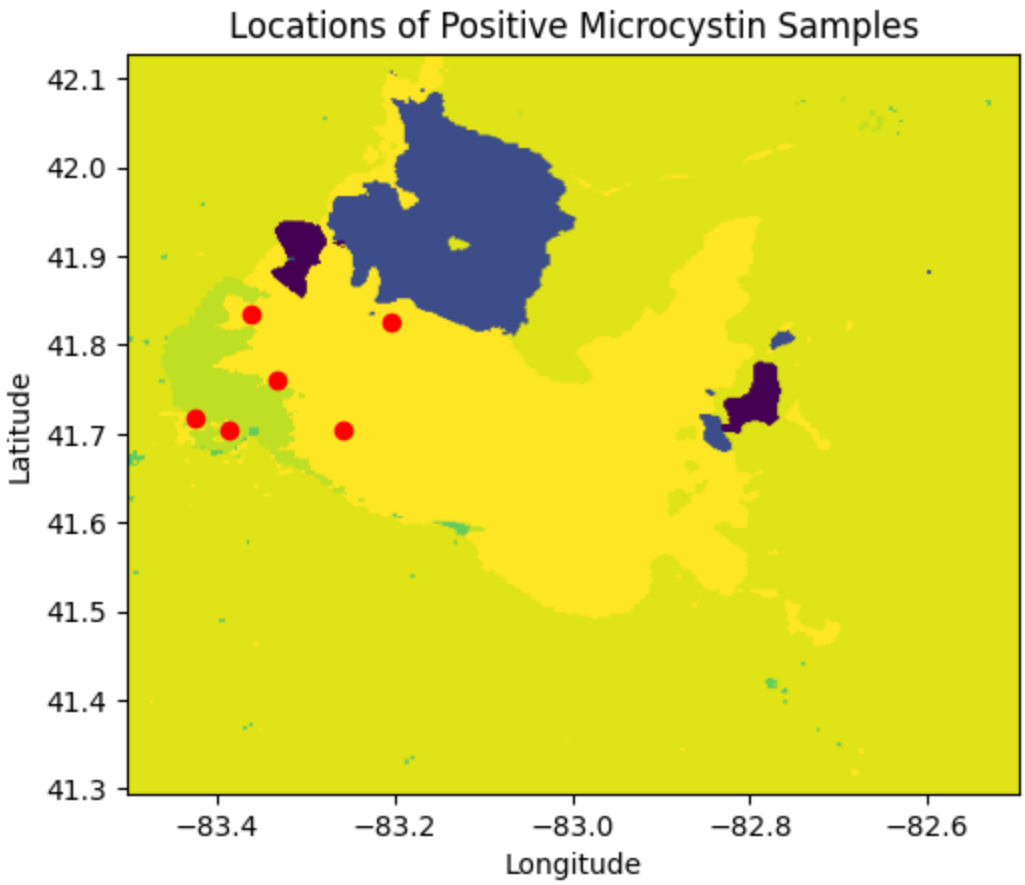


Figure 8: GPS coordinates of manual sampling in Western Lake Erie, indicated by red dots in the image above.

Discussion

Overall the STEGO model trained on images of HABs performed worse than existing methods of HAB extent prediction using satellite imagery. However, examination of predictions revealed that the STEGO model did seem to be assigning pixels associated with HABs into the same one or two predicted classes. Predictions could likely be improved with additional training data and model hyperparameter tuning.

We faced several challenges that limited our ability to achieve good model performance. There was a major class imbalance in our data. Pixels corresponding to land and water were much more common than pixels corresponding to HABs. The class imbalance combined with limited training data probably hampered our models ability to correctly identify HABs. Clouds in many images were also problematic for training and prediction. This makes our method less useful in cloudy places compared to places with fewer cloudy days.

We faced other challenges in addition to a class imbalance and limited training data. The STEGO model required a major rework to be useful for our task. Much of our time was spent adapting the model to our task rather than focusing on improving performance. We hypothesize

that the model can be further tuned by modifying certain hyperparameters. The following hyperparameters would be the set with which we would run a GridSearch and further refine the model performance: number of steps(max_steps), number of neighbors(num_neighbors), number of classes(dir_dataset_n_classes), learning rate(lr), batch size(batch_size), and model type(model_type).

Using STEGO to predict the extent of HABs shows promise but additional work is needed to improve predictions and validate results. It is not clear that the model we trained will correctly identify HABs in other locations outside of locations where we had image data. Future work would benefit from using state of the art networks such as Facebook's Segment Anything Model, with class down selection similar to our process (Alexander Kirillov et al, 2023).

The images we used were collected near populated areas over water bodies with many public water system intakes and lots of recreational use. The consequences of making poor predictions could be quite high. Before model results were provided to water system managers or the public, the model would need rigorous evaluation. The addition of a buffer around predictions might also be prudent to decrease public health risks.

References

Alexander Kirillov et al, 2023. Segment Anything. https://ai.facebook.com/research/publications/segment-anything/

Clark, J. M., Blake A. Schaeffer, John A. Darling, Erin A. Urquhart, John M. Johnston, Amber R. Ignatius, Mark H. Myer, Keith A. Loftin, P. Jeremy Werdell, Richard P. Stumpf. 2017. Satellite monitoring of cyanobacterial harmful algal bloom frequency in recreational waters and drinking water sources. Ecological Indicators 80: 84-95.

"HAB Data Explorer". National Oceanographic and Atmospheric Administration (NOAA), https://products.coastalscience.noaa.gov/habs_explorer/index.php. Accessed March 15, 2023.

"Learn about Cyanobacteria and Cyanotoxins." United States Environmental Protection Agency (US EPA) , https://www.epa.gov/cyanohabs/learn-about-cyanobacteria-and-cyanotoxins. Accessed March 20, 2023.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., & Freeman, W. T. 2022. Unsupervised semantic segmentation by distilling feature correspondences. arXiv preprint arXiv:2203.08414.

