

# Cyanobacteria Detection from Satellite Data

Phase 2 Concluding Report

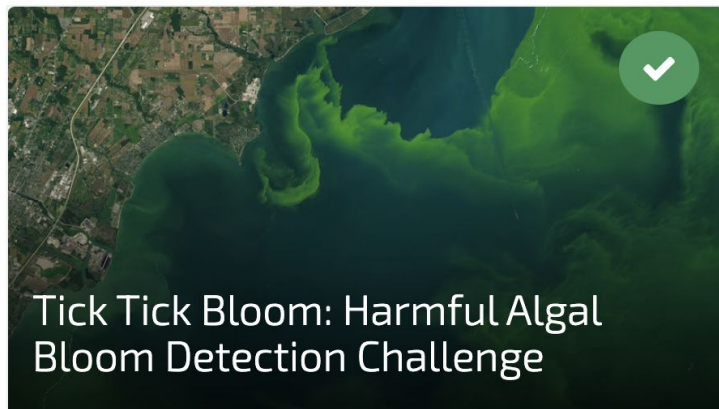
# Agenda

- Goals of Phase 2
- User interview synthesis
- Getting to a production-ready model
- CyFi: Cyanobacteria Finder
  - Overview
  - Live demo
  - Model performance
- Learnings + opportunities

# Goals of Phase 2

# Where we left off

## Phase 1



COMPETITION HAS ENDED

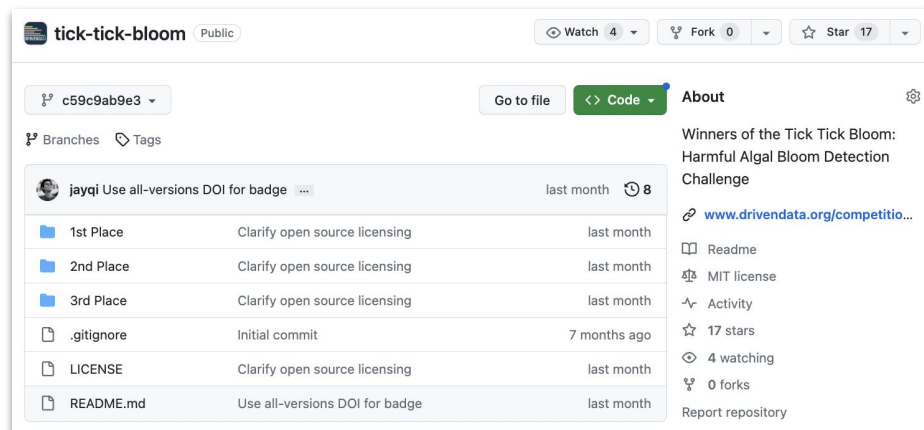
**\$30,000**

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

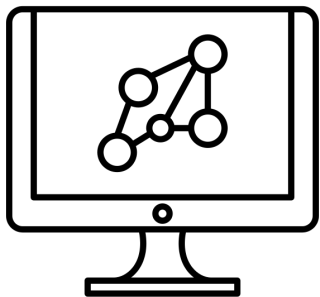
[RESULTS →](#)



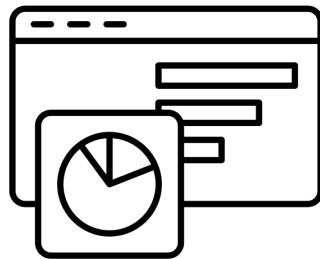
Successful machine learning competition to estimate cyanobacteria using satellite imagery, climate data, and elevation data. Winners' code provided in a public github repo.

# Carrying models forward for use

Competition github repo



Regular cyanobacteria  
predictions



There remains a “**missing middle**” of implementation that bridges the gap between static research code and the ability to use regularly generated predictions from the model

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

The outputs of machine learning competitions include winning model code, trained model, and write ups.

The desired outcome is regular predictions of cyanobacteria levels for given latitude and longitude points within small water bodies across the US.

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

## Why does this gap exist?

- Competitions rely on static data exported and processed once. Deployment requires **repeated, automatic use with new data**.
- Winning models are relatively unconstrained by the size and cost of their solutions. For ongoing use, **efficiency matters**.
- Competition code is validated once with anticipated, clean data. In the real world things break and change; use requires **basic robustness, testing and configurability**.
- There is substantial variability in the clarity and organization of competition-winning code. Usable code requires others to be able to **understand, maintain, and build on the codebase**.

# The “missing middle” for model use

..	Develop a competition model	Set up deployment-ready code package	Deploy and monitor
Task	<p>Explore data &amp; approaches to develop an accurate model</p> <p>Submit code assets and write up of winning approaches</p>	<p>Assess performance &amp; efficiency opportunities</p> <p>Simplify &amp; restructure code to transform it into a runnable pipeline</p> <p>Engineer clean, reproducible repository, access points for processing new data, tests, &amp; continuous integration</p>	<p>Set up infrastructure to run the model regularly</p> <p>Add monitoring &amp; error surfacing for effective maintenance</p> <p>Integrate predictions into dashboards &amp; user-facing analysis frameworks</p>
Output	Organized repo of “development code” (assorted notebooks and scripts) + weights file	Clean, configurable code package capable of generating predictions on new input data	Configured infrastructure, logging and monitoring, and optional code integrating with user interfaces

# Components of Phase 2

## Set up deployment-ready code package

**Dataset preparation:** Review and cleaning of dataset of ground observations of cyanobacteria severity (e.g., identifying incorrectly labeled data points) for public dissemination.

5%

**User interviews:** Understanding workflows of SMEs and end-users with specific attention to decision-making, actions, and data usage. Helps inform package configurations around optimal cadence of results, data formats, etc.

15%

**Algorithm improvements:** Assessment and improvement in accuracy, efficiency, and generalizability of submissions from Tick Tick Bloom Challenge

35%

**Code package:** Production of final data product(s) that generates estimates of cyanobacteria for specified longitude and latitude points in water bodies, tested with at least one SME/end-user.

45%



## High level goals of Phase 2

- Understand user needs and workflows
- Train a generalizable, production-ready model
- Reduce the barriers to using the model

# Dataset preparation

# Components of Phase 2

## Set up deployment-ready code package

**Dataset preparation:** Review and cleaning of dataset of ground observations of cyanobacteria severity (e.g., identifying incorrectly labeled data points) for public dissemination.

5%

**User interviews:** Understanding workflows of SMEs and end-users with specific attention to decision-making, actions, and data usage. Helps inform package configurations around optimal cadence of results, data formats, etc.

15%

**Algorithm improvements:** Assessment and improvement in accuracy, efficiency, and generalizability of submissions from Tick Tick Bloom Challenge

35%

**Code package:** Production of final data product(s) that generates estimates of cyanobacteria for specified longitude and latitude points in water bodies, tested with at least one SME/end-user.

45%

# Open dataset

Tick Tick Bloom featured a unique and valuable dataset of cyanobacteria in situ measurements from over 30 organizations across the U.S.

