Machine Learning for Outlier Detection in Algal and Cyanobacterial Fluorescence Signals

Husein Almuhtaram[1], Arash Zamyadi[2], and Ron Hofmann[1]

[1]University of Toronto
[2]Water Research Australia

November 24, 2022

Abstract

Many drinking water utilities drawing from waters susceptible to harmful algal blooms (HABs) are implementing monitoring tools that can alert them of the onset of potential blooms. Some have invested in fluorescence-based online monitoring probes to measure chlorophyll a and phycocyanin, two pigments found in cyanobacteria, but it is not clear how to best use the data generated this way. Previous studies have focused on correlating phycocyanin fluorescence and cyanobacteria cell counts. However, not all utilities collect cell count data, making this method impossible to apply in some cases. Instead, this paper proposes a novel approach to determine when a utility needs to respond to an HAB based on machine learning by identifying outliers in chlorophyll a and phycocyanin fluorescence data without the need for corresponding cell counts or biovolume. Four existing algorithms are evaluated on data collected at four buoys in Lake Erie from 2014-2019: k-means clustering, One-Class Support Vector Machine (SVM), elliptic envelope, and Isolation Forest (iForest). When trained and tested on data collected at different buoys, the iForest algorithm performed the best in terms of computation time for training and true positive rate, and second best for false positive rate. In a more realistic application where the algorithms are trained on historical phycocyanin data collected at the same location as the testing data, all the algorithms, except k-means, accurately identified anomalies in phycocyanin data coinciding with real cyanobacteria bloom events. Therefore, One-Class SVM, elliptic envelope, and iForest are promising algorithms for detecting potential HABs using fluorescence data.

Machine Learning for Outlier Detection in Algal and Cyanobacterial Fluorescence Signals

Husein Almuhtaram[1], Arash Zamyadi[2,3], and Ron Hofmann[1]

[1] Department of Civil and Mineral Engineering, University of Toronto, Toronto ON M5S 1A4 Canada

[2] Water Research Australia (WaterRA), Adelaide, SA 5001, Australia

[3] BGA Innovation Hub and Water Research Centre, School of Civil and Environment Engineering, University of New South Wales (UNSW), Sydney, NSW 2052, Australia

Corresponding author: Husein Almuhtaram (husein.almuhtaram@mail.utoronto.ca)

Key Points:

- A method for detecting potential algal and cyanobacterial activity using only fluorescence data and machine learning was developed.
- Four machine learning algorithms were assessed: k-means, One-Class SVM, elliptical envelope, and Isolation Forest.
- Isolation Forest performed the best and identified outliers that corresponded to a real cyanobacteria bloom in Lake Erie.

Abstract

Many drinking water utilities drawing from waters susceptible to harmful algal blooms (HABs) are implementing monitoring tools that can alert them of the onset of potential blooms. Some have invested in fluorescence-based online monitoring probes to measure chlorophyll a and phycocyanin, two pigments found in cyanobacteria, but it is not clear how to best use the data generated this way. Previous studies have focused on correlating phycocyanin fluorescence and cyanobacteria cell counts. However, not all utilities collect cell count data, making this method impossible to apply in some cases. Instead, this paper proposes a novel approach to determine when a utility needs to respond to an HAB based on machine learning by identifying outliers in chlorophyll a and phycocyanin fluorescence data without the need for corresponding cell counts or biovolume. Four existing algorithms are evaluated on data collected at four buoys in Lake Erie from 2014-2019: k-means clustering, One-Class Support Vector Machine (SVM), elliptic envelope, and Isolation Forest (iForest). When trained and tested on data collected at different buoys, the iForest algorithm performed the best in terms of computation time for training and true positive rate, and second best for false positive rate. In a more realistic application where the algorithms are trained on historical phycocyanin data collected at the same location as the testing data, all the algorithms, except k-means, accurately identified anomalies in phycocyanin data coinciding with real cyanobacteria bloom events. Therefore, One-Class SVM, elliptic envelope, and iForest are promising algorithms for detecting potential HABs using fluorescence data.

1 Introduction

Cyanobacteria are increasingly threating drinking water supplies worldwide (Fernández et al., 2015). There is a need for improved monitoring to trigger responses by authorities. Traditional monitoring relies on visual observation of the source water and cell counting by microscopy (Chorus & Bartram, 1999; EPA Office of Water, 2015; Health Canada, 2016). However, visual monitoring of the water surface does not necessarily capture the conditions at the intake of a drinking water treatment plant, and microscopy is a costly, labour-intensive, and slow technique. Consequently, approaches including gene quantification by quantitative polymerase chain reaction (qPCR), remote sensing, cell imaging, and real-time fluorescence monitoring have been developed (Pacheco et al., 2016; Srivastava et al., 2013). Of these, only cell imaging and fluorescence monitoring can be implemented online at the drinking water intake: qPCR kits must be used by treatment plant staff at appropriate measurement frequencies and remote sensing is limited to capturing the conditions at or near the water surface. Automated cell imaging and identification techniques are promising but are often highly dependent on the quality of the model calibration (Jin et al., 2018).

Fluorescence monitoring probes measure the fluorescence of the cyanobacteria-specific photosynthetic pigment phycocyanin and chlorophyll a, present in all photosynthetic organisms. There is a need to find a better way for utilities to use fluorescence data to trigger a response to mitigate the effects of a developing algal bloom. In their response, a utility can also determine whether the bloom is an HAB that poses a potential toxin or taste and odour risk. The primary approach to interpreting phycocyanin fluorescence data in the literature is to correlate it to cell counts or biovolume determined by microscopy. The resulting coefficients of determination in field samples have ranged from $R^2 = 0.41$ (n = 53) to 0.87 (n = 46) (Almuhtaram et al., 2018; Bastien et al., 2011; Brient et al., 2008; Florence Choo et al., 2018; Hodges et al., 2018; McQuaid et al., 2011; Pazouki, 2016; Zamyadi, Choo, et al., 2016; Zamyadi, MacLeod, et al., 2012; Zamyadi, McQuaid, Prévost, et al., 2012). Threshold values for early warnings for cyanobacteria blooms can be set based on guideline values for cell counts or biovolumes given by various jurisdictions including the World Health Organization (WHO). However, the correlations can be site- and, if the composition of the cyanobacteria community changes, season-specific, requiring periodic validation of their accuracy by additional cell counting (Chang et al., 2012; Loisa et al., 2015).

In practice, microscopically enumerating cyanobacteria on a recurring basis is expensive, making cell counting data scarce. Consequently, monitoring data are often used without quantitative correlations to cell counts by interpreting the fluorescence pattern in a more arbitrary or qualitative way, to trigger a response. For example, Zamyadi et al. (2016b) set an arbitrary fluorescence threshold of 10% above the baseline phycocyanin readings to trigger permanganate dosing in a full-scale trial to oxidize cyanobacteria cells and microcystins. However, this approach is subjective, and may be prone to bias and inefficiency. Therefore, there is a need to interpret real-time monitoring data in a more objective and useful way for a utility to be able to determine when to initiate their HAB response strategy without relying on slow and laborious manual methods. Anomaly detection using machine-learning algorithms is a novel approach that has not yet been investigated for cyanobacteria and algae monitoring. Moreover, machine learning reduces the subjectivity involved with users interpreting real-time probe readings independently, and it has the potential to be implemented in monitoring software to trigger alarms automatically.