Mishra, S., Stumpf, R.P., Schaeffer, B.A., Werdell, J., Loftin, K. A., & Meredith. A. 2019. Measurement of Cyanobacterial Bloom Magnitude using Satellite Remote Sensing. Sci Rep 9: 18310. https://doi.org/10.1038/s41598-019-54453-y

Wynne, T.T.; Stumpf, R.P.; Tomlinson, M.C.; Warner, R.A.; Tester, P.A.; Dyble, J.; Fahnenstiel,

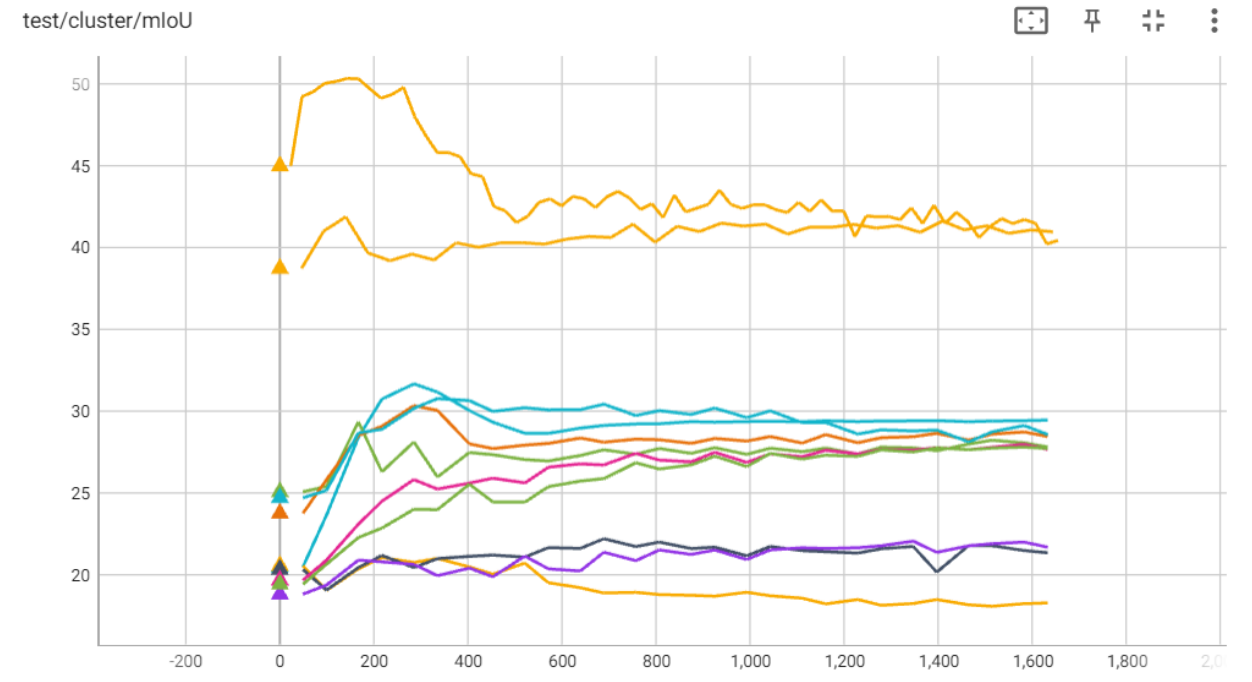
G.L. 2008. Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. International Journal of Remote Sensing 29: 3665–3672.

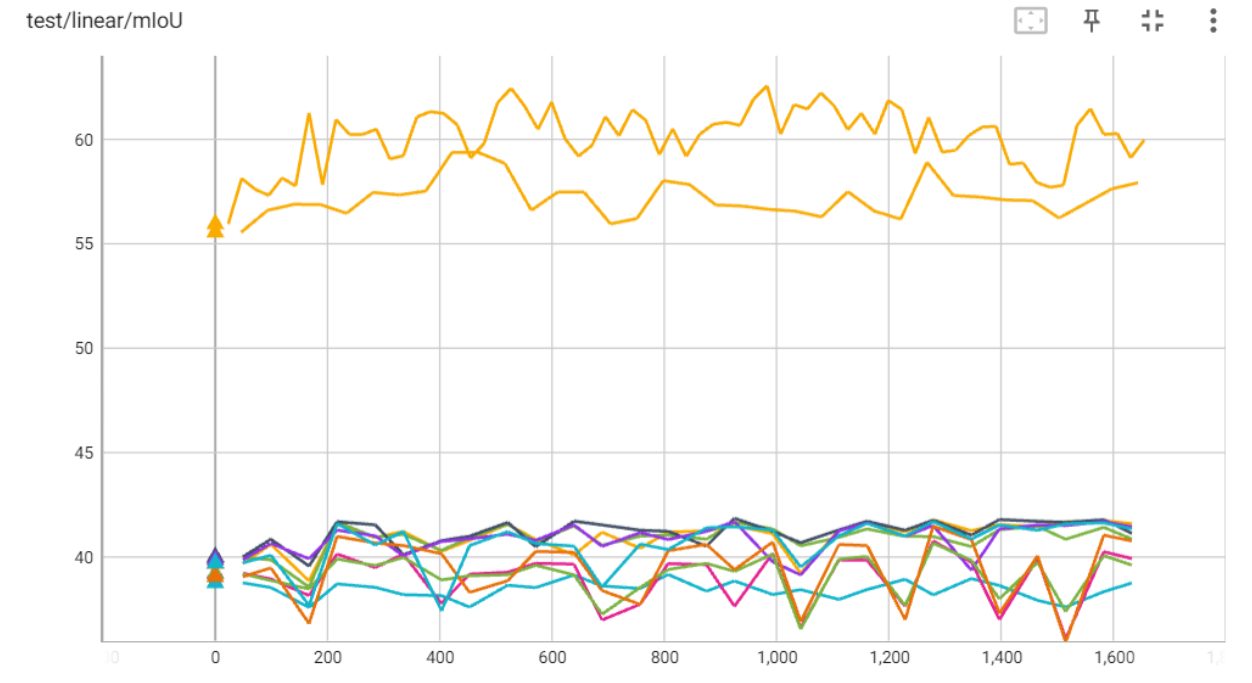
Wynne, T., A. Meredith, T. Briggs, W. Litaker, and R. Stumpf 2018. Harmful Algal Bloom Forecasting Branch Ocean Color Satellite Imagery Processing Guidelines. NOAA Technical Memorandum NOS NCCOS 252. Silver Spring, MD. 48 pp. doi:10.25923/twc0-f025