Making this dataset publicly available supports future research, testing, and benchmarking of models

## From competition data to public data

- Calculate a “distance to water” variable to identify noisy points\*
- Include all relevant metadata like data provider
- Provide documentation with variable descriptions and summary statistics

data_provider	region	latitude	longitude	date	density_cells_per_ml	severity	distance_to_water_m
Indiana State Department of Health	midwest	39.080319	-86.430867	2018-05-14	585.0	1	0.0
California Environmental Data Exchange Network	west	36.559700	-121.510000	2016-08-31	5867500.0	4	3512.0
N.C. Division of Water Resources N.C. Department of Environmental Quality	south	35.875083	-78.878434	2020-11-19	290.0	1	514.0
N.C. Division of Water Resources N.C. Department of Environmental Quality	south	35.487000	-79.062133	2016-08-24	1614.0	1	129.0
Bureau of Water Kansas Department of Health and Environment	midwest	38.049471	-99.827001	2019-07-23	111825.0	3	19.0

*Example rows from prepared dataset*

\* Using [ESA's World Cover](#) map which was more reliable at detecting water than scene classification band in Sentinel-2

## Description

This dataset provides cyanobacteria measurements for inland water bodies across the U.S. Measurements are based on "in situ" samples that were collected manually and then analyzed for cyanobacteria density. Each measurement is a unique combination of date and location (latitude and longitude). Samples were collected between 2013 and 2021.

The data was generated as part of DrivenData's [Tick Tick Bloom: Harmful Algal Bloom Detection Challenge](#) in 2023, which was created on behalf of NASA. The goal of the competition was to use satellite imagery to detect and classify the severity of cyanobacteria blooms in small, inland water bodies. Data were provided by [over 30 organizations](#) that manage public health and water quality, and then cleaned and aggregated.

## Fields

- `uid` (str): unique ID for each row
- `data_provider` (str): Name of the organization that provided the measurement
- `region` (str): Region of the U.S. Possible values are `midwest`, `northeast`, `south`, and `west`
- `latitude` (float): Latitude of the location where the sample was collected
- `longitude` (float): Longitude of the location where the sample was collected
- `date` (pd.datetime): Date when the sample was collected, in the format YYYY-MM-DD
- `density_cells_per_ml` (float): Raw measurement of total cyanobacteria density in cells per mL
- `severity` (int): Severity level based on the cyanobacteria density. The density ranges for each severity level are:

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

- `distance_to_water_m` (float) : Euclidean distance to the nearest water body in meters. This column can be used to identify sample locations that may be noisy or incorrect. For example, a distance of 0 indicates that the point is in a water body. Distances were calculated using [ESA's World Cover map](#) and [Google Earth Engine](#).

## Statistics

Breakdown by data provider

Data provider	Samples
N.C. Division of Water Resources N.C. Department of Environmental Quality	10,902
California Environmental Data Exchange Network	5,645
Bureau of Water Kansas Department of Health and Environment	1,465
US Army Corps of Engineers	1,145
EPA Central Data Exchange	1,086
EPA National Aquatic Research Survey	894
Pennsylvania Department of Environmental Protection	836
Indiana State Department of Health	649
Connecticut State Department of Public Health	325
Delaware National Resources and the University of Delaware's Citizen Monitoring Program	221
New Mexico Environment Department	152
EPA Water Quality Data Portal	139
Wyoming Department of Environmental Quality	96
Texas Commission on Environmental Quality	15

Breakdown by severity level

Severity level	Samples
1	9,761
2	4,083
3	3,812
4	5,824
5	90

# User interviews

# Components of Phase 2

## Set up deployment-ready code package

**Dataset preparation:** Review and cleaning of dataset of ground observations of cyanobacteria severity (e.g., identifying incorrectly labeled data points) for public dissemination.

5%

**User interviews:** Understanding workflows of SMEs and end-users with specific attention to decision-making, actions, and data usage. Helps inform package configurations around optimal cadence of results, data formats, etc.

15%

**Algorithm improvements:** Assessment and improvement in accuracy, efficiency, and generalizability of submissions from Tick Tick Bloom Challenge

35%

**Code package:** Production of final data product(s) that generates estimates of cyanobacteria for specified longitude and latitude points in water bodies, tested with at least one SME/end-user.

45%

# User interview setup

## Choosing users to talk to

- Worked with NASA to select states based on a range of:
  - Geographic locations
  - HAB severity
  - Reliance on federal aid (resource availability)
  - Population and number of inland water bodies in their region (impact)
- Invited the following states:
  - California
  - Michigan
  - Georgia
  - Louisiana
  - New York
  - Florida
  - Texas (no response)

## HCD strategy

Interviews were structured with guiding questions and a consistent set of topics to cover, but with ample space for free-flowing discussion. User interview guide included in appendix.

Transcripts recorded for detailed analysis

Most helpful HCD tool was “ride alongs” (or screensharing) to see the tools people were using.

Some of the key HCD resources we used to guide our approach to user interviews:

- [IDEO Field Design Kit](#)
- [Luma Innovating for People](#)



# User interview setup

## Areas of focus in user interviews

- What decisions are you making
- What tools are you using
- What data supports those decisions
- How is the data managed

## Learnings will inform

- Which available dashboard is the best fit for surfacing cyanobacteria predictions
- In what format should we make predictions available and how should the results be surfaced
- What should we focus on when improving / preparing the model
- Are there other computational constraints we need to work around

# User interview summary

## California

Has clear need + desire for a view of cyanobacteria in small lakes. Currently using satellite data as a screening tool to identify lakes of concern. Primary uses are regulatory and impairment. Has a custom dashboard to view satellite data.

## New York

Has multiple use cases: lake quality inventory, informing sampling decisions, confirming publicly reported blooms, and identifying impaired water bodies. Viewing satellite imagery to confirm public reports.

## Georgia

Interested in satellite-based tool but has many competing priorities (HABs is not top) and to date has seen few examples of high toxin samples.

## Louisiana

HABs work is limited to a pilot study evaluating the accuracy of satellite-based tools in their state. No systematic monitoring.

## Florida

Robust, systematic monitoring program. Main areas of concern are very large lakes.

## Michigan

Current workflows are relatively manual and rely on publicly reported blooms. Limited openness to satellite-based tools. 20

# User interview takeaways for CyFi

## **Continuous coverage of a lake is nice but not necessary**

- States tend to have designated sampling locations or locations of reported blooms that can be inputted to CyFi as sample points

## **Thresholds are not universal and actions vary by state**

- Density values are more helpful than severity buckets

## **States have their own tools for managing their data**

- Rather than being a standalone tool, CyFi estimates should be easy to feed into state-level dashboards; output csv contains columns for latitude, longitude, and date for joining with other data

## **Maximum cyanobacteria estimation cadence is daily**

- With single source of satellite imagery (Sentinel-2) and static land cover map, CyFi can generate estimates for thousands of points in a day on a normal laptop

## **Visualizing on imagery is a nice to have**

- Created CyFi explorer to help users review cyanobacteria estimates

## **Invite California for a pilot**

- California has a high levels of technical readiness, need, and interest

# Understanding use cases

## New York use cases

**Select a representative sample for lake inventory:** The New York Department of Environmental Conservation (DEC) is responsible for ~8,000 lakes in the state, only a small number of which are sampled every year (staff is 3-4 people). Their [Lake Classification and Inventory project](#) would benefit from a birds eye view of lakes across the state to determine how to select a representative sample for ground sampling (to calculate trophic state). Historically, lakes that were easy to access were the ones that were sampled, leading to bias in the results.

**Confirming public blooms:** New York DEC has been confirming public blooms based on visual assessment of satellite imagery. A CyFi density estimate provides an additional datapoint, which can increase the confidence of the assessment.

**Identifying places where sampling is not needed:** NY DEC says identifying water that is not impaired is just as helpful as identifying places where it is impaired, as it allows them to cross it off the list of where to sample.

# Understanding use cases

## General use cases

Many states have a version one of the following use cases. We see these as the main places where CyFi can be useful.