Anomaly detection informs utilities of when to investigate possible cyanobacteria or algae events. An anomalous data point could be due to either a change in the actual cyanobacteria or algae concentration where the monitoring probe is installed, or due to interference from turbidity or temperature, which have been shown to be significant but can be corrected for (Chang et al., 2012; F. Choo et al., 2019; Florence Choo et al., 2018; Zamyadi, Choo, et al., 2016). The objective of this study is to illustrate proof-of-concept for four unsupervised machine-learning algorithms to identify potential HABs from monitoring data collected in Lake Erie from 2014-2019 without the need for corresponding cell count data.

2 Materials and Methods

2.1 Data description

Phycocyanin and chlorophyll a are measured by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) in Lake Erie. The data from four buoys was obtained in relative fluorescence units (RFU), collected using YSI EXO2 (YSI, Yellow Springs, OH, USA) multiparameter water quality sondes equipped with Total Algae sensors. The buoy locations are shown in Figure 1. Hourly data collected in 2014 at the WE2 and WE4 buoys and data collected every 15 min in 2015-2019 at all four buoys was obtained from the GLERL publicly available archives (NOAA/GLERL, 2020a). The sondes were serviced on an approximately monthly basis during the monitoring period and replaced with cleaned and calibrated (two-point calibration with Rhodamine WT dye) sondes. The data are available in the NOAA GLERL Lake Erie real-time data archives (NOAA/GLERL, 2020b).

2.2 Machine learning algorithms

Four machine-learning algorithms for unsupervised anomaly detection were applied to the Lake Erie monitoring data: k-means clustering, One-Class Support Vector Machine (SVM), elliptic envelope, and Isolation Forest (iForest). These algorithms are unsupervised because they use unlabelled data. That is, the algorithms do not know whether any of the data points inputted to them are normal or anomalous. All of the data analysis was conducted in the Python programming language (V. 3.7.3) using the scikit-learn machine-learning package (V. 0.20.3) (Pedregosa et al., 2011).

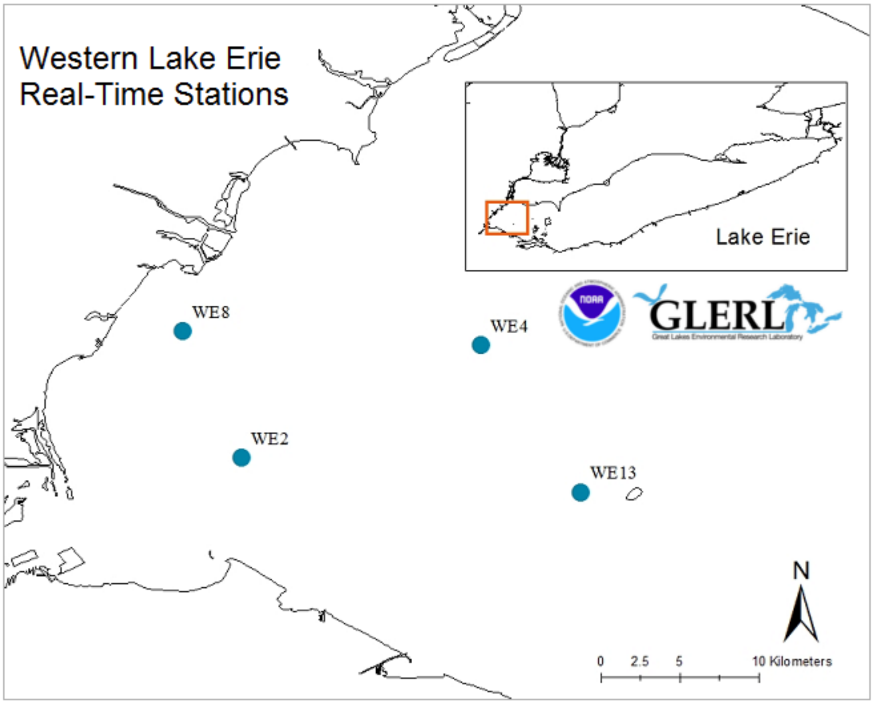


Figure 1 Location of the four monitoring buoys in Lake Erie.

The k-means clustering algorithm creates a specified number of clusters, k, of data points, iteratively adjusting the centroids of the clusters until the sum of the distance between the data points and the center of the clusters (sum of squares) is minimized. Points that fall outside of the clusters or in small clusters are deemed anomalous. The number of clusters can be optimized using an 'elbow curve' that shows the number of clusters beyond which improvements to the sum of squares become progressively smaller. Then, the phycocyanin and chlorophyll a data can be divided into that number of clusters to identify outliers. However, because the clustering iteration initializes k cluster centers at random, the output is different every time the algorithm is run. The accuracy of the algorithm can be improved by using the k-means++ method for initialization, included in the scikit-learn package, which randomly selects k centers but weighs the data according to the initial sum of squares to improve the initialization procedure (Arthur & Vassilvitskii, 2007).

In contrast, the One-Class SVM algorithm estimates a function that returns positive for normal (non-outlier) data points and negative for outliers when the probability that any data point is not an outlier is known. This is accomplished by mapping the data points into a feature space corresponding to the radial basis function kernel (commonly used in SVM algorithms) and separating them from the origin with a maximum margin hyperplane in a higher dimension feature space using a minimization formulation (Schölkopf et al., 2001).

For a dataset $\{(x_1, x_2, \ldots, x_i)\}$ where x_i is the i-th data point, the data points are lifted to a higher dimension feature space F via a non-linear function ϕ. This is so that a straight line, a hyperplane, can separate the data points into the two classes whereas a complex non-linear curve would have been required in the original dimension. Because noisy data can make the separation between the classes unclear, a slack variable, ξ, is introduced that allows data points to lie within the margin, in the proportion delimited by the parameter v. The minimization formulation

$$\min_{\omega,\xi,\rho} \frac{\|\omega\|^2}{2} + \frac{1}{vn}\sum_{i=1}^{n} \xi_i - \rho \qquad (1)$$

$$\text{subject to } (w \cdot \phi(x_i)) \geq \rho - \xi_i,\ \xi_i \geq 0$$

separates the data from the origin with a maximum margin, and a decision function

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i\, K(x, x_i) - \rho\right) \qquad (2)$$

assesses whether a data point is normal or anomalous, returning a value of +1 for normal and -1 for anomalous. In this equation, α_i are the Lagrange multipliers and $K(x, x_i)$ is the radial basis function kernel

$$K(x, x_i) = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{3}$$

where $\sigma \in R$ is a kernel parameter and $\|x - x'\|$ is the dissimilarity measure.

Unlike the maximum margin method of the One-Class SVM, the elliptic envelope models the data to a Gaussian distribution and identifies an ellipse that contains most of the data; a data point outside the ellipse is anomalous. The size and shape of the ellipse are determined by the FAST-Minimum Covariance Determinant algorithm (Rousseeuw & Driessen, 1999). This algorithm iteratively computes the Mahalanobis distance (a measure of how many standard deviations a data point is from the mean) of subsamples from the data until the determinant of the covariance matrix converges (Hoyle et al., 2015). An ellipse is constructed using the covariance matrix with the smallest determinant for all subsamples with the use of a hyper-parameter: the contamination rate of the dataset. The contamination rate is defined as the approximate proportion of data points that lie outside the ellipse, inputted before the algorithm is run. It has been reported that the contamination rate does not necessarily need a high degree of accuracy, so it can be estimated initially and adjusted in subsequent runs of the algorithm (Hoyle et al., 2015).

Lastly, a fourth and fundamentally different anomaly detection approach, iForest, is used. The preceding three methods rely on building a profile of the data set and identifying outliers by various metrics. In iForest, anomalous points are explicitly isolated based on the fact that they are few and different, and no profile of the normal data is constructed (Liu et al., 2008). The algorithm constructs isolation trees by recursively partitioning every feature (variable) of the data split randomly between its maximum and minimum values. It was found that normal data points require more partitions, termed the ‘path length’, to be isolated than anomalous points, which require a lower path length (Liu et al., 2008). Because each isolation tree is made up of random partitions, as the number of isolation trees increases, the average path length required to isolate a point converges. This value is used to calculate an anomaly score that identifies a data point as normal or anomalous.