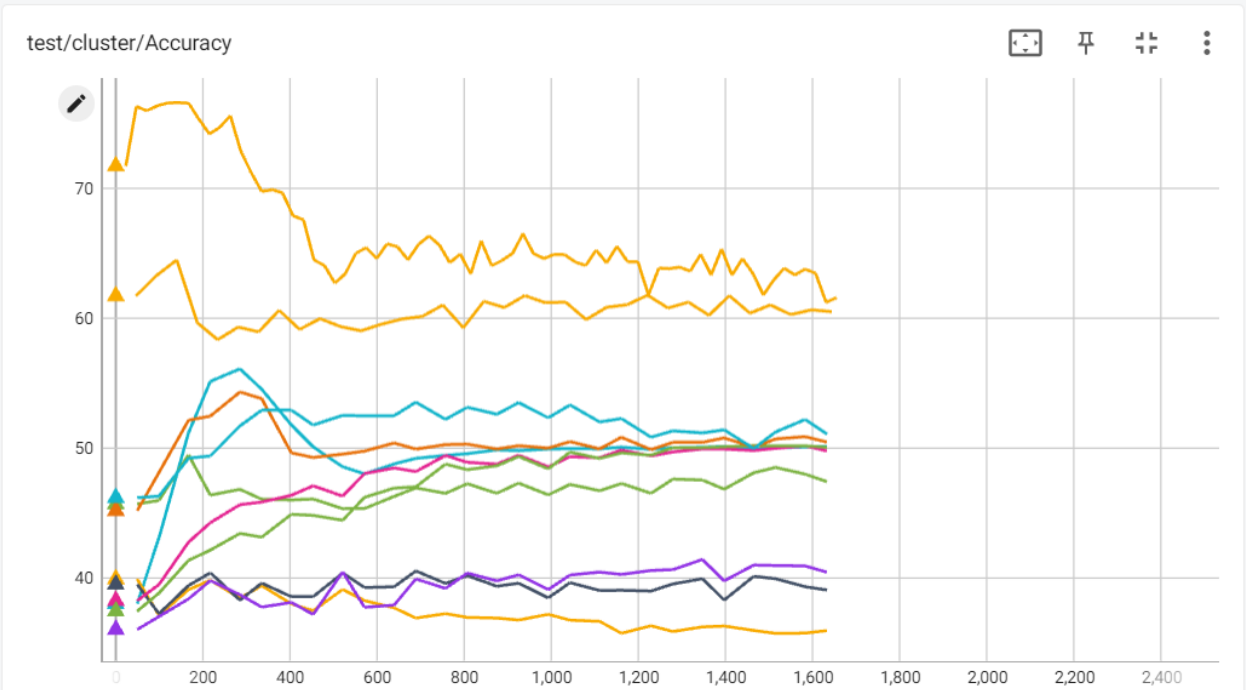
Statement of Work:

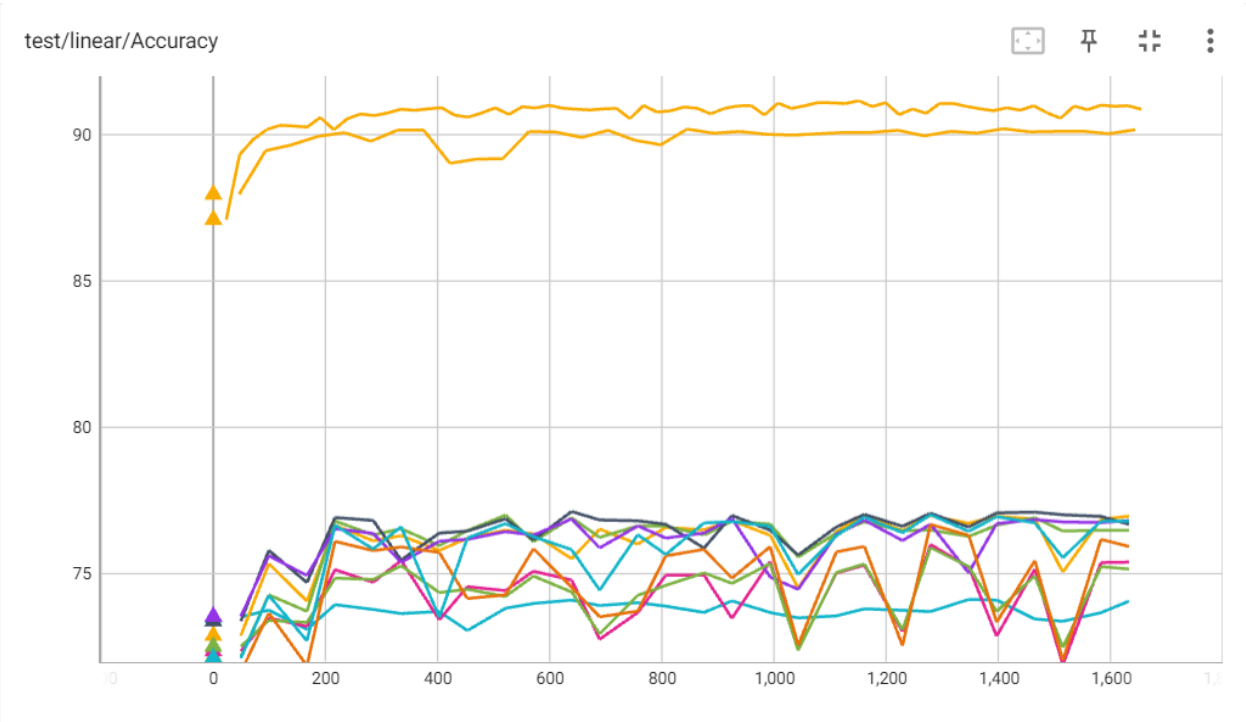
- Benjamin Schmitt
    - Image data collection
    - Model training and evaluation
    - Storytelling and visualization
- Charles Hatch
    - Rework of the STEGO repository
    - Color channel pixel analysis
    - Semantic map creation and formatting data for training
- Shyam Veerasankar
    - Image data collection
    - Model Training
    - Peer reviewer
    - Visualizations