- Identify areas of concern
- Identify areas of no concern
- Provide a birds-eye view of conditions in lakes across the state

Ultimately, CyFi can help:

- Inform interventions to support public health
- Help prioritize limited ground sampling staff and resources
- Provide a more comprehensive view of water bodies across the state for regulatory and impairment work

# Algorithm improvements

# Components of Phase 2

## Set up deployment-ready code package

**Dataset preparation:** Review and cleaning of dataset of ground observations of cyanobacteria severity (e.g., identifying incorrectly labeled data points) for public dissemination.

5%

**User interviews:** Understanding workflows of SMEs and end-users with specific attention to decision-making, actions, and data usage. Helps inform package configurations around optimal cadence of results, data formats, etc.

15%

**Algorithm improvements:** Assessment and improvement in accuracy, efficiency, and generalizability of submissions from Tick Tick Bloom Challenge

35%

**Code package:** Production of final data product(s) that generates estimates of cyanobacteria for specified longitude and latitude points in water bodies, tested with at least one SME/end-user.

45%

## Data sources used by winning models

	<b>Landsat Satellite</b>	<b>Sentinel 2 Satellite</b>	<b>HRRR Climate data</b>	<b>Copernicus DEM Elevation</b>	<b>Metadata Time, location</b>
1st Place		✓ Color value statistics	✓ Temperature		✓ Region Location
2nd Place		✓ Color value statistics		✓	✓ Clustered location
3rd Place	✓ Color value statistics	✓ Color value statistics	✓ Temperature Humidity		✓ Longitude

*All winners used statistics of the color values to generate features from satellite imagery, and supplemented with other environmental features.*



# Experiment summary table

Data		Satellite processing		Target variable		Model	
✓	Sentinel-2	✓	Bounding box size	✓	Severity category	✓	Hyperparameters
✓	Landsat	✓	Water filtering	✓	Density	✓	Number of models to ensemble (k-folds)
✓	Climate (HRRR)	✓	Cloud filtering	✓	Log density		
✓	Elevation (DEM)	✓	Date range (# of days prior to sample)				
✓	Land cover	✓	Number of images to use per observation				
✓	Metadata						

# Key decisions: data

## **Use only Sentinel-2 imagery**

The 3rd place winner was the only one to use Landsat imagery, suggesting that it was not critical for effective prediction. Using only one source of satellite imagery significantly speeds up generating predictions without a corresponding dip in accuracy.

## **Do not use climate or elevation features**

While climate and elevation do have an impact on how cyanobacteria blooms form, we find that adding simple climate and elevation features do not improve accuracy where satellite data is present. Winning models likely used these for data points prior to the launch of Sentinel-2.

## **Include land cover (300m resolution)**

Land cover helps capture farmland areas that are more likely to have runoff that contributes to blooms. We find that a static-map from 2020 is sufficient, which is much faster than querying a real-time satellite-derived product.

# Key decisions: satellite processing

## **Filter to water area and use a larger bounding box**

Land pixels will generate falsely high cyanobacteria estimates so we filter them out. We find that the scene classification band (while not perfect) is sufficient for masking non-water pixels. Since ground sampling points are often near land (taken from the shore or the dock), a larger bounding box (2,000m) is used to ensure the relevant water pixels are included.

## **Use a larger look-back window and filter to images with almost no clouds**

We use the scene classification band to calculate the percent of clouds in the bounding box and do not use any imagery that has greater than 5% clouds. Given the strict cloud threshold, we use a look-back window of 30 days before the sample. This increases the chances of getting a cloud-free image.

## **Use only one image per sample point**

Some winning solutions average predictions over multiple images within a specified range. We find that this favors static blooms. We use only the most-recent cloud-free image to better detect short-lived blooms.

# Key decisions: target variable

## **Estimate density instead of severity**

States use different thresholds for action, so predicting density instead of severity categories supports a broader range of use cases.

## **Train the model to predict log density**

We find transforming density into a log scale for model training and prediction yields better accuracy. This helps the model learn that incorrectly estimating a density of 100,000 when the true density is 0 is much more important than incorrectly estimating a density of 1,100,000 when the true density is 1,000,000. The estimate a user sees has been converted back into (non-log) density.

# Generalizability

## What it is and why it matters

Generalizability is the degree to which the model applies well to new data in new time ranges and new locations.

We care about generalizability because it's not possible (or reasonable) to train the model on a sample from every lake in the U.S. We expect the model to learn useful patterns that apply to new locations.

## Removing competition artifacts

**Longitude:** All winning solutions used a “longitude” feature in their solutions. This overfits to the data distribution of the data providers and won't generalize well.

Example:

- California only conducts toxin analysis for suspected blooms so in the competition dataset, most California lakes have high severity.
- With longitude as a feature, the model learns that data in this longitude always has high severity and so will predict high severity for lakes in California.
- This works well for the competition test set but doesn't generalize to non-competition data.

# 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



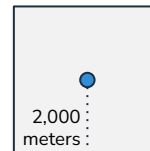
# Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify **bounding box** around point
  - 2,000 m

*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*

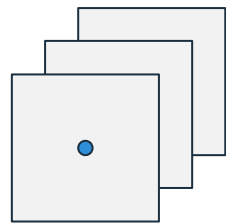
# Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify **time window** of imagery prior to sample date
  - 30 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*



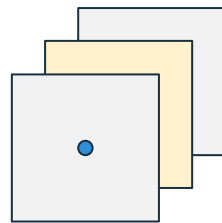
# Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select **most recent, least cloudy** image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made

*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*

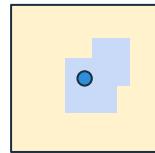
# Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select most recent, least cloudy image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made
- Filter to **water area**
  - 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*

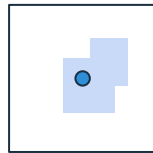
# Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select most recent, least cloudy image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made
- Filter to water area
  - Using scene classification band
- **Calculate features** from imagery bands in water area
  - Summary stats (mean, max, min)
  - Ratios (NDVI, etc.)

Sample points

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



Satellite imagery

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

Sample features

# Code package

CyFi: Cyanobacteria Finder

# Components of Phase 2

## Set up deployment-ready code package

**Dataset preparation:** Review and cleaning of dataset of ground observations of cyanobacteria severity (e.g., identifying incorrectly labeled data points) for public dissemination.

5%

**User interviews:** Understanding workflows of SMEs and end-users with specific attention to decision-making, actions, and data usage. Helps inform package configurations around optimal cadence of results, data formats, etc.

15%

**Algorithm improvements:** Assessment and improvement in accuracy, efficiency, and generalizability of submissions from Tick Tick Bloom Challenge

35%

**Code package:** Production of final data product(s) that generates estimates of cyanobacteria for specified longitude and latitude points in water bodies, tested with at least one SME/end-user.

45%

A satellite map of a large lake, likely Lake Erie, showing extensive green algal blooms. The blooms are concentrated in the western and southern parts of the lake. Two red dots mark specific sampling locations: one in the western basin and another in the southern basin. The surrounding land is a mix of urban areas, agricultural fields, and forested regions.

# CyFi

*Machine learning for harmful algal bloom detection*



Harmful algal blooms (HABs) are a **common threat to marine and human health.**

Existing automated detection tools focus on ocean and coastal areas. But blooms in smaller inland water bodies are still monitored manually, **which is very time intensive.**



**CyFi** (Cyanobacteria Finder) is an open-source Python package that uses satellite imagery and machine learning to detect cyanobacteria levels, one type of HAB.

