

Cyanobacteria Detection from Satellite Data

Using Sentinel-2 for small, inland water bodies

Our work

DrivenData has been collaborating with NASA to estimate cyanobacteria in small, inland water bodies with the ultimate goal of helping water quality managers better allocate resources and make more informed decisions around public health warnings.

Context

Harmful algal blooms (HABs) occur in lakes and reservoirs across the U.S. and threaten human and animal health, marine habitats, and recreation and economic opportunities.

“In situ” sampling is common and accurate, but time intensive to perform regularly.

There are methods to detect HABs from satellite data for large water bodies like oceans, but not small ones like lakes and reservoirs.

Phase 1

- Run machine learning competition to crowdsource top modeling approaches

Phase 2

- Carry winning models forward into a deployment-ready code package

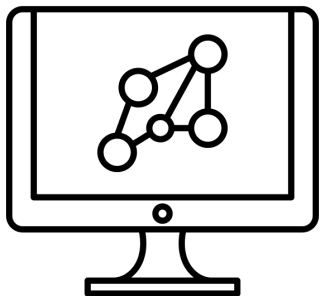
Goals

- Share capabilities of DrivenData-developed code package
- Build shared understanding of current CyAN capabilities and products
- Explore possible integration of Sentinel-2 based models into CyAN

Estimating cyanobacteria in
small, inland water bodies

Phase 1

Machine learning competition

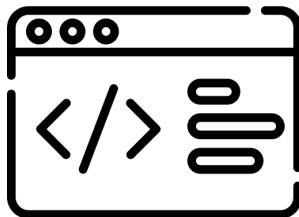


Ran a machine learning competition to develop models that detect and classify the severity of cyanobacteria blooms in small, inland water bodies.

Machine learning competitions are excellent for crowd-sourcing top approaches to complex predictive modeling problems.

Phase 2

Deployment-ready code package



Static research code needs to be transformed into a code package that is capable of generating predictions on new input data.

A production-ready code pipeline is efficient, generalizable, well-documented, and tested.

Handoff

Regular cyanobacteria predictions

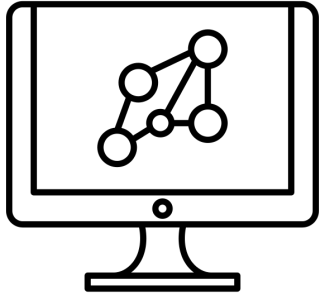


The desired outcome is accessible, regularly generated predictions of cyanobacteria levels for given latitude and longitude points in inland water bodies across the US.

This enables ongoing detection of unsafe bacteria counts to inform advisories and protect public health and safety.

Phase 1

Machine learning competition

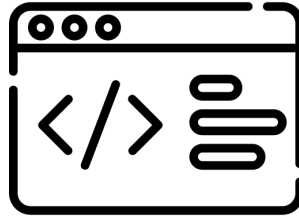


Ran a machine learning competition to develop models that detect and classify the severity of cyanobacteria blooms in small, inland water bodies.

Machine learning competitions are excellent for crowd-sourcing top approaches to complex predictive modeling problems.

Phase 2

Deployment-ready code package



Static research code needs to be transformed into a code package that is capable of generating predictions on new input data.

A production-ready code pipeline is efficient, generalizable, well-documented, and tested.

Handoff

Regular cyanobacteria predictions



The desired outcome is accessible, regular predictions of cyanobacteria levels for given latitude and longitude points in inland water bodies across the US.

This enables ongoing detection of unsafe bacteria counts to inform advisories and protect public health and safety.

Summary of code package

Overview

- Python package that generates an estimate of cyanobacteria severity for a given latitude, longitude, and date
 - “Nowcasting” that gives you a view of conditions on the ground
 - Estimates come from a machine learning model that uses Sentinel-2, climate, and elevation data
- Intended for use in small, inland water bodies where higher resolution imagery is needed
- Pip-installable python package that can be run on a user’s laptop or as part of a cloud-hosted deployment pipeline that outputs to a public dashboard

Summary of code package

	date	latitude	longitude
0	2021-05-18	35.650000	-78.682816
1	2018-10-22	37.564318	-101.335575
2	2021-05-17	36.050000	-76.700000
3	2016-08-31	35.705416	-79.164659
4	2015-06-27	41.287577	-80.424543