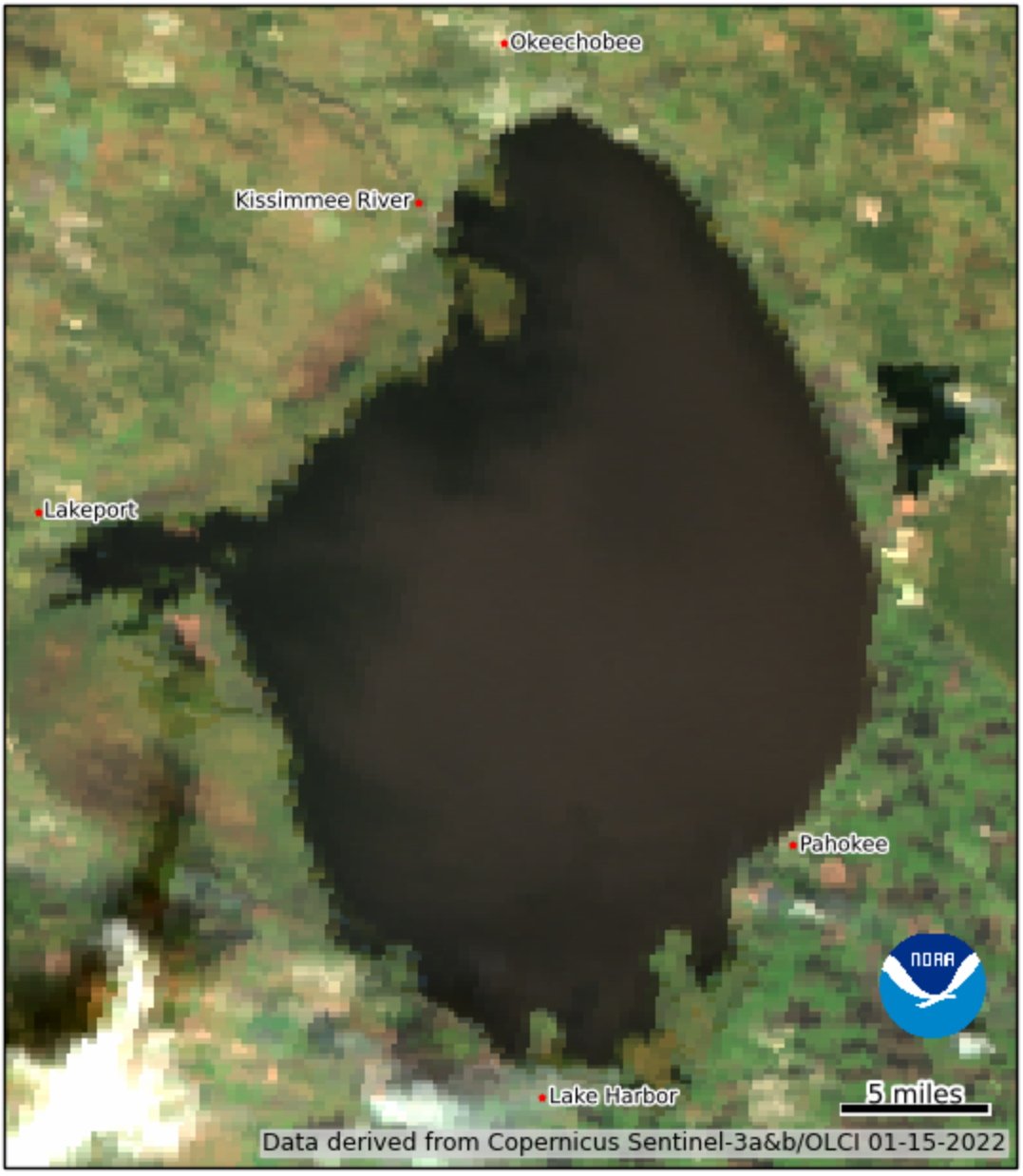
Appendix:


test/cluster/mIoU
50
45
40
35
30
25
20
-200
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800