CyFi can help decision makers protect the public by **flagging the highest-risk areas in lakes, reservoirs, and rivers quickly and easily.**



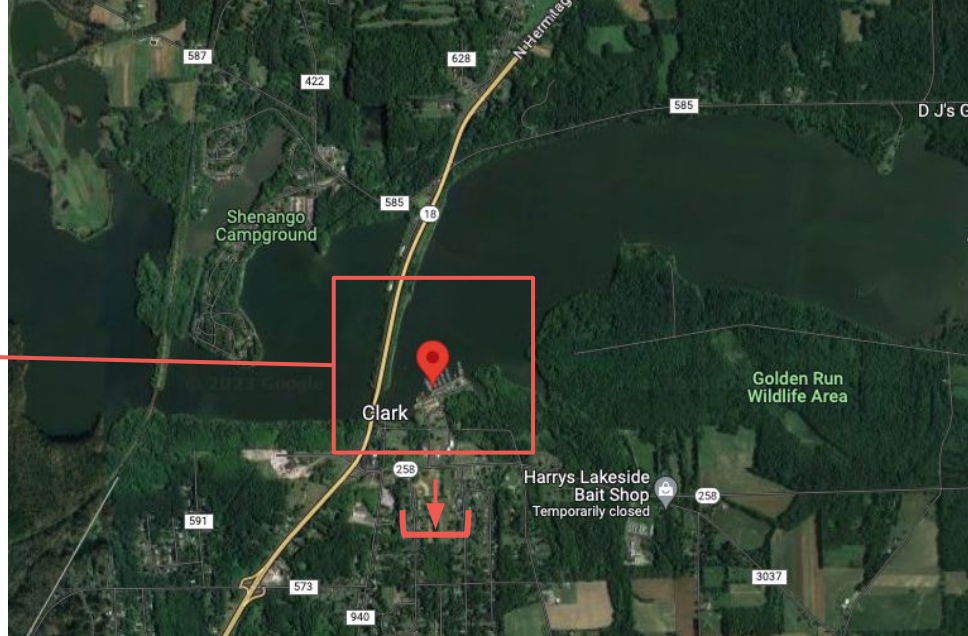




CyFi uses **high-resolution Sentinel-2 satellite imagery** (10-30m) to focus on smaller water bodies with rapidly changing blooms.

**Sentinel-3 is used by most existing tools**, but its resolution of 300-500m is often too coarse for small, inland water bodies.

	date	latitude	longitude
0	2023-06-27	41.287577	-80.424543
1	2023-07-10	35.650000	-78.682816
2	2023-08-31	35.705416	-79.164659
3	2023-09-22	37.564318	-101.335575



Generate estimates for many points at once with a simple CSV of **dates and locations**!

CyFi searches for and downloads publicly available **satellite imagery** around each point, which is passed into a machine learning model.

sample_id	Sample information			Predicted cyanobacteria density (cells/ml)	Severity level
	date	latitude	longitude	density_cells_per_ml	severity
89e12c14b5a131b82e9738932a7fa9c8	2023-06-27	41.287577	-80.424543	57,433	moderate
087d604d9d8568761513d26a47c94bc8	2023-07-10	35.650000	-78.682816	83,609	moderate
a0517780fa24874ebf166aefa17a0c1b	2023-08-31	35.705416	-79.164659	5,733	low
cde656c081bfe8fa99c7c8b20ff547f7	2023-09-22	37.564318	-101.335575	3,684,003	high

Cyanobacteria estimates are saved out as a CSV that can be plugged into any existing decision-making process.

For each point, the model provides a severity level based on World Health Organization (WHO) guidelines and an estimated density in cells per mL for detailed analysis.

Severity	Density range (cells/mL)
Low	0 - 20,000
Moderate	20,000 - 100,000
High	Over 100,000

Simply run one  
line of code to  
generate  
predictions

```
$ cyfi predict list_of_points.csv
```

```
SUCCESS | Loaded 5 sample points (unique combinations of date, latitude, and longitude) for prediction  
SUCCESS | Downloaded satellite imagery  
SUCCESS | Cyanobacteria estimates for 4 sample points saved to preds.csv
```

Or estimate  
cyanobacteria for  
a single point  
rather than  
providing a file

```
$ cyfi predict-point --lat 35.6 --lon -78.7 --date 2023-09-25
```

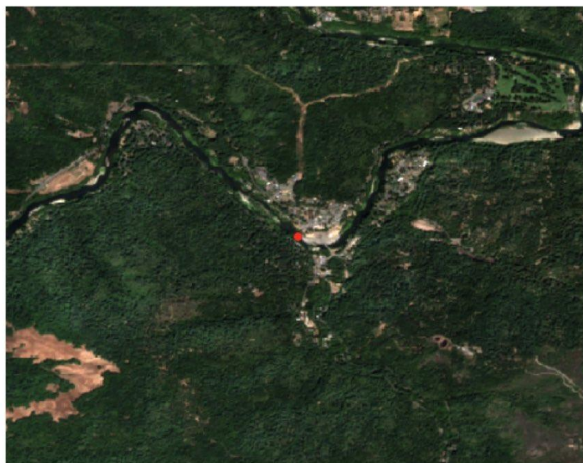
```
SUCCESS | Estimate generated:  
date                2023-09-25  
latitude            35.6  
longitude           -78.7  
density_cells_per_ml 22,836  
severity            moderate
```

Launch the CyFi Explorer to view cyanobacteria estimates alongside Sentinel-2 imagery!

#### CyFi estimates

sample_id ▲	date ▲	latitude ▲	longitude ▲	density_cells_per_ml ▲	severity ▲
6be1f8ed407e0ec7ab0c9a42394d9d44	2023-08-24	38.32629	-119.21121	7957	low
c485b9c41484d4d0b82b8580a215a43c	2023-08-23	34.24757	-117.2664	9234	low
3935648294a71be0197814c37de2f9a8	2023-08-23	38.466885	-123.01219	16141	low
389fee8dbca6759f0588dc842396c6b6	2023-08-22	37.7726963	-119.08373	17313	low
8b31451563d5ebd26cfa2cc5eb7357cc	2023-08-22	37.822007	-119.11976	17112	low

#### Sentinel-2 Imagery



#### Details on the selected sample



Estimated cyanobacteria density (cells/ml)

16141

Estimated severity level

low

Location

(-123.01219, 38.466885)

Sampling date

2023-08-23

Satellite imagery date

2023-08-12

CyFi makes it simple for water quality managers to take advantage of state-of-the-art machine learning.

Plus, the algorithm is open source so anyone can reuse, update, or contribute.

The screenshot shows the GitHub repository page for **cyfi** by **drivendataorg**. The repository is public and has 10 issues, 2 pull requests, 1 project, and 1 wiki. The main branch is selected. The file list shows the following items:

File/Folder	Commit Message	Time Ago
<b>ejm714</b> fix broken link ...		7 hours ago (66 comments)
<b>.github/workflows</b>	Prepare for release (#111)	17 hours ago
<b>cyfi</b>	release v1.0.0	7 hours ago
<b>docs</b>	release v1.0.0	7 hours ago
<b>tests</b>	Expand bbox to 2000 meters, filter to water are...	last week
<b>.gitignore</b>	Add docs (#107)	4 days ago
<b>CHANGELOG.md</b>	release v1.0.0	7 hours ago
<b>LICENSE</b>	initial commit	3 months ago
<b>Makefile</b>	Prepare for release (#111)	17 hours ago
<b>README.md</b>	fix broken link	7 hours ago
<b>pyproject.toml</b>	release v1.0.0	7 hours ago
<b>requirements_docs.txt</b>	Add docs (#107)	4 days ago

The right sidebar contains the following information:

- About**: Estimate cyanobacteria density based on Sentinel-2 satellite imagery. Link: [cyfi.drivendata.org/](https://cyfi.drivendata.org/). Tags: [satellite-imagery](#), [sentinel-2](#), [cyanobacteria](#), [habs](#).
- Readme**: MIT license.
- Activity**: 1 star, 1 watching, 0 forks.
- Releases**: 1 release, **v1.0.0** (Latest) 7 hours ago.