Input csv of
sample points

cyano predict input.csv

```
2023-08-11 15:50:48.147 | INFO | cyano.pipeline:_prep_predict_data:130 - Loaded 5 samples
for prediction
2023-08-11 15:50:48.147 | INFO | cyano.data.satellite_data:generate_candidate_metadata:19
2 - Generating metadata for all satellite item candidates
2023-08-11 15:50:48.147 | INFO | cyano.data.satellite_data:generate_candidate_metadata:21
1 - Searching ['sentinel-2-l2a'] within 30 days and 1000 meters
100%|████████████████████████████████████████████████████████████████████████████████| 5/5 [00:02<00:00, 1.76it/s]
2023-08-11 15:50:50.998 | INFO | cyano.data.satellite_data:generate_candidate_metadata:24
1 - Generated metadata for 46 Sentinel item candidates
2023-08-11 15:50:50.999 | INFO | cyano.data.satellite_data:identify_satellite_data:294 -
Selecting which items to use for feature generation
100%|████████████████████████████████████████████████████████████████████████████████| 5/5 [00:00<00:00, 197.52it/s]
2023-08-11 15:50:51.027 | INFO | cyano.data.satellite_data:identify_satellite_data:313 -
Identified satellite imagery for 4 samples
2023-08-11 15:50:51.028 | INFO | cyano.pipeline:_prepare_features:59 - 4 rows of satellit
e metadata saved to /tmp/tmpwoa5aufk/satellite_metadata_train.csv
2023-08-11 15:50:51.028 | INFO | cyano.data.satellite_data:download_satellite_data:338 -
Downloading bands ['B02', 'B03', 'B04']
100%|████████████████████████████████████████████████████████████████████████████████| 4/4 [01:05<00:00, 16.37s/it]
2023-08-11 15:51:56.494 | SUCCESS | cyano.pipeline:_prepare_features:71 - Raw source data sa
ved to /tmp/tmpwoa5aufk
2023-08-11 15:51:56.494 | INFO | cyano.data.features:generate_satellite_features:49 - Gen
erating features for 5 samples
100%|████████████████████████████████████████████████████████████████████████████████| 5/5 [00:00<00:00, 1128.05it/s]
2023-08-11 15:51:56.502 | INFO | cyano.data.features:generate_satellite_features:90 - Fil
ling missing satellite values for 1 samples
2023-08-11 15:51:56.507 | INFO | cyano.data.features:generate_features:179 - Generated 7
satellite features
2023-08-11 15:51:56.509 | SUCCESS | cyano.pipeline:_prepare_features:77 - 7 features for 5 s
amples saved to /tmp/tmpwoa5aufk/features_train.csv
2023-08-11 15:51:56.517 | SUCCESS | cyano.pipeline:_write_predictions:149 - Predictions save
d to preds.csv
```

	date	latitude	longitude	severity
0	2021-05-18	35.650000	-78.682816	2
1	2018-10-22	37.564318	-101.335575	2
2	2021-05-17	36.050000	-76.700000	2
3	2016-08-31	35.705416	-79.164659	2
4	2015-06-27	41.287577	-80.424543	2

Output csv with
estimated
cyanobacteria severity
level

Command line tool takes one line of code to run

Summary of code package

Key benefits

- Models use 10m imagery resolution
 - Ability to estimate cyanobacteria in small, inland lakes, rivers, and reservoirs
 - These may look like land pixels at 300m resolution
 - Less affected by cloudiness
 - More pixels per water body means an increased chance of having cloud free pixels
- Straightforward to get estimates
 - Easily configured to generate estimates for routine sampling locations (provide csv with lat, lon, date)
 - Easy to run batch queries (just add more rows)

Summary of code package

Example end-user workflows this supports

Allocating staff between routine sampling locations

Example: Florida routinely monitors many locations, but is constrained by number of trained staff to collect and analyze samples.

→ Regularly generated predictions help determine which locations are highest risk, enabling a more strategic allocation of finite sampling resources

Flagging new, non-routine incidence of algal blooms to supplement public reports

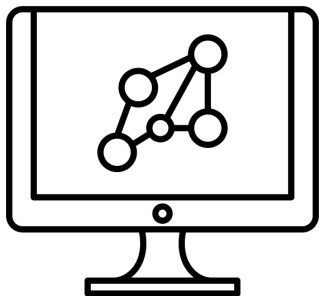
Example: Georgia relies on the public to trigger algal bloom sampling, and therefore has a limited view of blooms across the state

→ Regularly generated predictions across state water bodies helps to flag new areas of high risk and potentially confirm publicly reported blooms

Under the hood

Phase 1

Machine learning competition

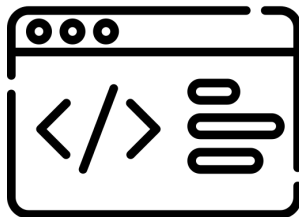


Ran a machine learning competition to develop models that detect and classify the severity of cyanobacteria blooms in small, inland water bodies.