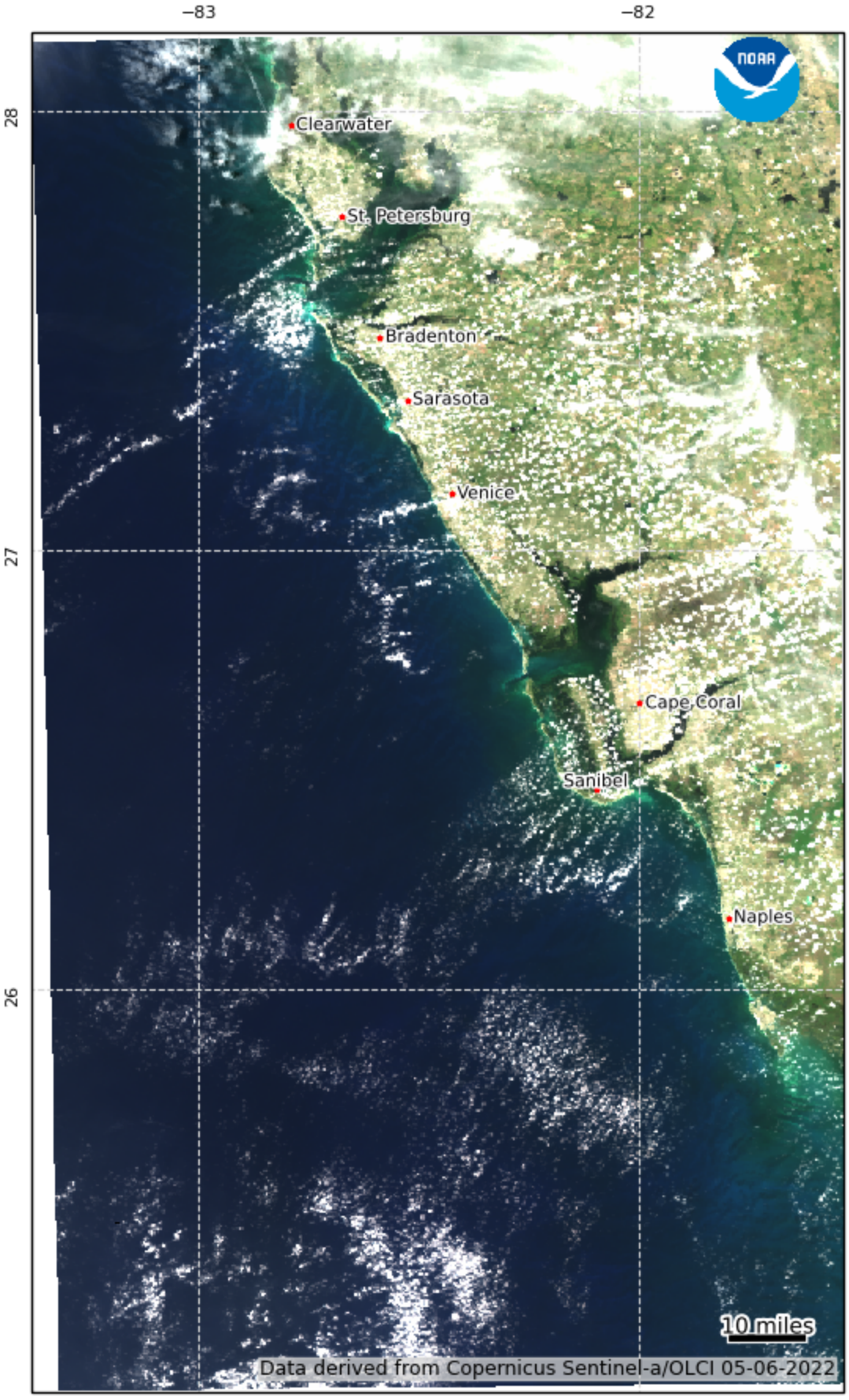

test/linear/mIoU
60
55
50
45
40
0
200
400
600
800
1,000
1,200
1,400
1,600


test/cluster/Accuracy
70
60
50
40
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400

