DS - GA 1017 RDS Draft Report

Fiona Chow and Jennah Gosciak

Spring 2023

Background

The purpose of this automated decisionmaking system (ADS) is to use satellite imagery to detect and classify the severity of cyanobacteria blooms in small, inland water bodies. This will help water quality managers better allocate resources for in situ sampling and make more informed decisions around public health warnings for drinking water and recreation. The primary goal of this ADS is to achieve accuracy, which is measured by the region-averaged root mean squared error (RMSE). The smaller the error value, the more accurate the model is. Region-averaged RMSE is calculated by taking the square root of the mean of squared differences between estimated and observed values for each region and averaging it by the number of regions.

The data and ADS come from a data science competition hosted by NASA and DrivenData. DrivenData is a company that hosts social impact data science competitions like those found on Kaggle. We chose to audit the solution of the second placed winner for the "Tick Tick Bloom: Harmful Algal Bloom Detection Challenge" that completed on Feb 17, 2023.

We selected this particular ADS because we were curious to see if there were sub-populations for which the ADS was less accurate in prediction. Algal blooms are an environmental justice issue. Exposure to high levels of cyanobacteria has been linked to cancer, birth defects, and even death (Gorham et al., 2020; Schaider et al., 2019). Prior research suggests that in the U.S. there are significant racial and socioeconomic disparities in access to clean drinking water (Schaider et al., 2019). Because this ADS could help water quality managers better allocate resources for in situ sampling and make informed decisions around public health warnings for drinking water and recreation, the algorithm must be accurate not only for the overall population but also for sensitive sub-populations.

Input and Output:

- **Description.** The competition provides 23,570 rows of training and test data. Since the test data do not have labels, we conduct our audit using the training data (n=17,060).Each row in the training data is a unique in situ sample collected for a given date and location going back to 2013. To illustrate this, we present the number of observations by year in figure 1. We sample from the training data to speed up the runtime of training and hyperparameter tuning. First, we restrict our data to instances recorded after 2016. From the filtered data, we sample so as to ensure proportional levels in each region to the original data while using a 72%/28% train/test split that is similar to what was done during the competition. We tune the hyperparameters on this smaller dataset and the results of our audit are based on these examples.
  - **Auxiliary Data.** We are using American Community Survey (ACS) 2015-2019 5-year estimates at the census tract level for our audit. We chose to use the 2015-2019 estimates, as there have been documented problems and delays associated with subsequent surveys due to COVID-19 (Wines and Cramer 10 March 2022). There are approximately 74,000 census tracts for all 50 U.S. states, Washington D.C., and Puerto Rico. However, only 1,660 census tracts are uniquely represented in the training data. This occurs because some in situ samples appear to come clustered in the same geographic areas. For example, we observed that 5,308 samples come from only a handful of census tracts in Chatham County, North Carolina. The features we have collected using census data include: race and ethnicity, median household income, and poverty rate.

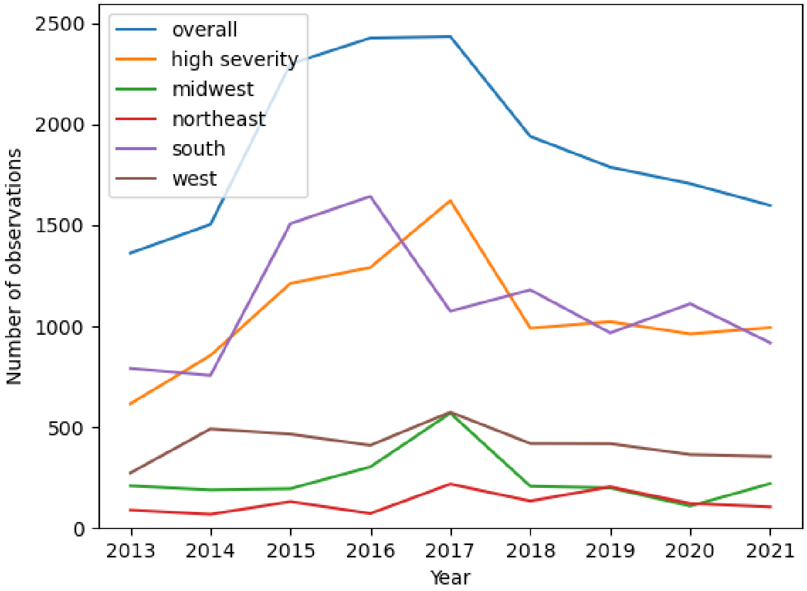

Figure 1: Number of training data observations by year

- **Input.** All the input datasets can be linked using a unique identifier, *uid*, which is a unique string. *uid* identifies the date and location (latitude and longitude) for each sample. We present detailed information on each of the input datasets used in the ADS. Note, as we explain below, we use a subsampled version of these datasets in our actual audit in order to reduce the runtime of training and hyperparameter tuning.