Machine learning competitions are excellent for crowd-sourcing top approaches to complex predictive modeling problems.

Phase 2

Deployment-ready code package



Static research code needs to be transformed into a code package that is capable of generating predictions on new input data.

A production-ready code pipeline is efficient, generalizable, well-documented, and tested.

Handoff

Regular cyanobacteria predictions



The desired outcome is accessible, regular predictions of cyanobacteria levels for given latitude and longitude points in inland water bodies across the US.

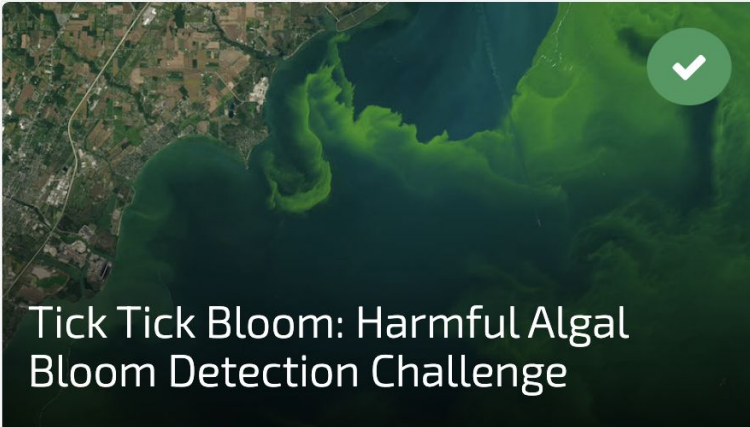
This enables ongoing detection of unsafe bacteria counts to inform advisories and protect public health and safety.

Machine learning competition

Goal: Detect and classify the severity of cyanobacteria blooms in small, inland water bodies

Participants used the following input data:


- Satellite imagery (Landsat or Sentinel-2)
- Climate data including temperature, wind, and precipitation (NOAA's HRRR)
- Elevation data (Copernicus DEM)



Tick Tick Bloom: Harmful Algal Bloom Detection Challenge

COMPETITION HAS ENDED **\$30,000**

Harmful algal blooms occur all around the world, and can harm people, their pets, and marine life. Use satellite imagery to detect dangerous concentrations of cyanobacteria, and help protect public health!

**sheep**
1ST PLACE

RESULTS →

Launch date: Dec 14, 2022

Submission closed: Feb 17, 2023

Machine learning competition

Model predictions were evaluated against manual cyanobacteria “in situ” measurement data from 14 data providers across the U.S., which were aggregated by NASA.

- ~17k labeled points provided for training
- ~6k labeled data points used for evaluation

Participants predicted a severity level category

severity level	Density range (cells/mL)
1	<20,000
2	20,000 — <100,000
3	100,000 — <1,000,000
4	1,000,000 — <10,000,000
5	≥10,000,000