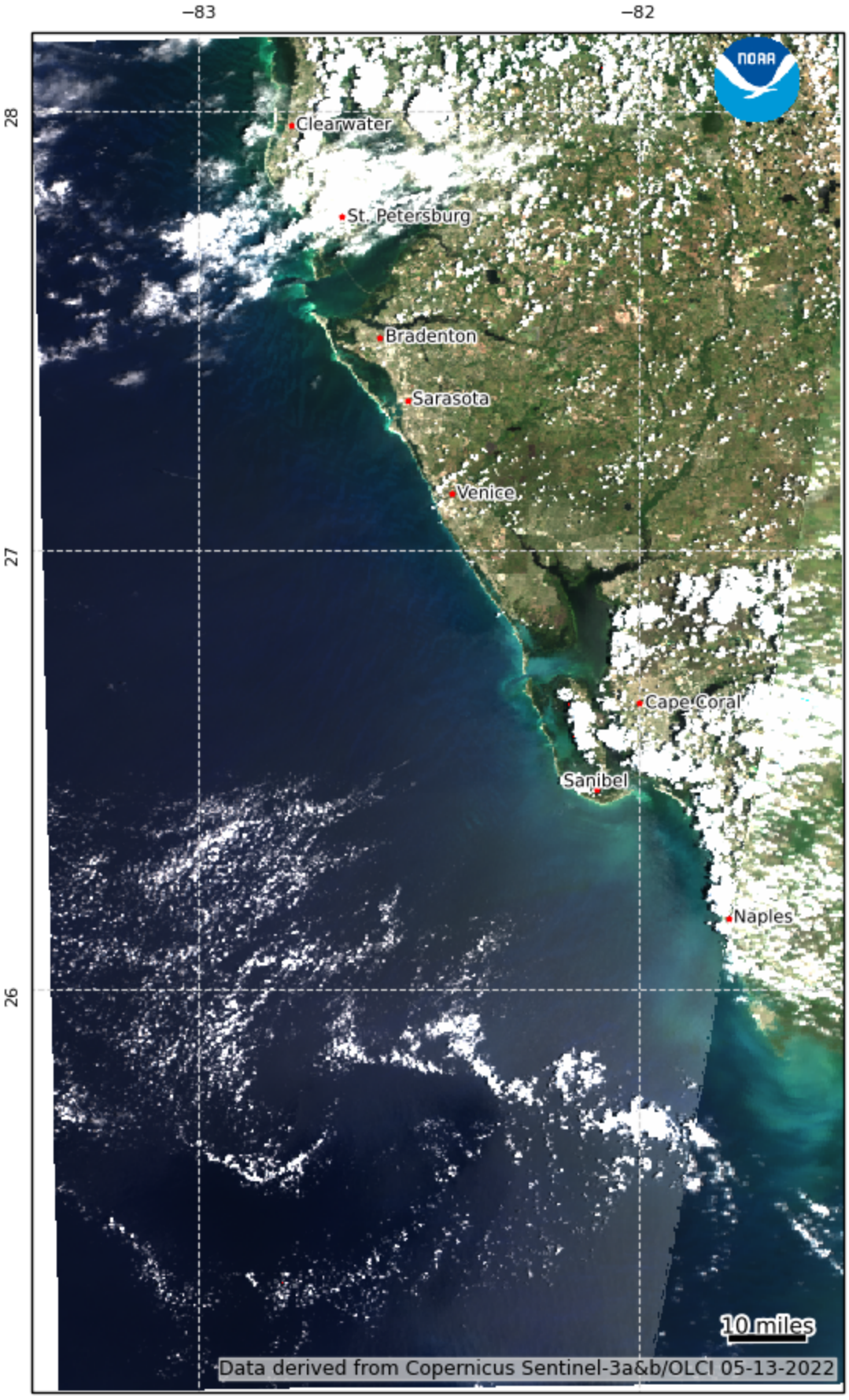

test/linear/Accuracy
90
85
80
75
0
200
400
600
800
1,000
1,200
1,400
1,600



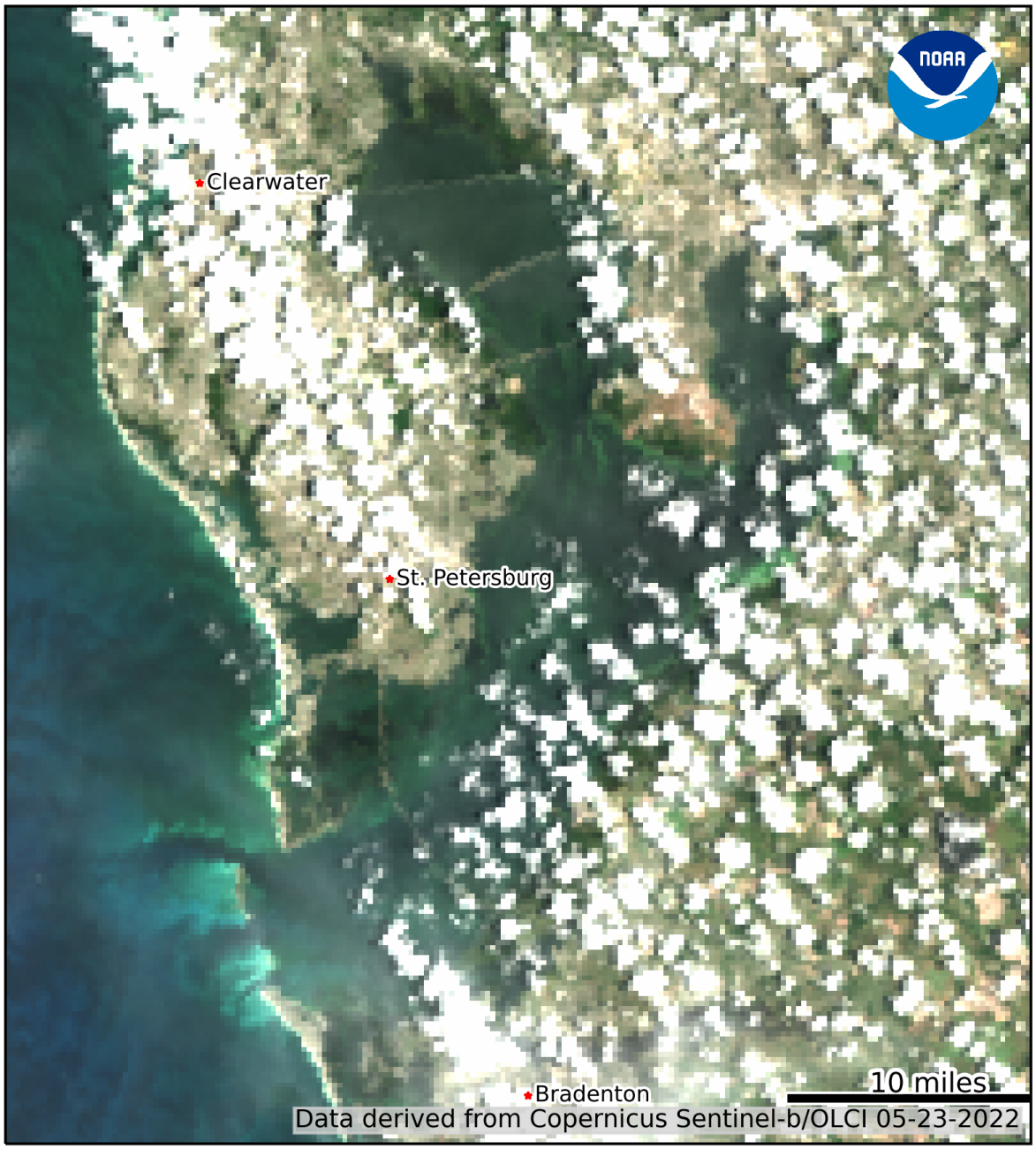
Composited Lake Okeechobee true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.



Southwest Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-a obtained from EUMETSAT.

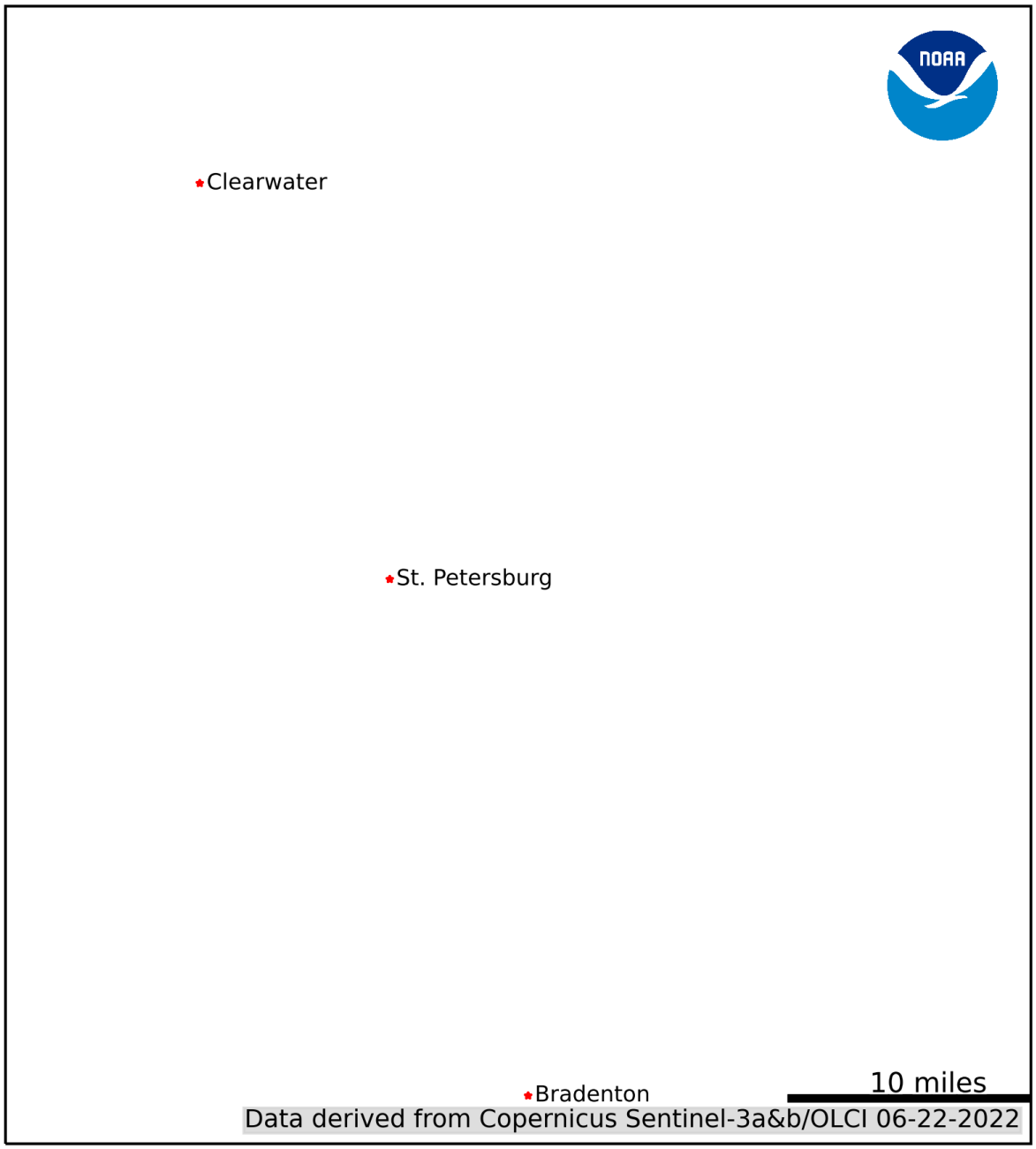


Composited Southwest Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.