User-friendly  
documentation at:

[cyfi.drivendata.org](https://cyfi.drivendata.org)

[cyfi](#) [Installation](#) [Quickstart](#) [Visualize](#) [Background](#) [Accuracy](#) [Changelog](#) [Search](#) [Edit on GitHub](#)

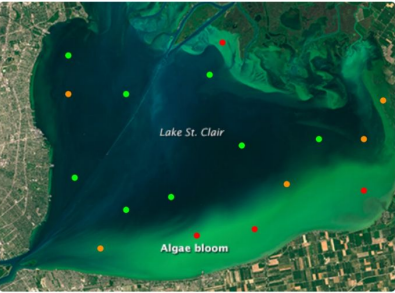
[CyFi: Cyanobacteria Finder](#)  
[Quickstart](#)  
[About the model](#)

## CyFi: Cyanobacteria Finder

CyFi is a command line tool that uses satellite imagery and machine learning to estimate cyanobacteria levels in small, inland water bodies. Cyanobacteria is a type of harmful algal bloom (HAB), which can produce toxins that are poisonous to humans and their pets, and can threaten marine ecosystems.

The goal of CyFi is to help water quality managers better allocate resources for in situ sampling, and make more informed decisions around public health warnings for critical resources like lakes and reservoirs.

Ultimately, more accurate and more timely detection of algal blooms helps keep both the human and marine life that rely on these water bodies safe and healthy.



*Stylized view of severity estimates for points on a lake with a cyanobacteria bloom.*  
Base image from [NASA Landsat Image Gallery](#)

## Quickstart

### Install

Install CyFi with pip:

```
pip install cyfi
```

For detailed instructions for those installing python for the first time, see the [Installation](#) docs.

### Generate batch predictions

Generate batch predictions at the command line with `cyfi predict`.

First, specify your sample points in a csv with the following columns:



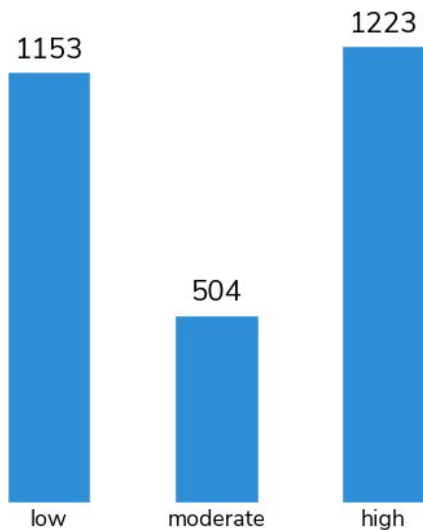
# Model performance

# Evaluation set

CyFi was evaluated  
using **2,880 ground  
measurements** from  
12 data providers  
spanning the time  
range August 2015  
to December 2021

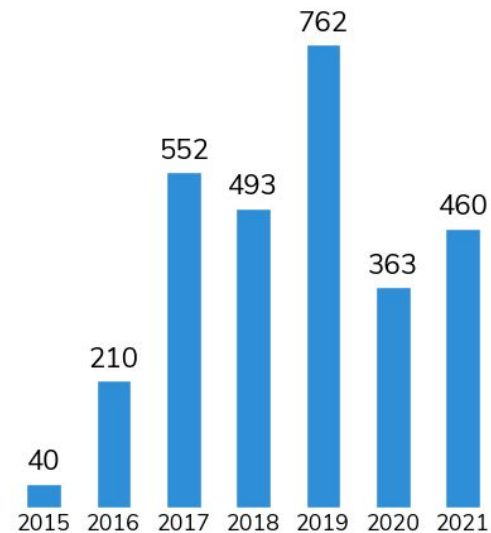


Observation counts by severity



Some states only conduct toxin analysis when blooms are suspected, which may account for the large number of high-severity observations in the evaluation set

Observation counts by year



Given that CyFi relies on Sentinel-2 imagery, the earliest date in the evaluation set aligns with the launch of Sentinel-2 (mid 2015)

# Metrics

Can detect 48% of non-bloom with 63% accuracy

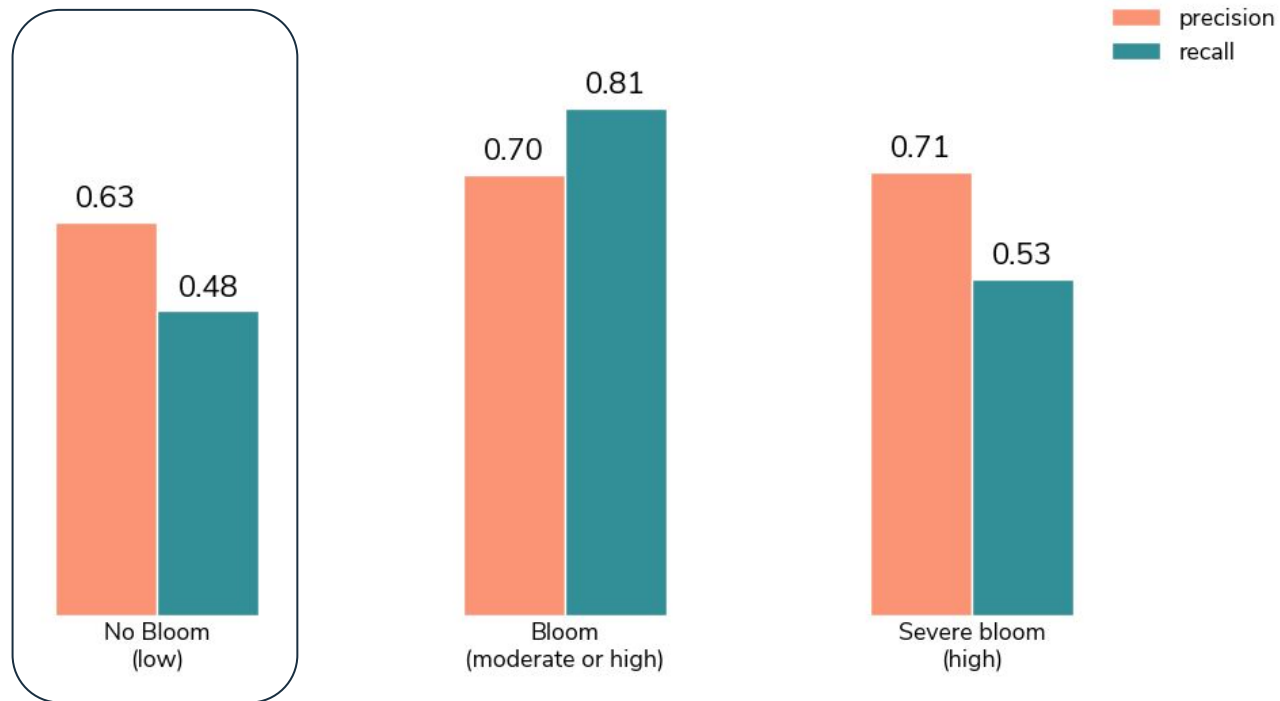
**Use case:** de-prioritize areas likely to not contain blooms to better allocate limited sampling resources