Modeling approaches

All winning models used **tree-based methods**

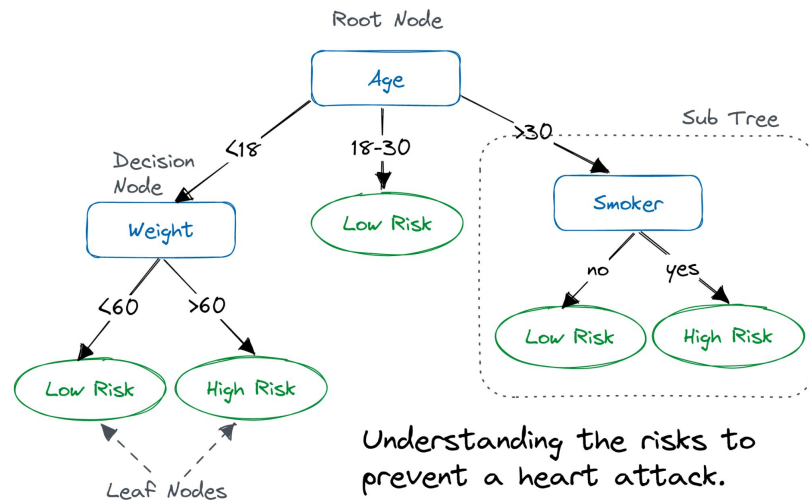
- Light GBM
- XGBoost
- CatBoost

Benefits

- Quick inference time once features are created
- Doesn't require a GPU

Drawbacks

- Produces only point-estimates, not a continuous heatmap



Mock example of a decision tree model

Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000

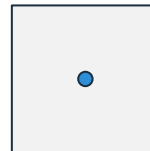


Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon
 - Specify **bounding box** around point
 - 200m to 2,500m range

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



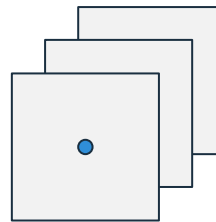
Satellite imagery

Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon
 - Specify bounding box around point
 - 200m to 2,500m range
 - Specify **time window** of imagery prior to sample date
 - 15-60 day window range

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



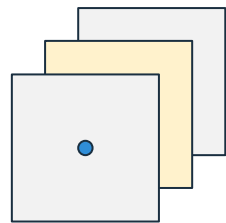
Satellite imagery

Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon
 - Specify bounding box around point
 - 200m to 2,500m range
 - Specify time window of imagery prior to sample date
 - 15-60 day window range
 - Select **least cloudy** image
 - e.g. cloud_cover property

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



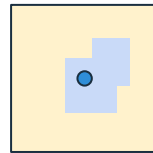
Satellite imagery

Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon
 - Specify bounding box around point
 - 200m to 2,500m range
 - Specify time window of imagery prior to sample date
 - 15-60 day window range
 - Select least cloudy image
 - e.g. cloud_cover property
 - **Filter to water** area
 - e.g. using scene classification band

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



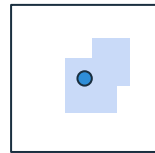
Satellite imagery

Feature generation from satellite imagery

- Each observation is a combination of date + lat/lon
 - Specify bounding box around point
 - 200m to 2,500m range
 - Specify time window of imagery prior to sample date
 - 15-60 day window range
 - Select least cloudy image
 - e.g. cloud_cover property
 - Filter to water area
 - e.g. using scene classification band
 - **Calculate features** from imagery bands
 - Summary stats (mean, max, min)
 - Ratios (NDVI, etc.)

Sample points

	date	latitude	longitude
uid			
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



Satellite imagery

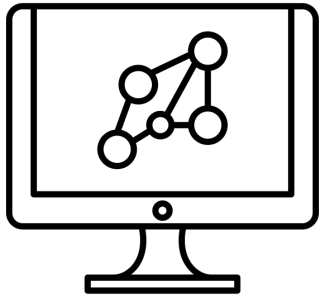
	B02_mean	B02_min	B02_max	B03_mean	B03_min	B03_max	B04_mean
uid							
bmdk	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475
obdp	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475
fmjb	418.988123	175.0	2686.0	604.710812	267.0	2934.0	509.557734
xyht	161.532712	50.0	1182.0	312.350417	69.0	1382.0	186.135216
gstw	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475

Sample features

Putting models into
production

Phase 1

Machine learning competition

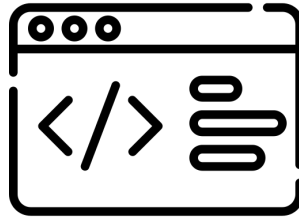


Ran a machine learning competition to develop models that detect and classify the severity of cyanobacteria blooms in small, inland water bodies.

Machine learning competitions are excellent for crowd-sourcing top approaches to complex predictive modeling problems.

Phase 2

Deployment-ready code package



Static research code needs to be transformed into a code package that is capable of generating predictions on new input data.

A production-ready code pipeline is efficient, generalizable, well-documented, and tested.