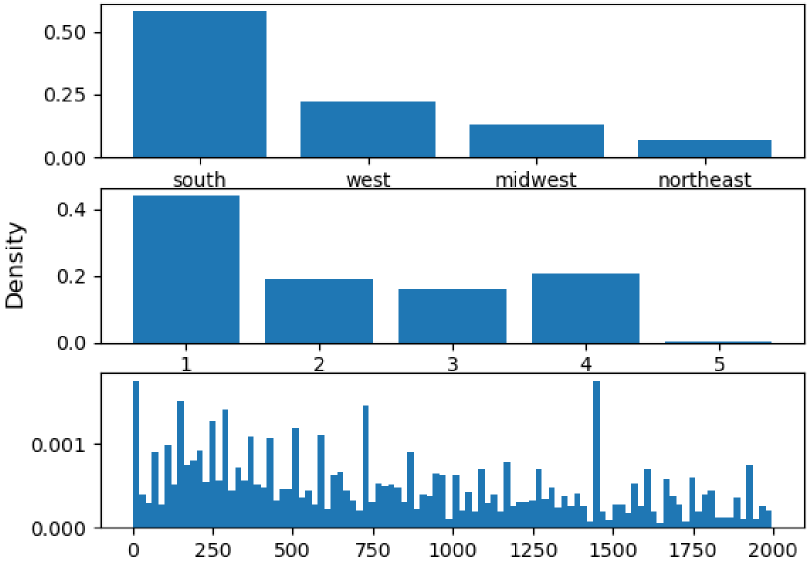

Figure 2: Value distributions of input features from training labels

  - Train labels: These data only cover the training records and provide information on the outcome for testing and validation purposes. The columns include:

  - region (categorical, taking values northeast, south, west, or midwest)
  - severity (categorical, a score from 1 to 5)
  - density (float, a value that corresponds to the severity score)

  We plot the value distributions in figure 2. We do not look at correlations as the relationship between severity and density is clearly defined.

- Metadata: The metadata provide information on the context of the sample. The columns include:
  - latitude (float)
  - longitude (float)
  - cluster (integer, construct which was generated from latitude/longitude to account for spatial variation)
  - date (string, which one can convert to a datetime object)
  - split (string, taking values "train" or "test")
- Elevation: Elevation data for the ADS come from the Copernicus Digital Elevation Model (DEM) with 30-meter resolution. The DEM is a digital surface model that can provide information on buildings, infrastructure, and vegetation. There are 23,570 entries of elevation data and no missing data in any of the columns. The columns, some of which are constructs that the author of the ADS generated, are:
  - latitude (float)
  - longitude (float)
  - box (integer)
  - elevation (float)
  - mine (float, indicates the minimum elevation)
  - maxe (float, indicates the maximum elevation)
  - dife (float, indicates the difference in elevation)
  - avge (float, indicates the average elevation)
  - stde (float, indicates the standard deviation in the elevation)
  - uid (string)
  - DateTime (string, indicates the date of the data download)

  Each input feature of the elevation data has the following value distribution:

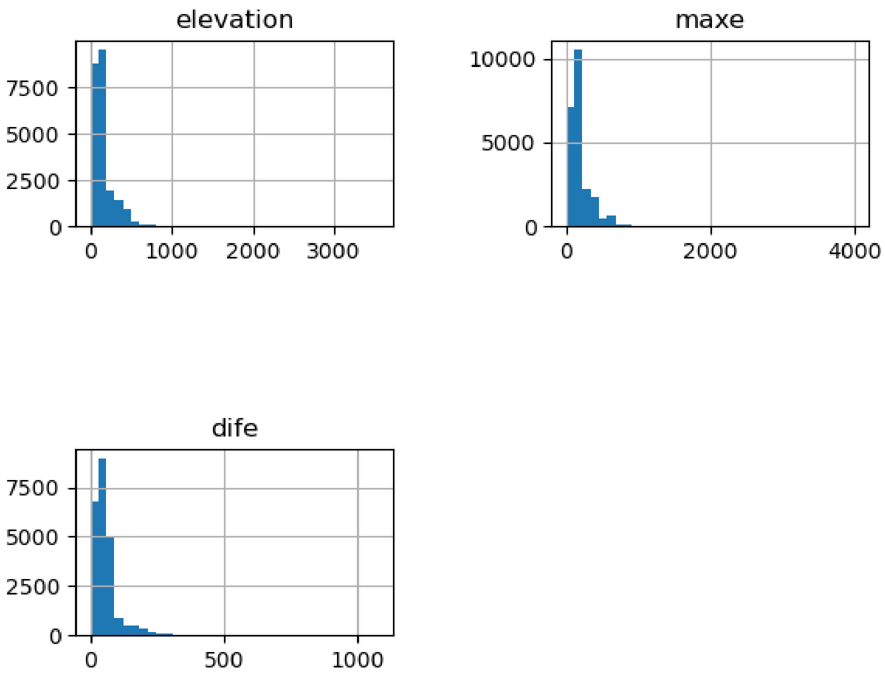

Figure 3: Value distributions of input features from elevation data

Both box and DateTime only have one unique value, 1,000 and December 22, 2022 respectively. In figure 3, elevation data range from 0 to 4000 meters. The median value for elevation, maximum elevation and difference in elevation is 124, 163 and 51 meters respectively.

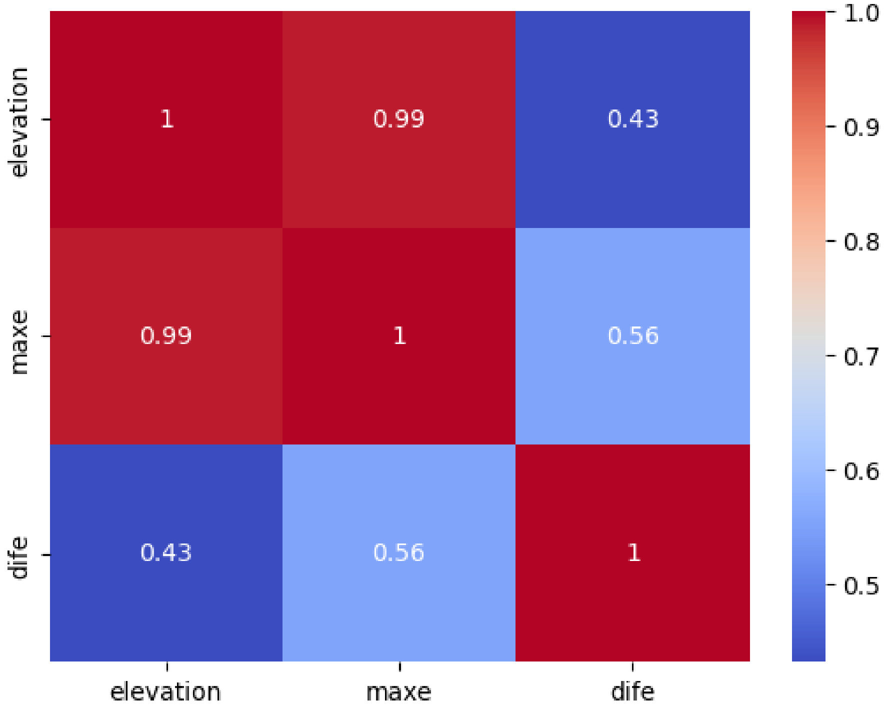

Figure 4: Correlation of input features from elevation data

In figure 4, the correlation matrix shows that there is a strong relationship between maximum elevation and elevation data (r = 0.99).