The iForest algorithm comprises two phases: training and testing. In the training phase, up to 100 isolation trees are constructed using a subset of the data: 256 data points per tree by default. At the end of the training phase, an isolation forest is returned. Next, the testing phase passes each vector (data point) through each isolation tree, and the path length to termination is stored and used to calculate the anomaly score.

2.3 Validation

The four algorithms detect outliers in fluorescence data, but it is not known if the outliers identified correspond to real harmful algal bloom (HAB) events: that is, blooms of cyanobacteria. Real HAB events in Lake Erie 2019 are identified using the NOAA GLERL Experimental Lake Erie HAB Tracker and compared to the algorithm outputs when trained and tested only on phycocyanin (NOAA/GLERL, 2020a). Chlorophyll a is omitted from the training and testing data because it represents not only cyanobacteria but also green algae, whereas the HAB Tracker is specifically designed to identify cyanobacteria blooms.

3 Results

3.1 Training data

Four unsupervised algorithms were implemented for anomaly detection of phycocyanin and chlorophyll a RFU data. Unsupervised algorithms have no knowledge of the correct classes for each data point, but they derive internally generated error measures to classify data based solely on the statistics of the training data (Kyan et al., 2014). Therefore, a training dataset should be abundant and diverse (Gong et al., 2019). As such, there are two possible training and testing scenarios for the current data: the algorithms can be trained on data from one of the four buoys and tested on the other three, or they can be trained on historical data for one buoy and tested on the most recent data collected in that location. In the former, the WE2 buoy data is the best candidate. It contains over 75,000 data points collected from the 2014-2019 algae seasons, shown in Figure 2. Although the WE4 buoy data contains approximately the same amount of data, it is not as diverse; the variances of the phycocyanin and chlorophyll a values, both 0.48, were lower than those of the WE2 buoy data at 1.9 and 1.3, respectively. Additionally, the WE4 buoy data results in the most conservative decision functions for each algorithm (Figure S1), and the data from the WE13 buoy produced similar decision functions (Figure S2). On the other hand, the WE8 buoy data was the least conservative (Figure S3). Therefore, training was performed using the more diverse WE2 data, which resulted in moderate decision functions. The diversity of this data set is likely due to the buoy's position in Lake Erie (Figure 1) being closest to the Maumee River, which is believed to be the most significant source of nutrients contributing to HAB development in Lake Erie (Stumpf et al., 2012). The lake experienced diverse conditions with blooms of varying severity in terms of algal biomass and bloom extent from 2014-2019, and it appears that these conditions were best represented by the WE2 buoy data (NOAA & NCWQR, 2019).

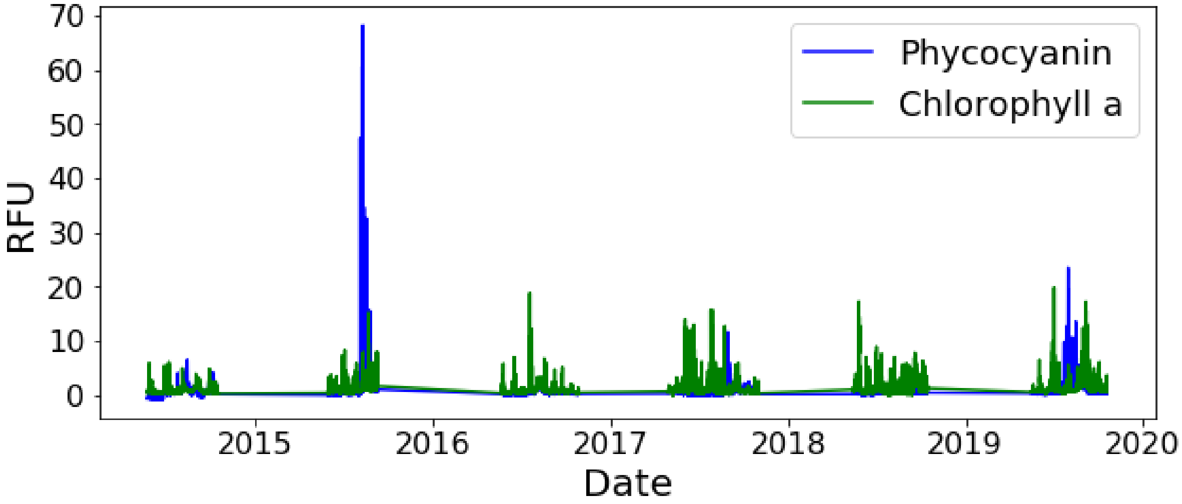


Figure 2 2014-2019 WE2 buoy data were used to train all four algorithms.

The classification of data by unsupervised algorithms is not only dependent on the training data but also a sensitivity parameter. Each of the algorithms used in this work has such a parameter: in k-means clustering a predefined number of the data points farthest from the cluster centres are identified as anomalous, starting with the farthest; in One-Class SVM a value is inputted that sets an upper bound on the proportion of outliers; and, in elliptic envelope and iForest a contamination rate is inputted that approximates the proportion of outliers in the dataset. Notably, the actual proportion of outliers determined is not necessarily equal to the predefined contamination rate (Hoyle et al., 2015). The optimal contamination rate varies for every dataset. In this study, the optimal contamination rate was determined to be 0.05 by visually examining the sensitivity of each algorithm in response to inputting contamination rates of 0.01, 0.05, and 0.1. This evaluation is external to the algorithms and relies on user judgement to validate the model. Baseline data and fluorescence peaks are understood to be 'normal' and 'anomalous', respectively. Figures S4 and S5 show that for all four algorithms using values of 0.01 and 0.1 result in the classification of baseline data as anomalous and the classification of data peaks as normal, respectively. Therefore, a contamination rate of 0.05 was used throughout this study.

The training results for all four algorithms with a contamination rate of 0.05 are shown in Figure 3. Note that it appears that the encircled regions in cyan seem to be smaller than the red regions outside, implying that more than half of the readings are 'anomalous'. However, the cyan region contains many more data points that are overlapped than the red region. Figure S6 shows that 95% of the chlorophyll a and phycocyanin RFU data are below 3.75 and 2.24, respectively, for the 2014-2019 WE2 buoy data.