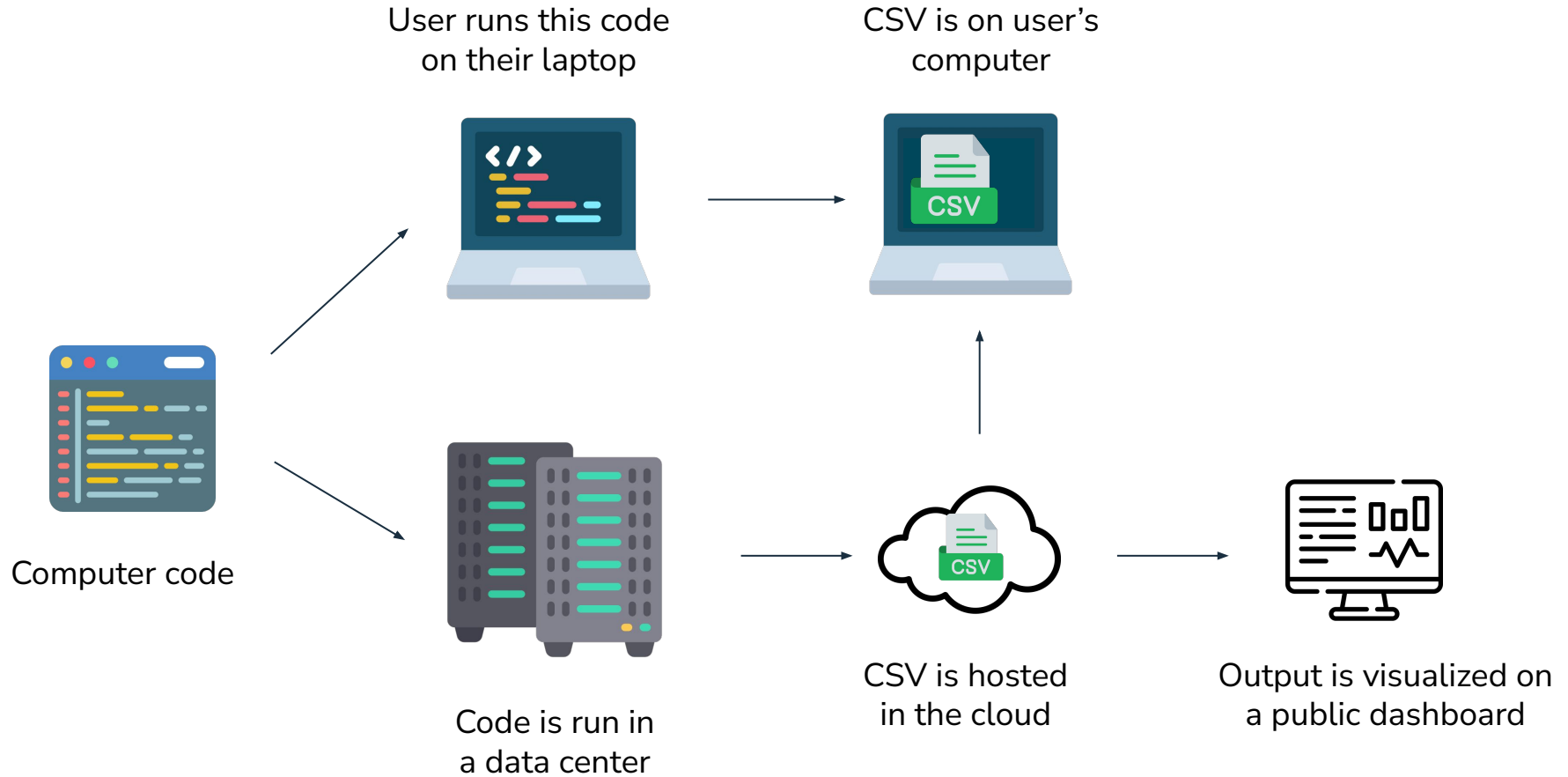
Handoff

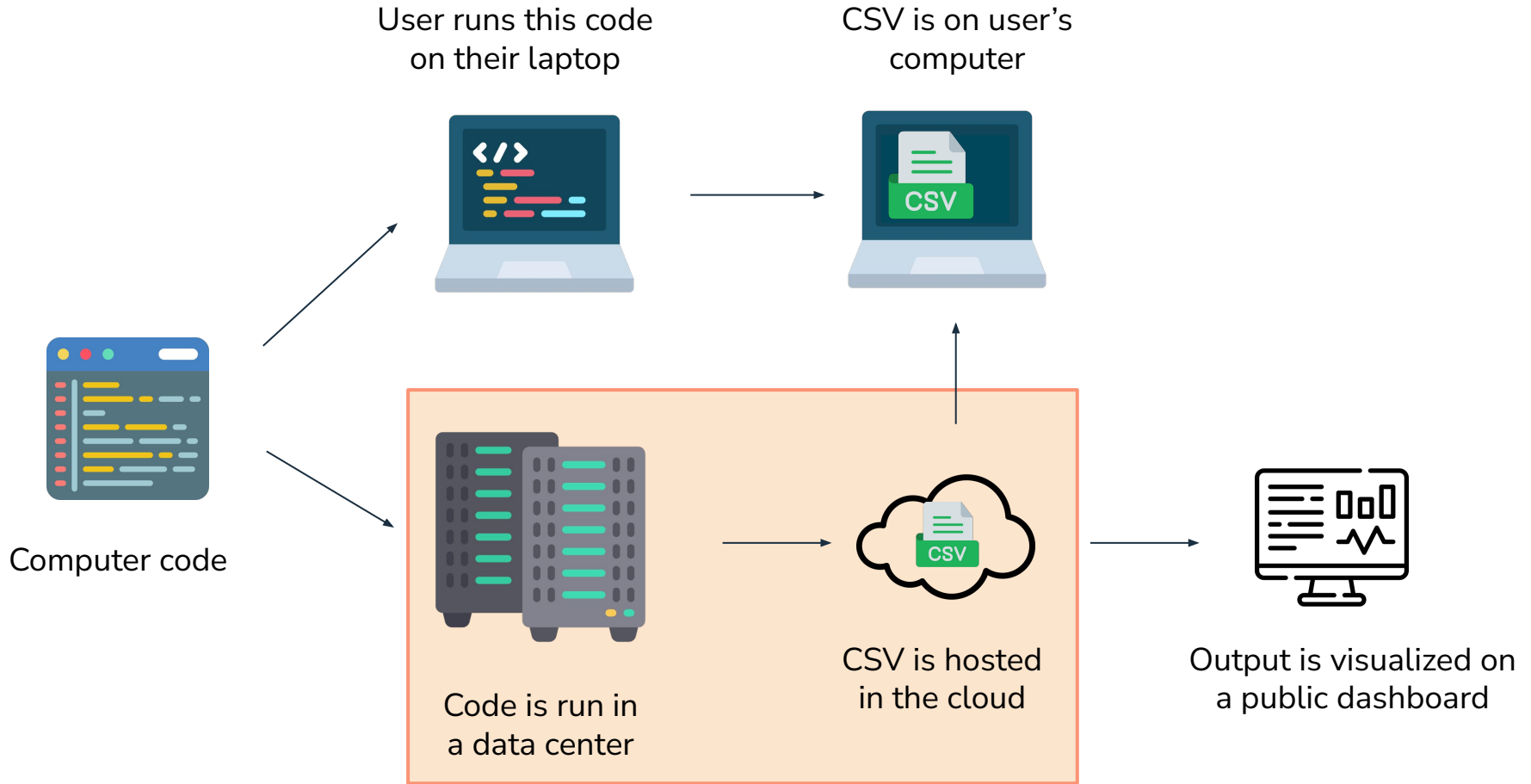
Regular cyanobacteria predictions



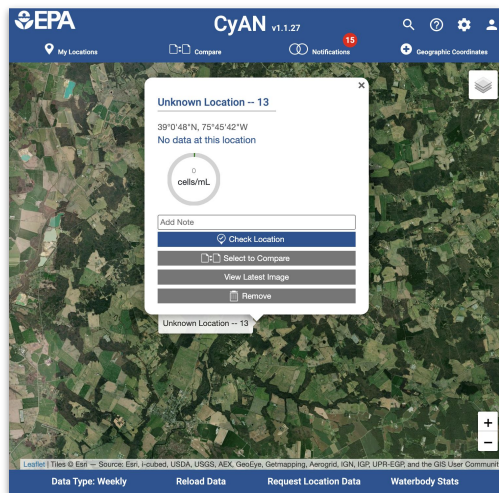
The desired outcome is accessible, regular predictions of cyanobacteria levels for given latitude and longitude points in inland water bodies across the US.

This enables ongoing detection of unsafe bacteria counts to inform advisories and protect public health and safety.