- Climate: Data come from the National Oceanic and Atmospheric Administration (NOAA) and provide information on temperature, wind, and precipitation. The author of the ADS acknowledged that he has not used climate data, and so we are not presenting any additional information on this dataset.
- Satellite: Sentinel-2 Level-2A satellite imagery come from the European Commission in partnership with the European Space Agency (ESA) and it provides information on the spectral bands – red, blue and green– at 1000 and 2500 meter radius from latitude and longitude and at ten-day intervals. The water areas were identified by k-means image segmentation and the satellite images were selected based on low cloud cover of less than five percent so as to increase the likelihood of capturing algal bloom detection in water surfaces. The columns, some of which are constructs that the author of the ADS generated, are:
  - imtype (object)
  - prop_lake_1000 (float, indicates estimate of water area at 1000 meters from latitude/longitude)
  - r_1000 (float, indicates estimate of red inside water area at 1000 meters)
  - g_1000 (float, indicates estimate of green inside water area at 1000 meters)
  - b_1000 (float, indicates estimate of blue inside water area at 1000 meters)
  - prop_lake_2500 (float, indicates estimate of water area at 2500 meters from latitude/longitude)
  - r_2500 (float, indicates estimate of red inside water area at 2500 meters)
  - g_2500 (float, indicates estimate of green inside water area at 2500 meters)
  - b_2500 (float, indicates estimate of blue inside water area at 2500 meters)

  Missing values were encoded with value '-1'. There may be missing data because satellite images could be limited due to cloud cover. In this case, there are 9,353 entries of satellite images and no missing data in any of the columns.

  Each input feature of the satellite data has the following value distribution:
  In figure 5, the red, green and blue (R/G/B) pixel values range from 0 (representing the minimum possible color intensity) to 255 (representing the maximum possible color intensity) and is the actual normalized range from which the Sentinel-2 images are created with. The proportion of image that is classified as lake values range from 0 (representing the entire image is not lake) to 1 (representing the entire image is lake).