No bloom: < 20,000

Bloom: >= 20,000

Severe bloom: >= 100,000

## CyFi bloom detection



# Metrics

Can detect 81% of blooms with 70% accuracy

**Use case:** identify areas for ground sampling where there are likely to be public health impacts due to cyanobacteria

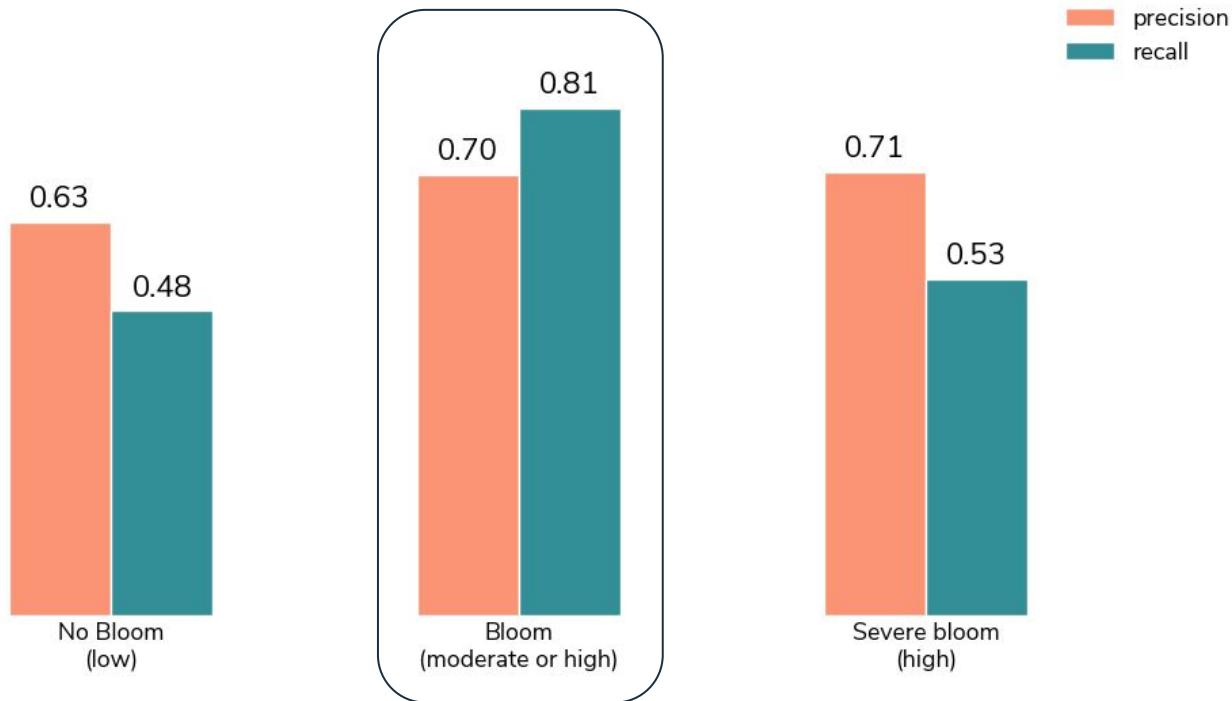
Can also help confirm public reports of blooms

No bloom: < 20,000

Bloom:  $\geq 20,000$

Severe bloom:  $\geq 100,000$

## CyFi bloom detection



# Metrics

Can detect 53% of severe blooms with 71% accuracy

**Use case:** identify areas for ground sampling where public health impacts are likely to be most severe

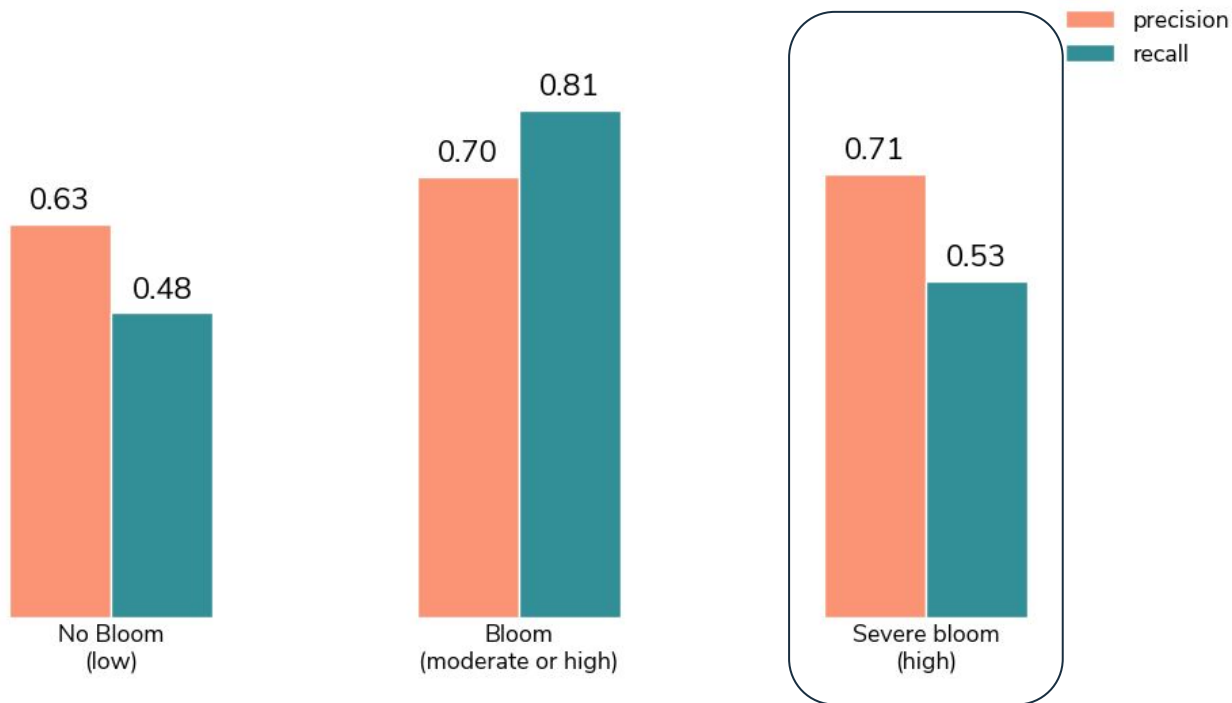
In most severe cases, may issue advisory without sampling

No bloom: < 20,000

Bloom:  $\geq 20,000$

Severe bloom:  $\geq 100,000$

## CyFi bloom detection



# CyAN comparison

## Generating the dataset

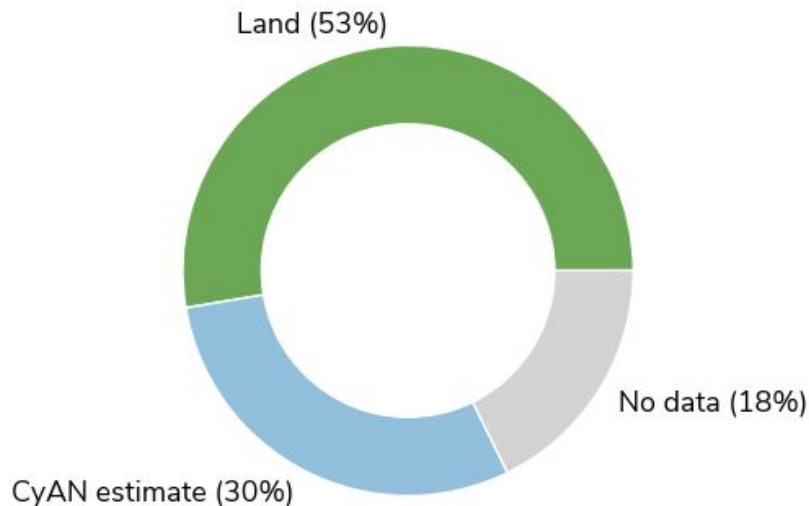
To provide an apples-to-apples comparison with CyAN, we looked up CyAN estimates for the evaluation set.