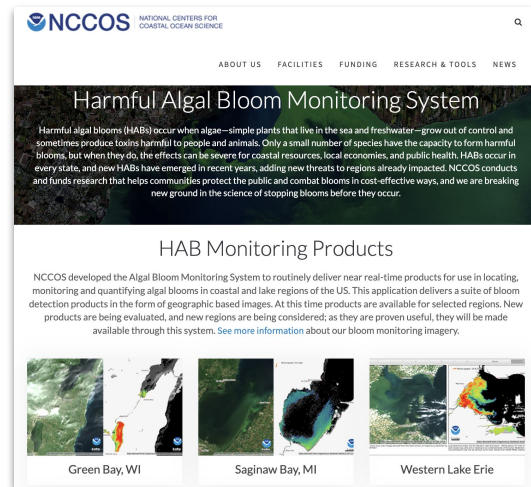
CyAN products we've explored



EPA's CyAN dashboard

The screenshot shows the 'CyAN File Search' interface on the Ocean Color website. It features a search form with fields for 'Date' (Start: 'yyyy-mm', End: 'yyyy-mm'), 'Period' (dropdown: '14 Day [2002-2007]'), and 'Product' (dropdown: 'cyanobacteria index'). There are checkboxes for 'Additional Options': 'Download results as a text file', 'Add URL prefix to results text file', and 'Generate checksum text file'. To the right, there are checkboxes for 'CONUS ALL - Single GEOTIFF' and 'CONUS ALL - Individual Tiles', with a note: 'Click one of the checkboxes to select all tiles or click your area(s) of interest on the map.' Below this is a map of the United States with a grid overlay, and a 'Submit' button at the bottom.

Ocean Color website where geotiffs can be downloaded



NOAA's HAB monitoring products for specific locations

Comparison to CyAN

Methodology

Using a test set of ~5k observations at locations our model has not been trained on, we evaluated our model's predictions and compared those against CyAN.

Querying the **CyAN dashboard**, we were only able to match 208 points, using a very generous +/- 90 day window from the sample date.

Using the **CyAN file server**, we were able to get data from either the sample day or week so we report those numbers instead.

We converted CyAN estimates to severity levels for an apples-to-apples comparison.

Comparison to CyAN

Daily CyAN estimates

Number of data points matched: 351 of 4,860 (7%)

CyAN RMSE: 0.94

DrivenData RMSE: 0.79

16% reduction in RMSE

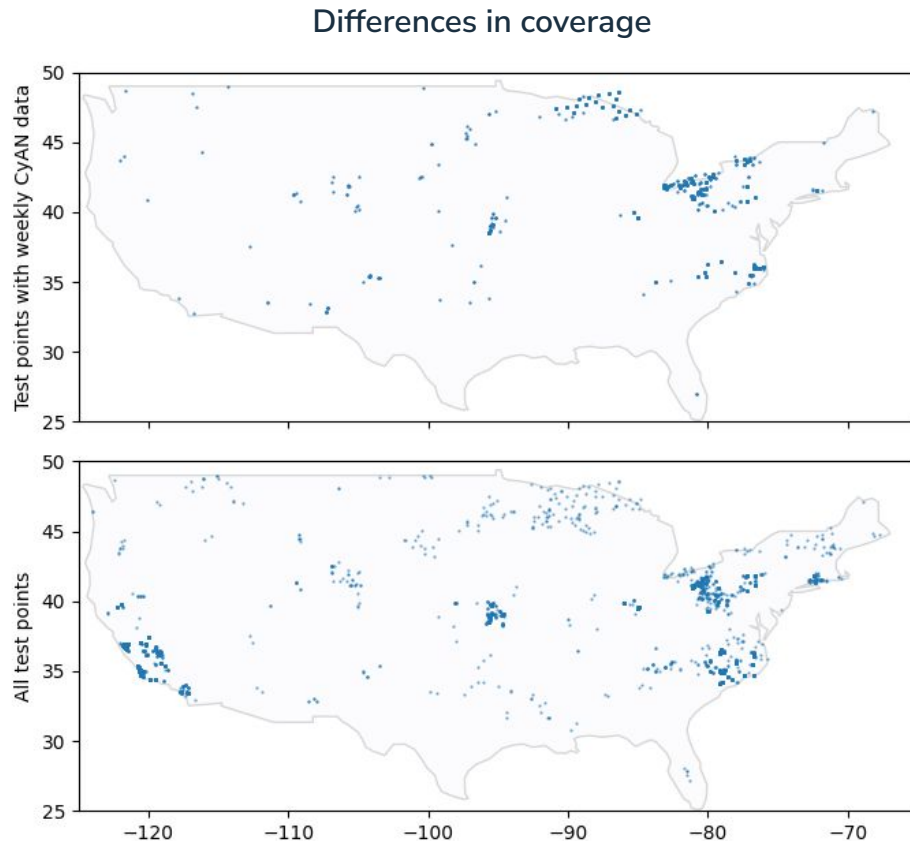
Weekly CyAN estimates

Number of data points matched: 893 (18%)