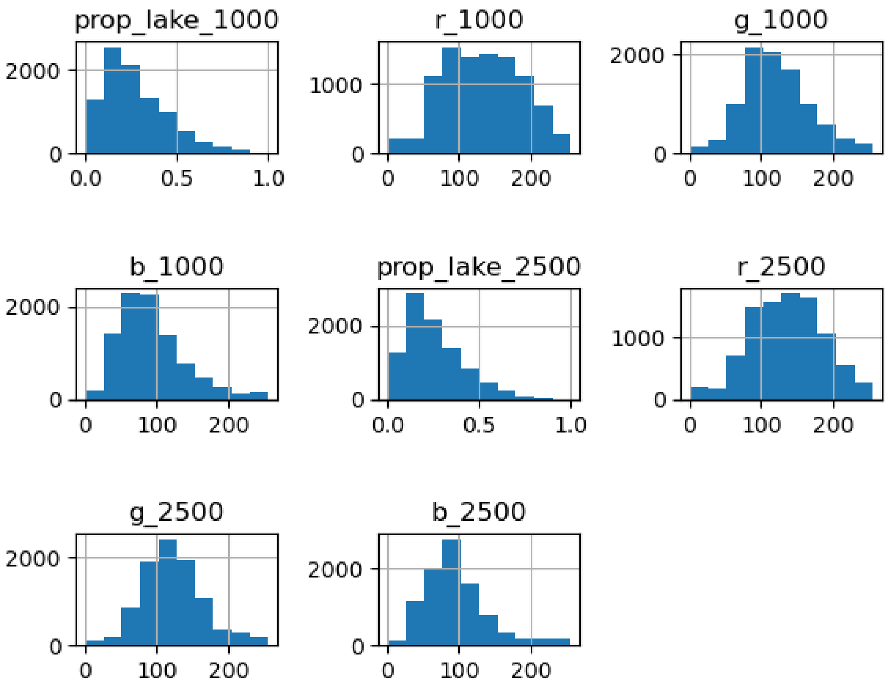

Figure 5: Value distributions of input features from satellite images

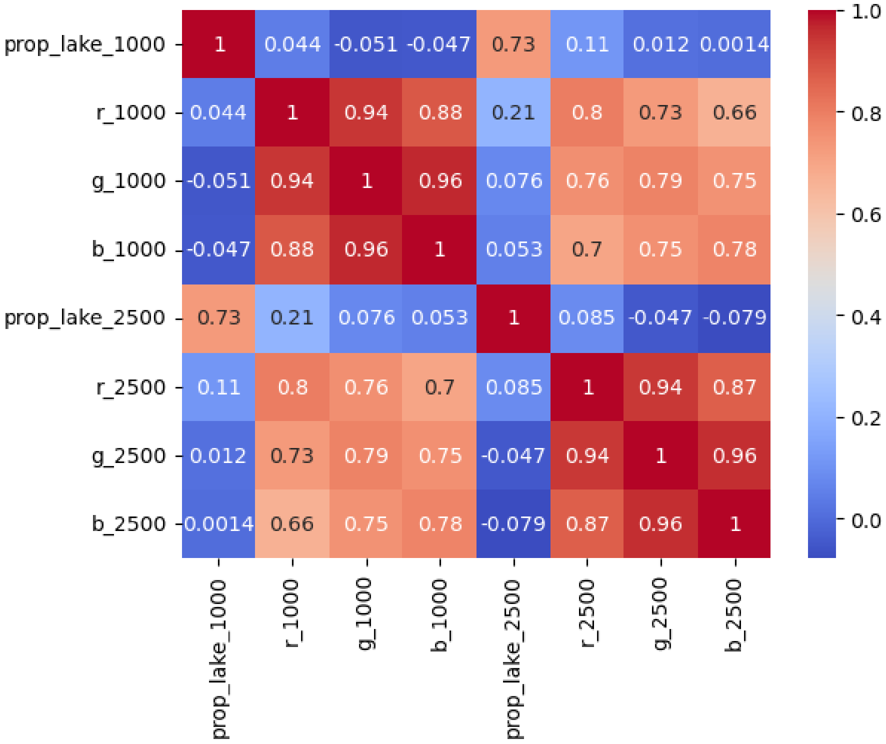

Figure 6: Correlation of input features from satellite images

In figure 6, the correlation matrix shows that there is a strong relationship between blue and green spectral bands ($r = 0.96$) and red and green spectral bands ($r = 0.94$) at their respective distance from latitude and longitude. In addition, there is a strong relationship between the spectral bands of the same color at different distances from the latitude and longitude ($r =$ 0.8).

- **Output.** The output of the system formally is a severity level that takes integer values 1 through 5. According to DrivenData, the severity is based on cyanobacteria density. Cyanobacteria density is another column in the training data, which ranges from 0 to 804,667,500 cells per mL. As table 1 and 7 illustrate, density values are non-overlapping for distinct severity levels and higher density values are associated with higher severity levels. For example, severity level five encompasses the the highest density values, greater than 10 million cells per mL. According to the World Health Organization (WHO), moderate and high risk health exposures occur at density levels $\geq 20,000$ cells per mL (WHO, 2003). Using this classification approach, we can construct a binary outcome label for a high risk health exposure. The binarized label aligns with severity levels 2-5, as table 2 indicates. Slightly less than half (≈ 44 percent) of the training data are low-risk. DrivenData specifies that the submission is based on severity levels and not the raw, underlying densities.

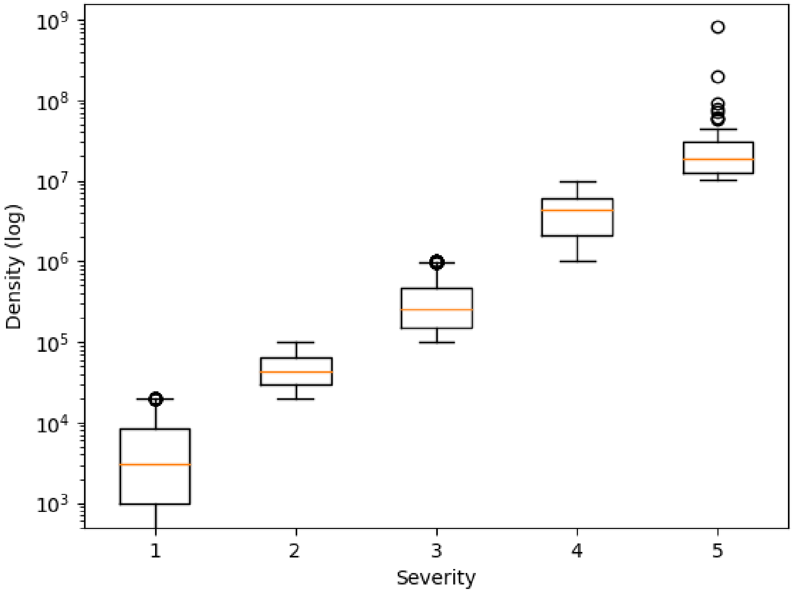

Figure 7: Range of density values for each severity level

Severity	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	$\geq 10,000,000$

Table 1: Formal severity level ranges

Severity (Binarized)	Severity	Percent
Low	1	44
High	4	21
High	2	19
High	3	16
High	5	< 1

Table 2: Binarized severity based on estimated risk level

Implementation and Validation

The data and code implementing the ADS can be found on the winner's GitHub page. We are using the second-place winner's solution, which is currently available with documentation as a public GitHub repository. DrivenData requires that all winning solutions are open source under The MIT License. The solution uses an ensemble of three different boosted tree models – XGBoost, CatBoost and LightBoost – and features such as region, date, location cluster, elevation and Sentinel-2 satellite images – red, blue and green spectral bands at 1000 and 2500 meters from latitude and longitude.