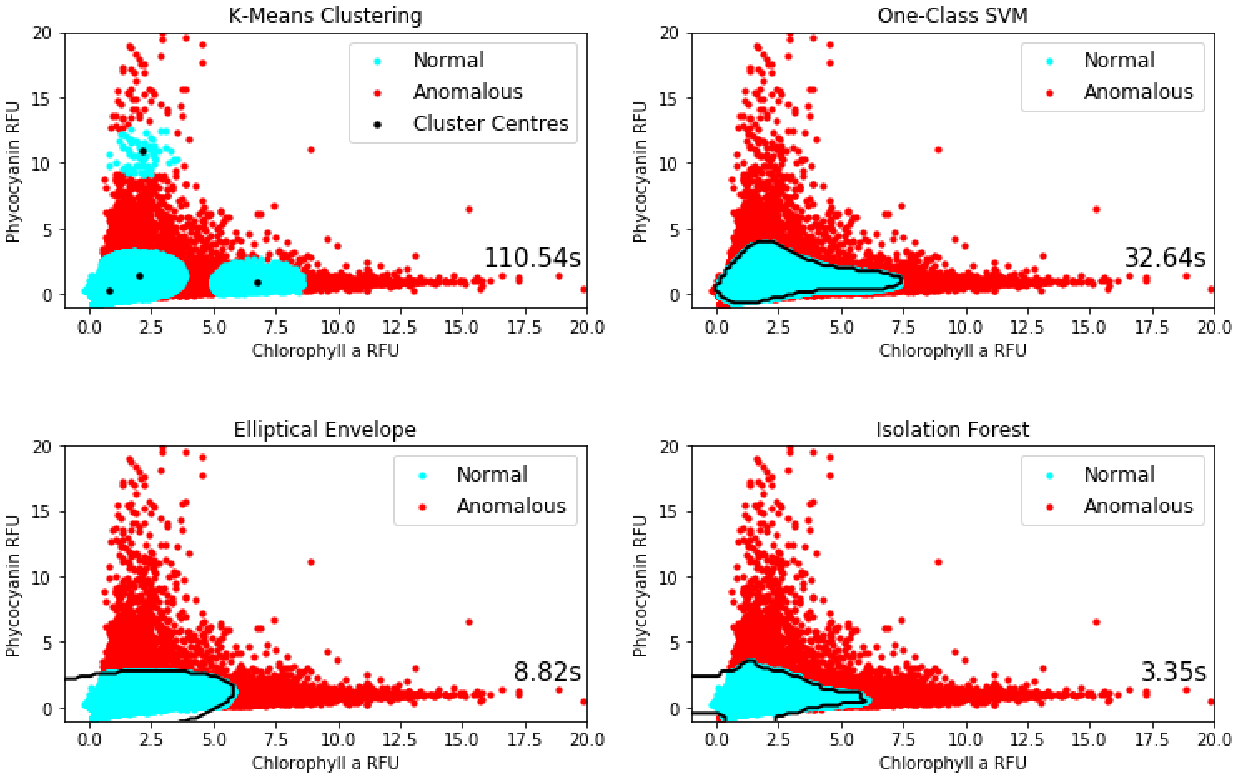


Figure 3 Normal and anomalous data points determined by the algorithms following testing on the 2014-2019 WE2 dataset. The solid lines represent the contours that separate the inliers from the outliers. The k-means method differs in that it uses the distance between a data point and the nearest cluster center of the training data to determine if it is an outlier, hence no contour is displayed. The computation time for training and testing both datasets is included in the bottom-right corner.

3.2 K-means clustering

The 2014-2019 data collected at the WE2 buoy were split into four clusters by the k-means algorithm. The optimal number of clusters was determined using an elbow curve, shown in Figure 4. The curve illustrates that beyond four clusters, the k-means score, the negative of the sum of distances between data points and their cluster centres, improves at a much slower rate for every additional cluster.

The top 5% of data points with the largest distance to the nearest centroid centre are labelled as anomalies. Consequently, the anomaly detection relies heavily on the success of the clustering procedure in the training dataset to be applicable to the testing dataset. Additionally, the computation time of the k-means method was over 110 s because after constructing the clusters in the training set and assigning the testing set data points to them, it checks the distance between each data point and its cluster center one by one. Figure 3 presents the outcome of training on the WE2 dataset, illustrating that generally points with high phycocyanin and chlorophyll a fluorescence are considered anomalous. However, the presence of a cluster centre from the training data in the upper phycocyanin range prevents the identification of some of the data as anomalous. When tested on the 2014-2019 WE4 data and plotted as a time-series, shown in Figure 5, it is apparent that some of the chlorophyll a and phycocyanin peaks are not identified as anomalous, indicating that this algorithm has a dissatisfactory true positive rate. Therefore, k-means clustering is not suitable for this data, although it may perform better in other cases.

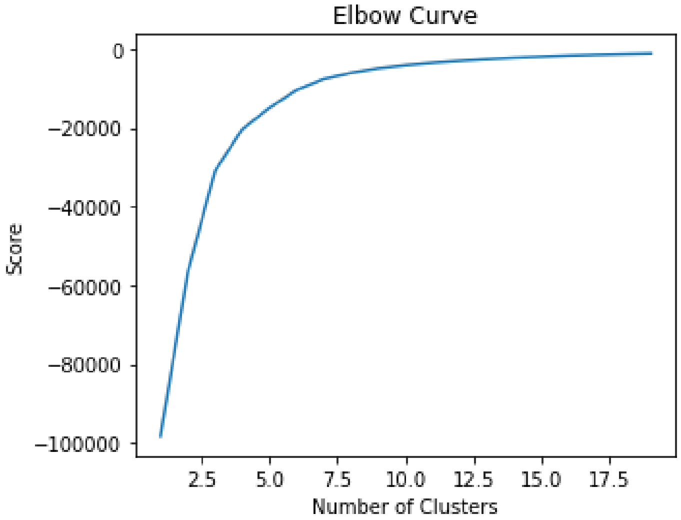


Figure 4 Elbow curve for the WE2 2014-2019 phycocyanin and chlorophyll a RFU data. The optimal number of clusters for this data appears to be 4.