**Over half of the points in the evaluation set were identified as “land” by CyAN due to the coarse resolution of Sentinel-3.**

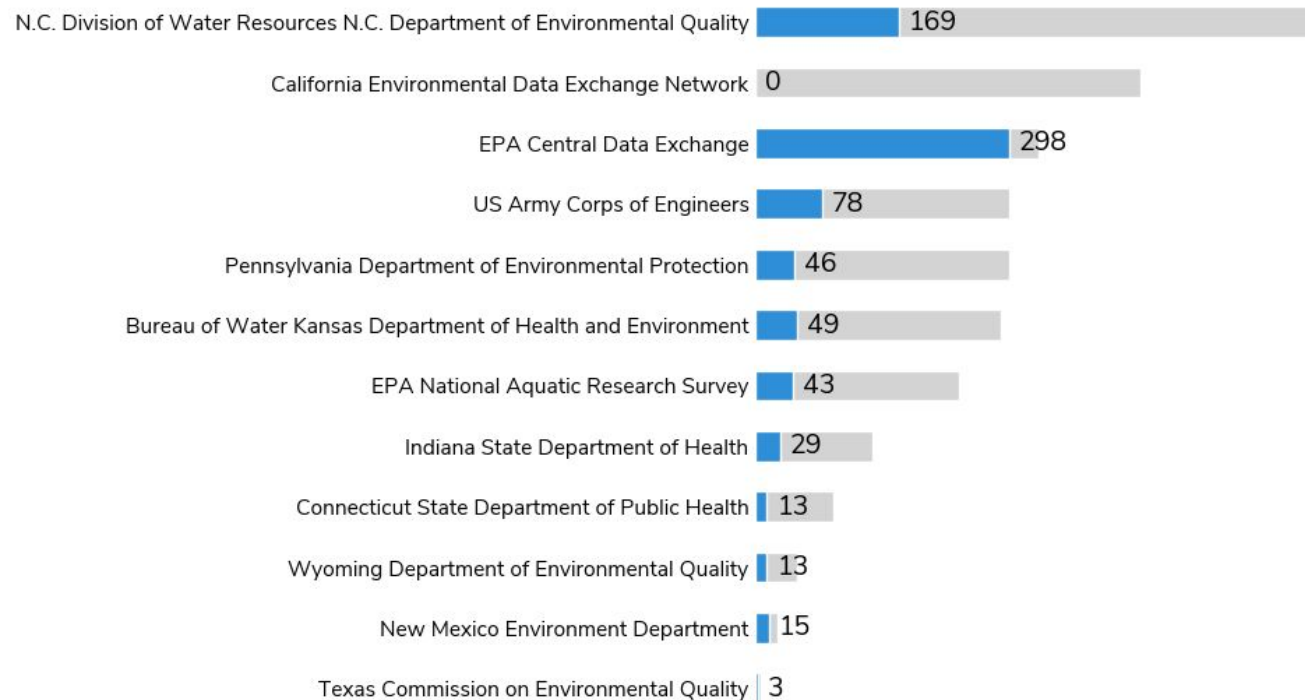
CyAN produces an estimate for 30% of points in the evaluation set. We'll call this the “CyAN subset.”

Notes: CyAN estimates were looked up using the [file search](#) tool on the [Ocean Color page](#). Weekly data estimates were used as this allowed us to get CyAN estimates for 30% of points in the evaluation set, compared to only 12% using daily data. Points in the evaluation set where distance to water is > 300m are excluded from the plot.

## CyAN predictions for evaluation set

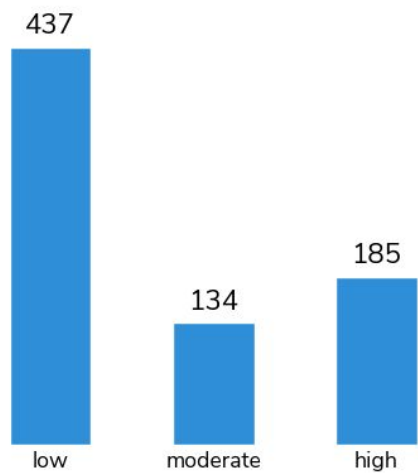


The CyAN subset has  
**756 observations,**  
**predominantly from the**  
**EPA Central Data**  
**Exchange.**



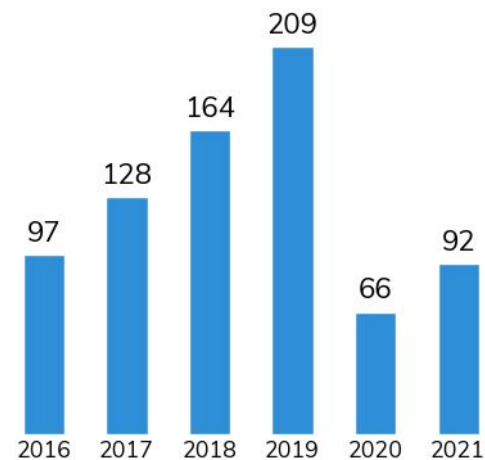


Observation counts by severity



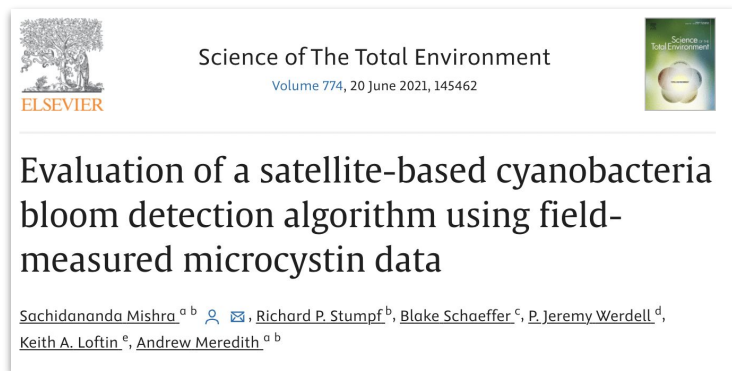
Low-severity observations account for over half (57%) of the observations in the CyAN subset.

Observation counts by year



The CyAN subset includes observations from all years between 2016 and 2021.

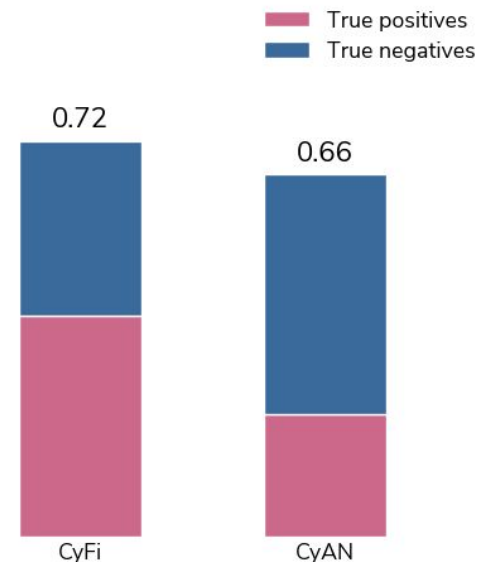
# Presence / absence



Using a cutoff of 10,000 per the evaluation paper, we find CyFi has a presence/absence accuracy of 72% compared to 66% for CyAN.

The improved accuracy is largely due to a higher correct classification of true positive cases (blooms).