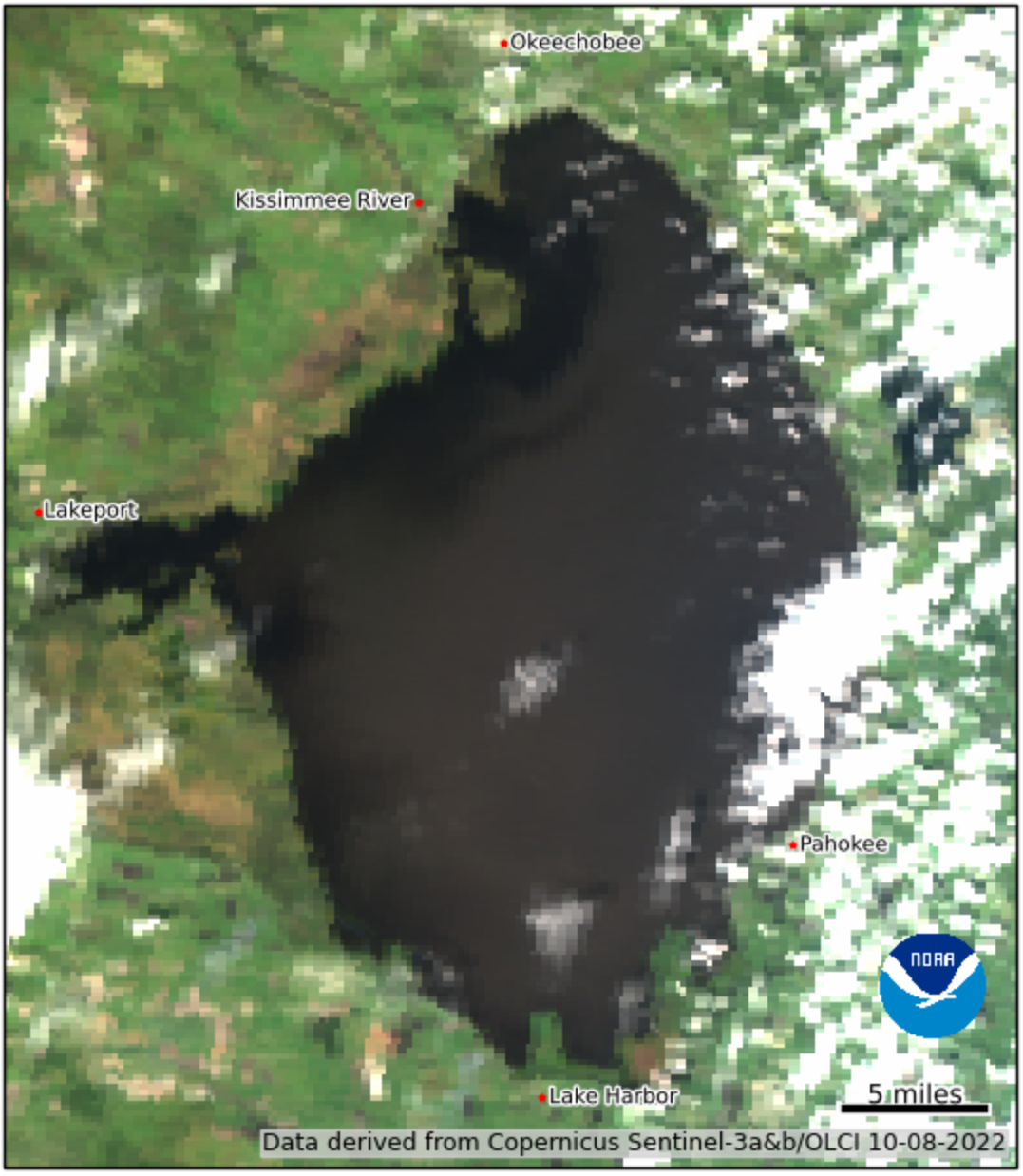


Tampa Bay, Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-b obtained from EUMETSAT.