3.3 One-Class SVM\

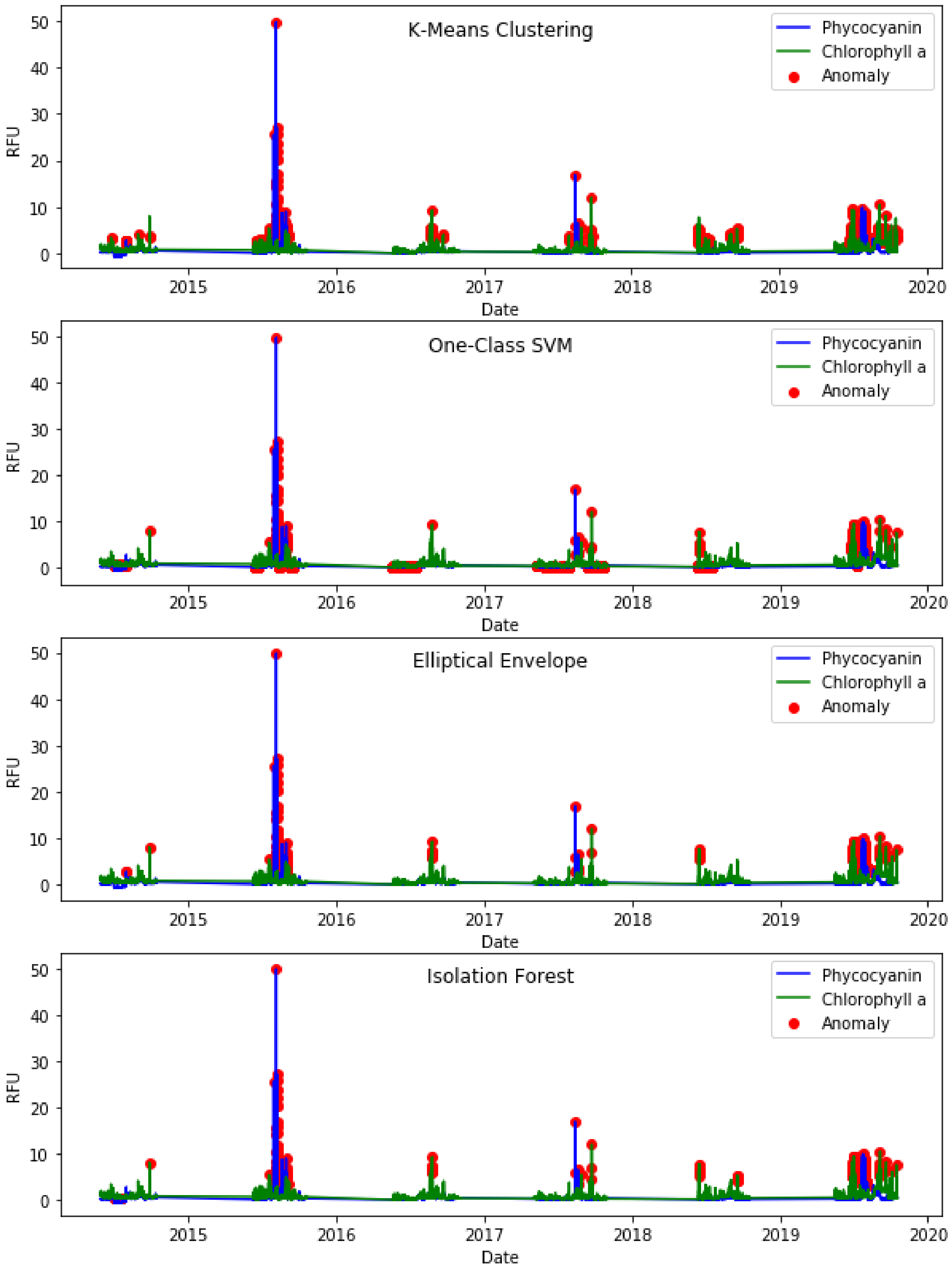
The same 2014-2019 WE4 dataset was analyzed using the One-Class SVM algorithm. Like in the k-means approach, the 2014-2019 WE2 training data were standardized and a contamination rate of 5% was assumed. Then, the algorithm mapped the data according to the feature space corresponding to the radial basis function kernel and separated the data from the origin with a maximum-margin hyperplane in a higher dimension (Schölkopf et al., 2001). Returning to the two-dimensional data, the hyperplane appears as the contour shown in Figure 3. The One-Class SVM algorithm learned the training data in a computation time of 33 s, a significant improvement over the k-means algorithm. Data points within the boundaries of the solid line are inliers and any points outside of it are outliers. This plot reveals that phycocyanin fluorescence values between 3-4 RFU are considered anomalous when the chlorophyll a fluorescence is low, but as chlorophyll a increases to 2 RFU, that phycocyanin range is considered normal. A similar pattern is seen for chlorophyll a fluorescence where values from 2.5-7.5 RFU are anomalous when phycocyanin is near zero but normal when phycocyanin is at 2 RFU. In Figure 5, the One-Class SVM algorithm is tested on the 2014-2019 WE4 dataset, showing that almost all the large and small phycocyanin and chlorophyll a fluorescence peaks were identified as anomalous by the One-Class SVM algorithm (i.e. high true positive rate), but so were some of the baseline data (i.e. high false positive rate), indicating that this algorithm's performance is not satisfactory for this dataset.

3.4 Elliptic envelope

The elliptic envelope technique computes the Mahalanobis distance, mean, and covariance matrices between randomly selected non-overlapping subsamples. Then, subsamples with low Mahalanobis distances are selected and their means, the covariance between them, and the Mahalanobis distances are calculated again. This process iterates until the determinant of the covariance matrix converges, and the subsamples forming the covariance matrix with the smallest determinant form an ellipse. The ellipse for the 2014-2019 WE2 dataset is shown in Figure 3 and was computed in only 9 s, a significant improvement over the computation time of the One-Class SVM algorithm. Unlike the One-Class SVM contour, it includes most of the data with elevated chlorophyll a fluorescence and a near uniform cut-off for phycocyanin above about 3.5 RFU. Additionally, when tested on the 2014-2019 WE4 dataset (Figure 5), the algorithm is less sensitive than the One-Class SVM algorithm in the lower range, such that none of the baseline data is identified as anomalous (i.e. low false positive rate), but it labels all points with chlorophyll a fluorescence above 5 RFU as anomalous illustrating more sensitivity than the One-Class SVM algorithm when phycocyanin RFU is below 2.5. Therefore, this algorithm appears to be a promising tool for detecting outliers in chlorophyll a and phycocyanin fluorescence data.

3.5 Isolation Forest

By default, the iForest algorithm builds an isolation forest from 100 isolation trees using 256 data points each. Each data point in the testing dataset passes through the forest and based on the path length to the terminating node an anomaly score is derived. The decision function of the iForest algorithm is comparable in shape to the One-Class SVM model for the 2014-2019 WE2 data, shown in Figure 3, but with a computation time about 10x faster. Moreover, it extends beyond the axes into the negative range so that fewer near-zero data points are mistakenly labelled anomalous, as opposed to the One-Class SVM decision function. Using the same contamination rate as the three previous methods (5%), the algorithm is sensitive to both chlorophyll a and phycocyanin fluorescence. In the time-series visualization for the 2014-2019 WE4 dataset, the performance is apparently satisfactory as all of the large and most of the small peaks were identified as anomalous (i.e. high true positive rate), as shown in Figure 5, although a few data points with negative phycocyanin RFU values were also labelled as anomalous (i.e. low false positive rate). Overall, iForest appears to be the most promising algorithm for unsupervised anomaly detection in this type of data.



280

281 **Figure 5** Time-series representation of anomaly detection on the 2014-2019 WE4 dataset with a contamination rate of 5%.

3.6 Validation

Although visually the iForest algorithm appears to perform the best, the anomalies identified must correspond to real cyanobacteria events for the approach to be useful. However, this is contingent upon not only the effectiveness of the algorithm, but also the collection of accurate measurements by the fluorometer used to generate the data. The experimental Lake Erie HAB Tracker was used to retroactively identify real bloom events occurring in the approximate position of the WE2 buoy shown in Figure 1. HAB Tracker data is available from July 2 to October 4, 2019.

Figure 6 illustrates a snapshot of the 2019 WE2 buoy data tested for outliers using the iForest algorithm, trained using data collected at the same buoy from 2014-2018, with yellow shaded regions superimposed denoting bloom events identified using the HAB Tracker. It is evident that the iForest results correctly identified outliers during the major bloom event identified by the HAB Tracker from July 11 to September 6, 2019 as well as the recurring blooms at the start of this period. However, at the end of this event, from September 6 to 10, no anomalous events were detected although the bloom persisted. Similarly, a short-lived cyanobacteria event occurred from September 19 to 21 but was not detected by the phycoyanin sensor and consequently not identified as anomalous. In both cases, it is possible that the limited spatial resolution of the satellite image prevent an accurate alignment of the WE2 buoy position with the discontinuous bloom patches such that the buoy was positioned outside of the bloom and therefore no phycocyanin was detected by the probe (Figure S7).

This illustrates the validity of the application of the iForest algorithm to detect cyanobacteria events from phycocyanin monitoring data. When the same process is carried out for the other three algorithms, the k-means approach fails to identify numerous peaks as anonymous but the One-Class SVM and elliptic envelope results are similar to the iForest results, although not as sensitive (Figure S8). So, while iForest performs the best when trained on data collected in a different location (i.e., can be generalized, Figure 5), One-Class SVM and elliptic envelope are satisfactory when trained on historical data for the testing location, which represents a more likely application.