CyAN RMSE: 0.98

DrivenData RMSE: 0.81

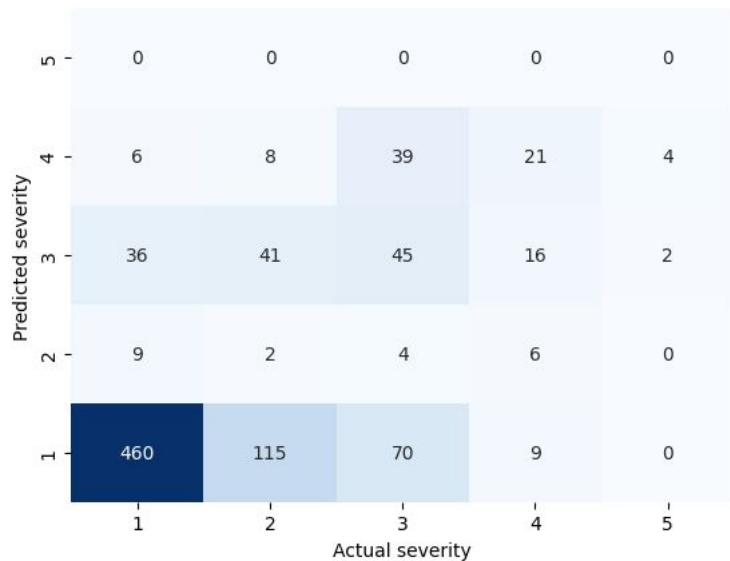
17% reduction in RMSE



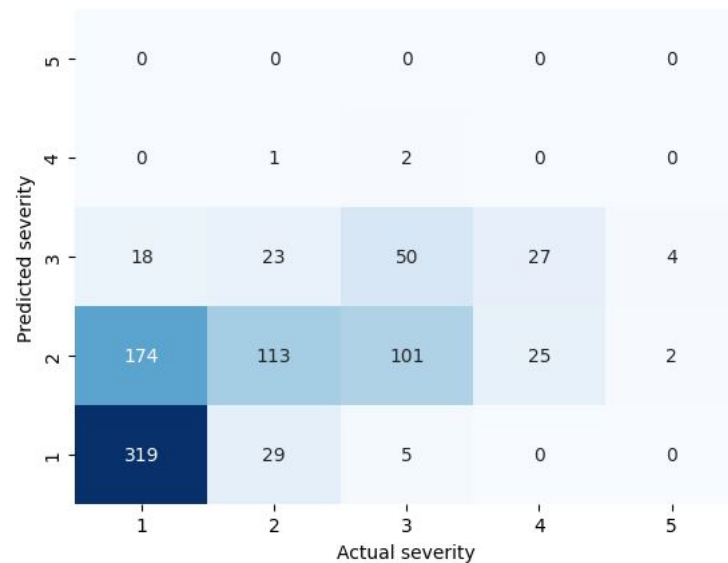
*Top map shows test set locations with a CyAN prediction.
Bottom map shows location of test set points.*

Comparison to CyAN

Weekly CyAN estimates vs. actual



DrivenData estimates vs. actual



Complementarities

CyAN works well for large blooms in oceans and coastal areas

- Uses Sentinel-3
 - Has bands more directly relevant to cyanobacteria
 - 300m resolution
 - ~2 day refresh rate
- Generates continuous map
- Outputs a geotiff file per date, supporting “side by side” view with imagery

Our model works well for localized blooms in smaller, inland water bodies

- Uses Sentinel-2
 - Ability to see features in small water bodies
 - 10m resolution
 - 5 day refresh rate
- Generates prediction for specific points
- Easy to query multiple date / location combinations
- Much greater coverage for water bodies

CyAN Overview

Integration considerations

Identifying if, how, and where our code package can supplement CyAN

- Where would Sentinel-2 based estimates provide the most value?
- For the various CyAN products, are there differences in:
 - Current user base
 - Target audience
 - Funding / resource allocation
 - Cloud infrastructure
- Is there a low-lift way to integrate regular runs of our code package into existing cloud workflows?
 - Are there specific runtime or output file format constraints?

Questions?

DRIVEN DATA