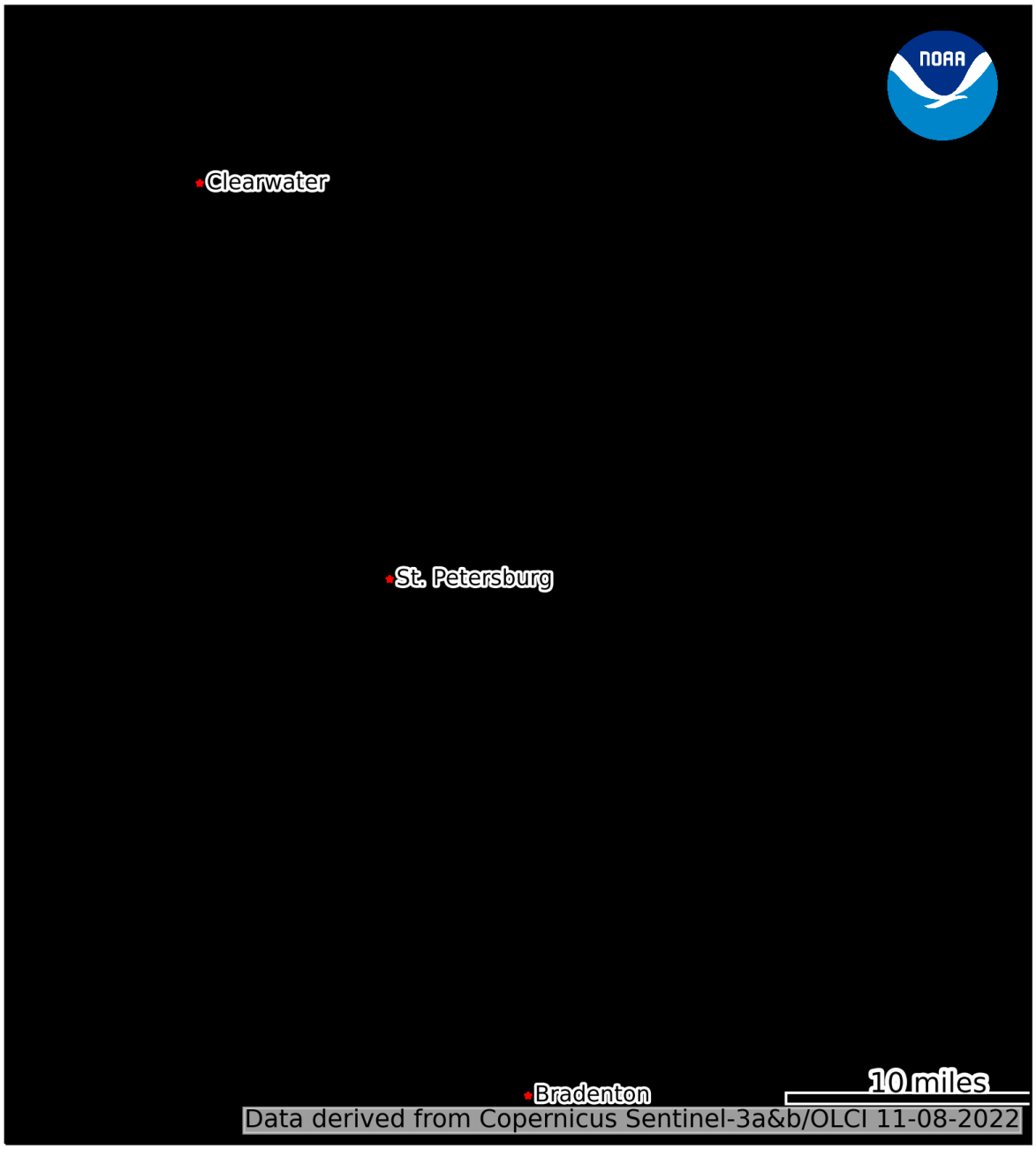
Tampa Bay, Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-a obtained from EUMETSAT.



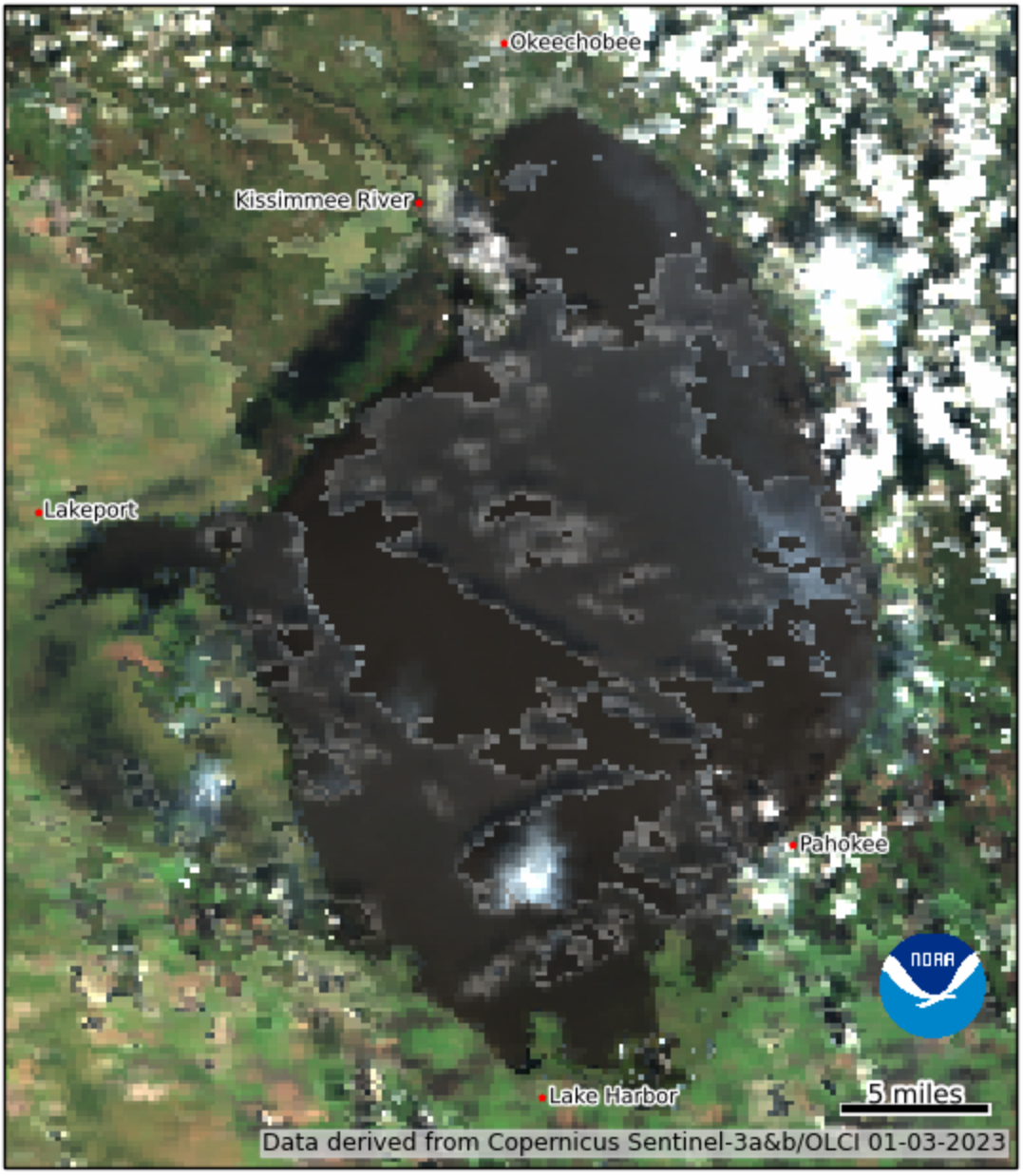
Composited Tampa Bay, Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.



Composited Lake Okeechobee true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.



Composited Tampa Bay, Flordia true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.



Composited Lake Okeechobee true color image derived from the OLCI sensor on Copernicus Sentinel-3a&b obtained from EUMETSAT.