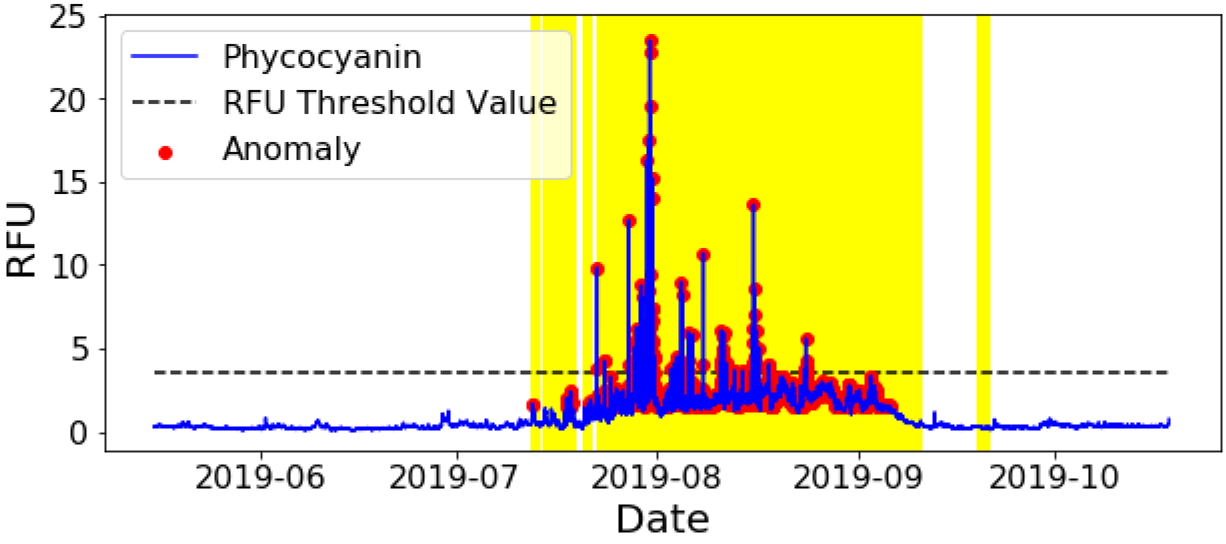


Figure 6 Outlier detection using the iForest algorithm trained on the WE2 phycocyanin data from 2014-2018 and tested on the 2019 phycocyanin data. The yellow shaded regions indicate the presence of a cyanobacteria bloom identified by the NOAA Experimental HAB Tracker using satellite data. The 3.6 RFU threshold value corresponds to the WHO Alert Level 1, previously determined for four Great Lakes region treatment plants in Almuhtaram et al. (2018).

4 Discussion and Conclusion

The One-Class SVM, elliptic envelope, and iForest algorithms might be valuable to utilities that already have fluorescence-based probes installed at their intakes or a raw water sampling line. They may eliminate the need for pairing microscopically enumerated cyanobacteria to phycocyanin fluorescence to establish a fluorescence threshold that, if surpassed, triggers reactive measures at the utility. The advantage of using machine learning is that continuously collected monitoring data can be added to an unsupervised algorithm's training set so that the underlying structure of the data distribution is learned (Abou-Moustafa & Schuurmans, 2015). This means that the algorithm can be applied to any source water provided that enough data is available. Moreover, the outlier detection by iForest conducted on the Lake Erie datasets illustrates generalization: that is, different datasets were used for training and testing to demonstrate that the algorithm is effective across a range of inputs. In practice, however, it is not necessary to separate the training and testing datasets; all the chlorophyll a and phycocyanin monitoring data collected at a utility can be used for training, and testing can focus on the newest data points added to that set at a desired frequency (e.g. testing for outliers in the past hour's data). Additionally, the algorithms can be used to conduct outlier detection for only one variable. An example of this is if a utility is interested specifically in the occurrence of cyanobacteria blooms and not green algae, phycocyanin can be used as the sole variable for training and testing (Figure 6).

Another possible advantage of using machine learning is the ability to detect the onset of a cyanobacteria bloom if the data are suitable. The gradual growth of cyanobacteria will result in gradual fluorescence increases, which can be theoretically identified as anomalies using machine learning. However, this was not observed in Figure 6 as the first bloom occurred suddenly such that no gradual increase was detected by the probe. Still, machine learning may result in better sensitivity and earlier detection than existing practices. A previous study employed a fluorescence-based probe to monitor cyanobacteria across four Great Lakes region treatment plants to assess cyanobacteria breakthrough and accumulation (Almuhtaram et al., 2018). Phycocyanin fluorescence was correlated to cyanobacteria cell counts among the raw water samples of the plants, which included samples from Lake Erie. A phycocyanin fluorescence threshold value of 3.6 RFU for a cyanobacteria biovolume of 0.2 mm^3/L, corresponding to the WHO Alert Level 1, was determined, and is added to Figure 6. This example demonstrates that the machine learning approach is more sensitive than the 3.6 RFU threshold and corresponds to the start of the bloom whereas the threshold may only detect the bloom when it becomes severe.

The current approach may therefore provide water utilities, and health and regulatory agencies with a management tool to detect the presence of cells in time to implement their management strategies prior to bloom breakthrough. Also, equivalent fluorescence threshold values such as that reported by Almuhtaram et al. (2018) are site-specific; hence authorities need to have a proper understanding of their water body and HAB dynamics and use long term monitoring data to develop their threshold values by matching fluorescence readings with taxonomic analysis data (Macário et al., 2015). The machine learning approach, in contrast, is not reliant on finding threshold values that represent a predetermined risk level, such as the WHO Alert Level 1. Instead, statistical anomalies are brought to the attention of the utility that must then investigate the associated risk in terms of cell, toxin, or taste and odour concentration.

This approach has the potential to be adopted by fluorometer manufacturers. Outlier detection can be conducted in real time provided that an outlier detection algorithm is included in the monitoring software. Data measured by the sensors can be used to create a training set, and as individual measurements are logged (e.g., every 15 min), the algorithm can test the newest point to immediately determine whether it is an outlier and trigger an alarm. The freedom to select which parameters to test for should also be included, such as both chlorophyll a and phycocyanin for detecting all algal events or only phycocyanin for detecting cyanobacteria. In theory, this approach could be applied to any continuously monitored water quality parameter such as turbidity or dissolved oxygen, although this requires further investigation. Moreover, machine learning makes use of all the data collected by a sensor to characterize the entire range of conditions encountered in its lifetime. In contrast, the existing approach uses a small subset of the available phycocyanin data to pair with grab sample measurements of cyanobacteria cell concentration.

When a potential HAB is identified using either the novel machine learning approach or the correlation approach, utilities need to conduct further investigations to determine whether harmful algae or cyanobacteria are present. This is because fluorescence-based probes are inherently susceptible to false positives due to interference by turbidity, temperature, and dissolved pigments (Zamyadi, McQuaid, Dorner, et al., 2012). Nonetheless, use of either method results in an improvement over traditional HAB and cyanotoxin monitoring practices that rely on visual monitoring and weekly microcystins measurements by enzyme-linked immunosorbent assays (ELISA) in that a sample can be taken in response to a potential HAB event, rather than on an arbitrary day of the week. Machine learning is therefore a promising way to utilize fluorescence data to alert a utility to potential HAB events.

Acknowledgments and Data

- The authors declare no conflict of interest.
- The data used and the Python code used to generate the figures can be found at (https://zenodo.org/record/3887222).
- This work was supported by the Natural Sciences and Engineering Research Council of Canada Project CRDPJ 482052-15. The authors also acknowledge the funding and support from the City of Toronto, the City of Hamilton, the Regional Municipality of Niagara, the Regional Municipality of York, the Regional Municipality of Durham, and the Union Water Supply System.

References

Abou-Moustafa, K. T., & Schuurmans, D. (2015). Generalization in Unsupervised Learning. In *Efficient Learning Machines* (pp. 300–317). Berkeley, CA: Apress. https://doi.org/10.1007/978-3-319-23528-8_19

Almuhtaram, H., Cui, Y., Zamyadi, A., & Hofmann, R. (2018). Cyanotoxins and Cyanobacteria Cell Accumulations in Drinking Water Treatment Plants with a Low Risk of Bloom Formation at the Source. *Toxins*, *10*(11), 430. https://doi.org/10.3390/toxins10110430

Arthur, D., & Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics* (pp. 1027–1035). New Orleans, Louisiana.