## Presence / absence accuracy (>10,000 cells/ml)



**Key takeaway: CyFi meets a reasonable baseline for performance and has broad applicability due to the use of Sentinel-2 imagery.**

# California

## Out of sample

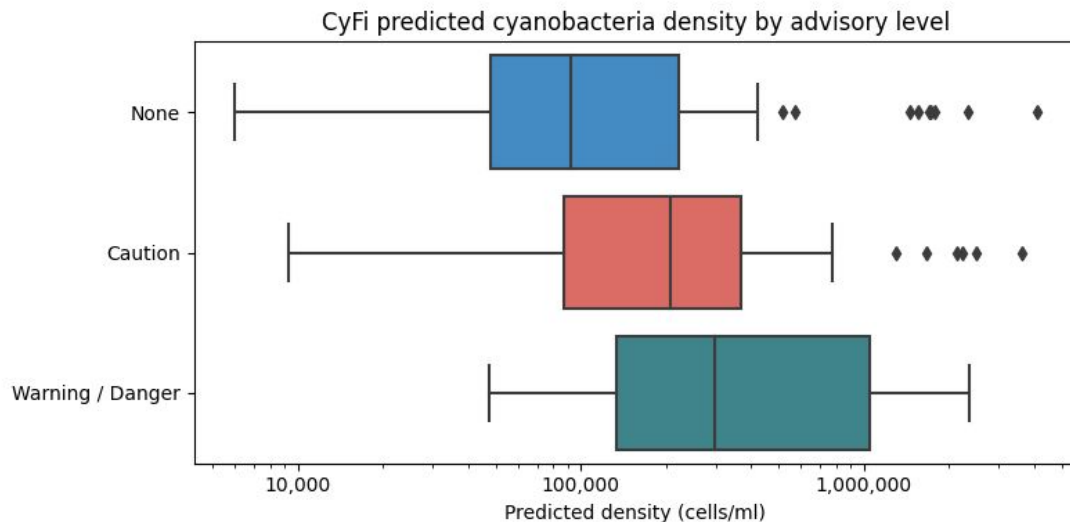
To test CyFi's ability to generate predictions on new data, we ran a test case with California

**CyFi successfully generated estimates for summer 2023 California sampling points in under an hour (231 sample points)**

Show promise for **identifying which areas can safely not be sampled and where most severe blooms are likely**

Estimated cyanobacteria density increases with advisory level

Predicted densities are quite high so custom thresholds should be used



Carrying CyFi forward

# Carrying CyFi forward

Here are some of the cases where CyFi is less reliable:



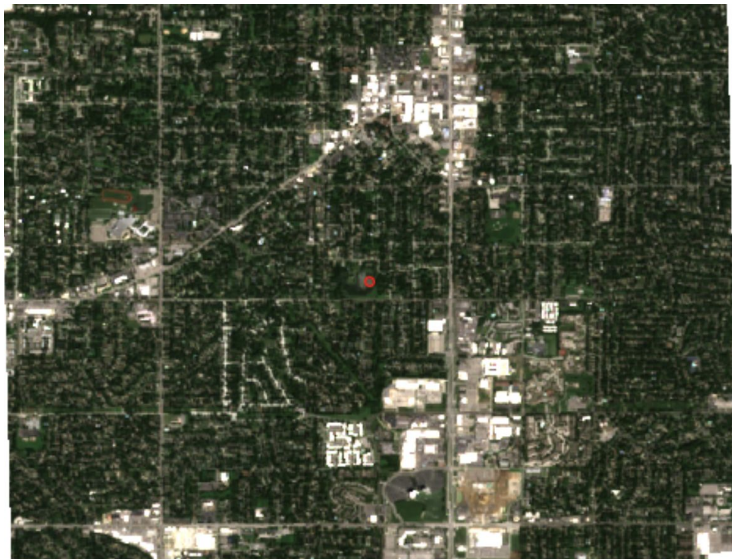
Other water bodies in the image



Very narrow waterways

# Carrying CyFi forward

Here are some of the cases where CyFi is less reliable:



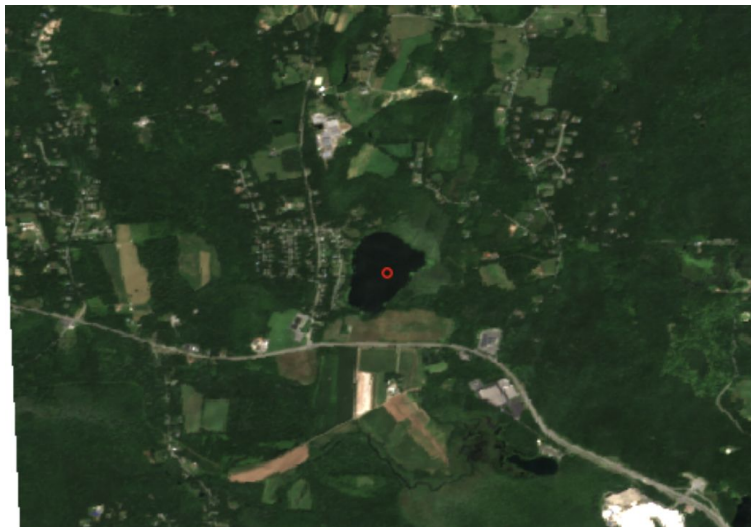
Very small water ways



Clouds

# Carrying CyFi forward

For best results, specify points in the water body rather than along the shore.



# Carrying CyFi forward

## Recommended modeling next steps

- Collect additional ground measurement data for true negative cases
- Water body segmentation to exclude non-contiguous water bodies in bounding box
- Cloud segmentation to remove cloud pixels
- Incorporate time-series climate features
- Control for naturally greener waterways
- Error analysis by water body size (determine minimum width for CyFi)
- Accept polygons as input and produce grids of sample points



# Carrying CyFi forward

## Recommended use

CyFi does best at identifying low and high densities.

### **Identify areas of no concern**

De-prioritize the lowest density predictions from ground sampling as these are likely not to be blooms.

### **Identify areas of concern**

If state advisories are based on presence/absence, consider issuing an advisory without ground sampling for severe blooms.

If state advisories are based on toxin ranges, prioritize ground sampling staff at suspected bloom locations.

Learnings + opportunities

# Reflections (for discussion)

## What went well

- **User interviews** helped align package design with decision-making workflows, and built foundations for sharing final products
- **Focusing on generalizability** outside of the competition dataset significantly changed and improved the winning models
- **User-friendly documentation and a simple interface** makes it easy to run cyanobacteria estimation models
- **Production code** is simpler, more efficient, and more robust than competition code
- Timing this work **immediately after the competition** allowed us to take advantage of existing work processes and fresh knowledge

## What were some challenges

- The wide **variety in decision-making processes** across states made it difficult to identify a reasonable nationwide approach to sharing predictions
- Our **labeled dataset was not fully representative** of the U.S., and additional data would have helped us build for generalizability
- There are many groups working on HAB estimation from satellite data. Connecting with relevant projects earlier may have fostered collaboration and **decreased duplication with existing efforts**

# Opportunities

## How can we continue growing CyFi?

- Work with an **individual state to pilot CyFi** and integrate with their specific workflows
- Gather **additional labeled cyanobacteria data** to continue improving the model and get more insight into performance
- Continue working to identify **existing platforms** into which CyFi could be integrated and share CyFi with groups undertaking related projects
- Collect user feedback on the CyFi package to inform **additional tutorials and user-interface improvements**

# Questions?

[cyfi.drivendata.org](http://cyfi.drivendata.org)

[emily@drivendata.org](mailto:emily@drivendata.org)

**DRIVEN**DATA