We plan to study the performance of the ADS across subpopulations. Since we are using a binary version of the outcome variable, we will primarily use metrics commonly studied in binary classification problems, which we outline below. As a comparison, to assess the validity of our findings, we may compare our results to the performance of the ADS across subpopulations using RMSE averaged over the four regions.

Subpopulations

Based on the ACS data we have, we propose the following construct definitions for sensitive groups across which we will analyze the performance of the ADS. We also include a construction definition for segregation, which we hope to explore if we have sufficient time.

- **Above average poverty rate:** We have ACS data at both the census tract and state level. We create indicator variables that denote whether the poverty rate in a given census tract is above the statewide average.
- **Above average shares of racial and ethnic subgroups:** We create a series of indicator variables, corresponding to racial and ethnic subgroups like Black, Asian, white, or Hispanic/Latinx, that denote whether a given census tract has an above average population share of each subgroup. We generate these indicator variables for each subgroup in comparison to the statewide average.
- **Low Income Community:** this is a designation from Internal Revenue Code §45D(e). Broadly, it refers to any census tract where the poverty rate is greater than or equal to 20 percent or the median family income is less than 80 percent of either the statewide median family income or the metropolitan area median family income–whichever is greater. We propose a modified version of this definition, given time constraints, that only compares the median family income to the statewide median.
- **Segregation:** the most commonly used measure of segregation is the dissimilarity index, which refers to the percentage of the population within a group that would have to move for each area to have the same percentage of that group as the larger metropolitan area or county. Time permitting, we plan to construct a measure of segregation by race and ethnicity using a binary version of the dissimilarity index.

Accuracy

Since we have turned this ADS into a binary classification problem, we plan to assess performance across metrics like recall, precision, and accuracy. We will evaluate performance overall, by region, and across the various subpopulations outlined above. We will prioritize and closely investigate recall as the outcomes for failing to detect algal blooms, and thus failing to issue a public health warning, could be consequential.

Fairness

According to the Aequitas Fairness Tree (Ghani, accessed April 17, 2023), the best fairness metric for this ADS use case is to have recall parity among the different subpopulations. Water quality managers can only allocate limited resources for in situ sampling because in situ sampling is labor intensive and expensive (Granger et al., 2018). Consequently, we should attempt to ensure the results of the ADS is distributed in a representative way. Since the false negative rate is equal to 1−recall, we will also look at false negative rate parity among the less privileged subpopulations because we want to see their chances of not receiving assistance given their group membership.

Additional

We will use SHAP to measure feature importance on the entire input data set of the ADS, which helps with model interpretability. And we will use LIME to understand the local behavior of the model, i.e., how it behaves around specific data points such as inconsistent feature contributions. With LIME, we can compare the explanations for similar instances to identify features whose contributions vary significantly across instances. Inconsistent feature contributions might indicate that the model is sensitive to small changes in the input data or that it's overfitting. We will look at features across severe and not severe predicted labels for different subpopulations.

References

R. Ghani. Aequitas. `http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/`, accessed April 17, 2023.

T. Gorham, E. D. Root, Y. Jia, C. Shum, and J. Lee. Relationship between cyanobacterial bloom impacted drinking water sources and hepatocellular carcinoma incidence rates. *Harmful algae*, 95:101801, 2020.

S. J. Granger, J. A. Qunicke, P. Harris, A. L. Collins, and M. S. Blackwell. Comparison of high frequency, in-situ water quality analysers and sensors with conventional water sample collection and laboratory analyses: Phosphorus and nitrogen species. *Hydrology and Earth System Sciences Discussions*, 22(1):1–33, 2018. doi: 10.5194/hess-2017-684.

L. A. Schaider, L. Swetschinski, C. Campbell, and R. A. Rudel. Environmental justice and drinking water quality: are there socioeconomic disparities in nitrate levels in us drinking water? *Environmental Health*, 18:1–15, 2019.

WHO. Guidelines for safe recreational water environments. *Coastal and Fresh Waters*, 1:1–219, 2003.