Bastien, C., Cardin, R., Veilleux, É., Deblois, C., Warren, A., & Laurion, I. (2011). Performance evaluation of phycocyanin probes for the monitoring of cyanobacteria. *J. Environ. Monit.*, *13*(1), 110–118. https://doi.org/10.1039/C0EM00366B

Brient, L., Lengronne, M., Bertrand, E., Rolland, D., Sipel, A., Steinmann, D., et al. (2008). A phycocyanin probe as a tool for monitoring cyanobacteria in freshwater bodies. *J. Environ. Monit.*, *10*(2), 248–255. https://doi.org/10.1039/B714238B

Chang, D. W., Hobson, P., Burch, M., & Lin, T. F. (2012). Measurement of cyanobacteria using in-vivo fluoroscopy - Effect of cyanobacterial species, pigments, and colonies. *Water Research*, *46*(16), 5037–5048. https://doi.org/10.1016/j.watres.2012.06.050

Choo, F., Zamyadi, A., Stuetz, R. M., Newcombe, G., Newton, K., & Henderson, R. K. (2019). Enhanced real-time cyanobacterial fluorescence monitoring through chlorophyll-a interference compensation corrections. *Water Research*, *148*, 86–96. https://doi.org/10.1016/j.watres.2018.10.034

Choo, Florence, Zamyadi, A., Newton, K., Newcombe, G., Bowling, L., Stuetz, R., & Henderson, R. K. (2018). Performance evaluation of in situ fluorometers for real-time cyanobacterial monitoring. *H2Open Journal*, *1*(1), 26–46. https://doi.org/10.2166/h2oj.2018.009

Chorus, I., & Bartram, J. (1999). *Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management*. *Retrieved March*. https://doi.org/10.1046/j.1365-2427.2003.01107.x

EPA Office of Water. (2015). *Recommendations for Public Water Systems to Manage Cyanotoxins in Drinking Water*.

Fernández, C., Estrada, V., & Parodi, E. R. (2015). Factors Triggering Cyanobacteria Dominance and Succession During Blooms in a Hypereutrophic Drinking Water Supply Reservoir. https://doi.org/10.1007/s11270-014-2290-5

Gong, Z., Zhong, P., & Hu, W. (2019). Diversity in Machine Learning. *IEEE Access*, *7*, 64323–64350. https://doi.org/10.1109/ACCESS.2019.2917620

Health Canada. (2016). *Cyanobacterial Toxins in Drinking Water*.

Hodges, C. M., Wood, S. A., Puddick, J., McBride, C. G., & Hamilton, D. P. (2018). Sensor manufacturer, temperature, and cyanobacteria morphology affect phycocyanin fluorescence measurements. *Environmental Science and Pollution Research*, *25*(2), 1079–1088. https://doi.org/10.1007/s11356-017-0473-5

Hoyle, B., Rau, M. M., Paech, K., Bonnett, C., Seitz, S., & Weller, J. (2015). Anomaly detection for machine learning redshifts applied to SDSS galaxies. *Monthly Notices of the Royal Astronomical Society*, *452*(4), 4183–4194. https://doi.org/10.1093/mnras/stv1551

Jin, C., Mesquita, M. M. F. F., Deglint, J. L., Emelko, M. B., & Wong, A. (2018). Quantification of cyanobacterial cells via a novel imaging-driven technique with an integrated fluorescence signature. *Scientific Reports*, *8*(1), 9055. https://doi.org/10.1038/s41598-018-27406-0

Kyan, M., Muneesawang, P., Jarrah, K., & Guan, L. (2014). *Unsupervised Learning*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118875568

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

Loisa, O., Kääriä, J., Laaksonlaita, J., Niemi, J., Sarvala, J., & Saario, J. (2015). From phycocyanin fluorescence to absolute cyanobacteria biomass: An application using in-situ fluorometer probes in the monitoring of potentially harmful cyanobacteria blooms. *Water Practice and Technology*, *10*(4), 695–698. https://doi.org/10.2166/wpt.2015.083

Macário, I. P. E., Castro, B. B., Nunes, M. I. S., Antunes, S. C., Pizarro, C., Coelho, C., et al. (2015). New insights towards the establishment of phycocyanin concentration thresholds considering species-specific variability of bloom-forming cyanobacteria. *Hydrobiologia*, *757*(1), 155–165. https://doi.org/10.1007/s10750-015-2248-7

McQuaid, N., Zamyadi, A., Prévost, M., Bird, D. F., & Dorner, S. (2011). Use of in vivophycocyanin fluorescence to monitor potential microcystin-producing cyanobacterial biovolume in a drinking water source. *J. Environ. Monit.*, *13*(2), 455–463. https://doi.org/10.1039/C0EM00163E

NOAA/GLERL. (2020a). Experimental Lake Erie Harmful Algal Bloom (HAB) Tracker. Retrieved from https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/habTracker.html

NOAA/GLERL. (2020b). Lake Erie Archived Real-Time Data. Retrieved from https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/rtMonSQL.php

NOAA, & NCWQR. (2019). *Western Lake Erie Harmful Algal Bloom Early Season Projection*.

Pacheco, A., Guedes, I., & Azevedo, S. (2016). Is qPCR a Reliable Indicator of Cyanotoxin Risk in Freshwater? *Toxins*, *8*(6), 172. https://doi.org/10.3390/toxins8060172

Pazouki, P. (2016). *Cyanobacteria in surface and bank filtered drinking water sources: application of phycocyanin probes for monitoring blooms*. ÉCOLE POLYTECHNIQUE DE MONTRÉAL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://arxiv.org/abs/1201.0490

Rousseeuw, P., & Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*(3), 212–223.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965

Srivastava, A., Singh, S., Ahn, C.-Y., Oh, H.-M., & Asthana, R. K. (2013). Monitoring Approaches for a Toxic Cyanobacterial Bloom. *Environmental Science & Technology*, *47*(16), 8999–9013. https://doi.org/10.1021/es401245k

Stumpf, R. P., Wynne, T. T., Baker, D. B., & Fahnenstiel, G. L. (2012). Interannual Variability of Cyanobacterial Blooms in Lake Erie. *PLoS ONE*, *7*(8), e42444. https://doi.org/10.1371/journal.pone.0042444

Zamyadi, A., McQuaid, N., Dorner, S., Bird, D. F., Burch, M., Baker, P., et al. (2012). Cyanobacterial detection using in vivo fluorescence probes: Managing interferences for improved decision-making. *Journal - American Water Works Association*, *104*(8). https://doi.org/10.5942/jawwa.2012.104.0114

Zamyadi, A., McQuaid, N., Prévost, M., & Dorner, S. (2012). Monitoring of potentially toxic cyanobacteria using an online multi-probe in drinking water sources. *Journal of Environmental Monitoring*, *14*(2), 579–588. https://doi.org/10.1039/c1em10819k

Zamyadi, A., MacLeod, S. L., Fan, Y., McQuaid, N., Dorner, S., Sauvé, S., & Prévost, M. (2012). Toxic cyanobacterial breakthrough and accumulation in a drinking water plant: A monitoring and treatment challenge. *Water Research*, *46*(5), 1511–1523. https://doi.org/10.1016/j.watres.2011.11.012

Zamyadi, A., Choo, F., Newcombe, G., Stuetz, R., & Henderson, R. K. (2016). A review of monitoring technologies for real-time management of cyanobacteria: Recent advances and future direction. *TrAC Trends in Analytical Chemistry*, *85*, 83–96. https://doi.org/10.1016/j.trac.2016.06.023

Zamyadi, A., Henderson, R. K., Stuetz, R., Newcombe, G., Newtown, K., & Gladman, B. (2016). Cyanobacterial management in full-scale water treatment and recycling processes: reactive dosing following intensive monitoring. *Environ. Sci.: Water Res. Technol.*, *2*(2), 362–375. https://doi.org/10.1039/C5EW00269A

AGU PUBLICATIONS

Water Resources Research

Supporting Information for

Machine Learning for Outlier Detection in Algal and Cyanobacterial Fluorescence Signals

Husein Almuhtaram[1], Arash Zamyadi[2,3], and Ron Hofmann[1]

[1] Department of Civil and Mineral Engineering, University of Toronto, Toronto ON M5S 1A4 Canada

[2] Water Research Australia (WaterRA), Adelaide, SA 5001, Australia

[3] BGA Innovation Hub and Water Research Centre, School of Civil and Environment Engineering, University of New South Wales (UNSW), Sydney, NSW 2052, Australia

Corresponding author: Husein Almuhtaram (husein.almuhtaram@mail.utoronto.ca)

Contents of this file

Figures S1 to S8.

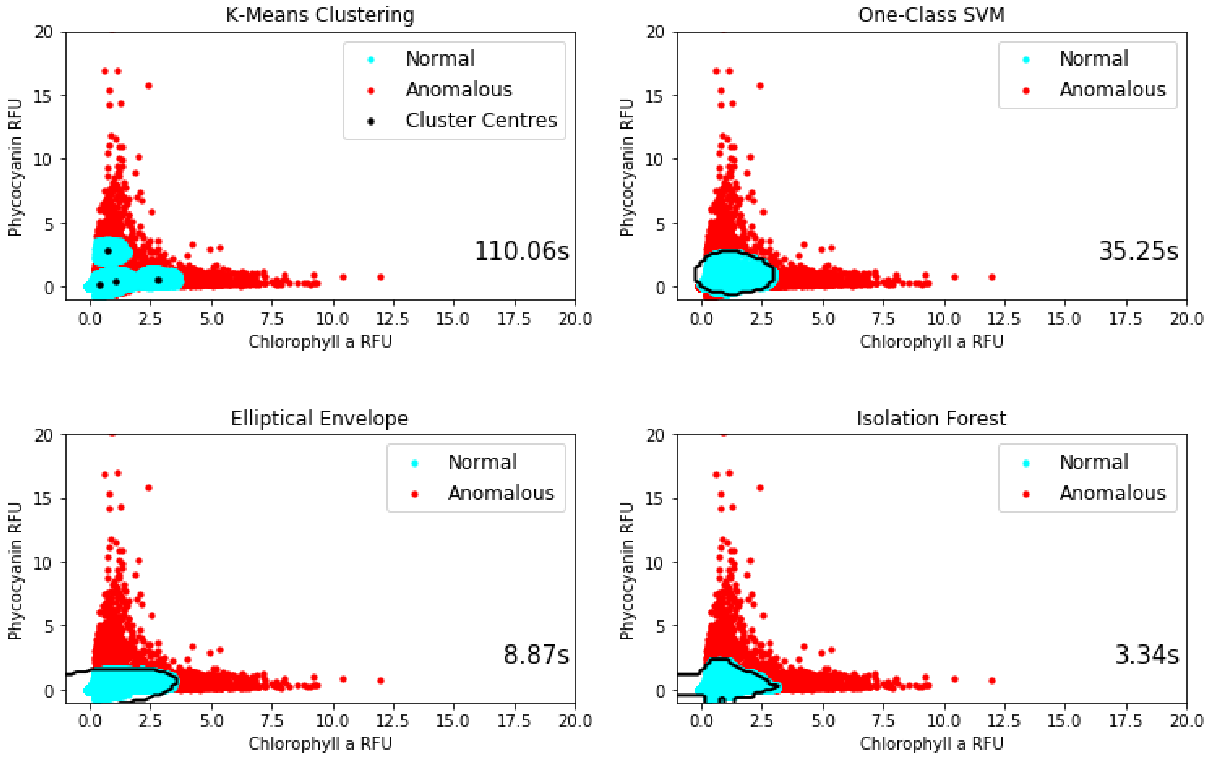


Figure S1. Algorithms trained on the 2014-2019 WE4 dataset.

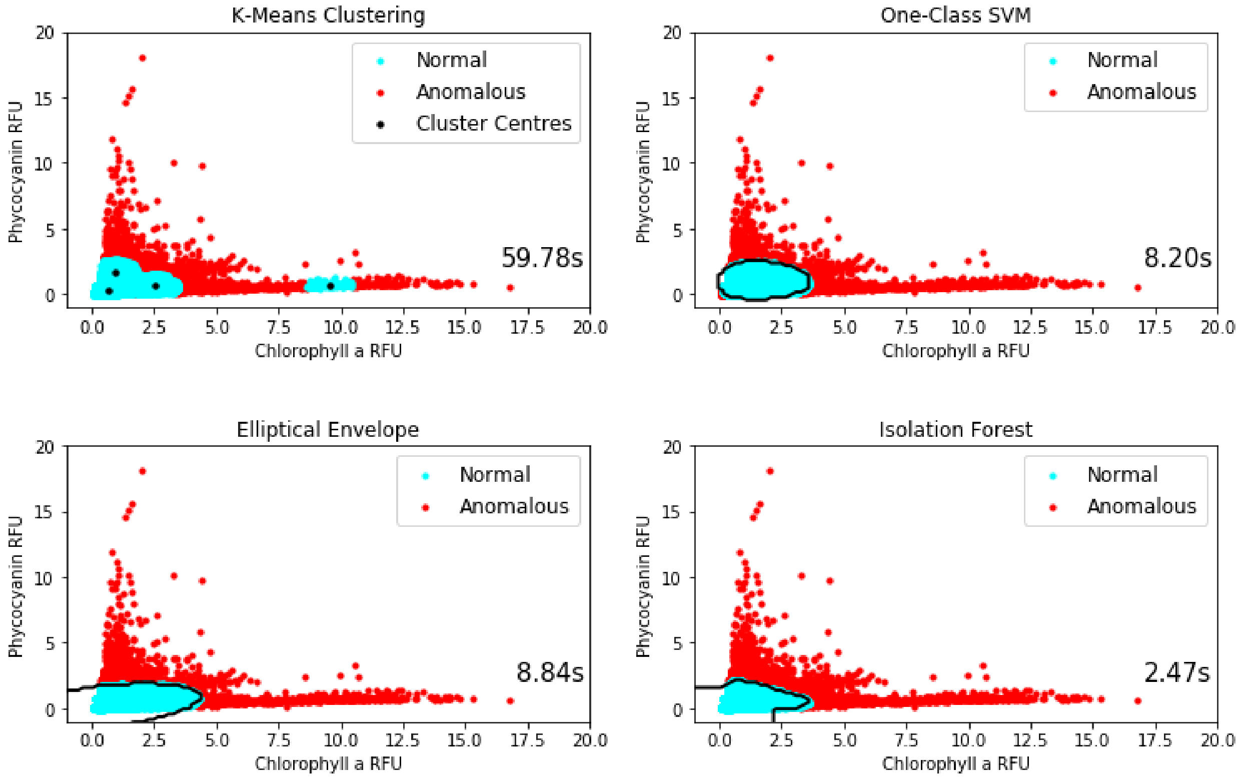


Figure S2. Algorithms trained on the 2015-2019 WE13 dataset.

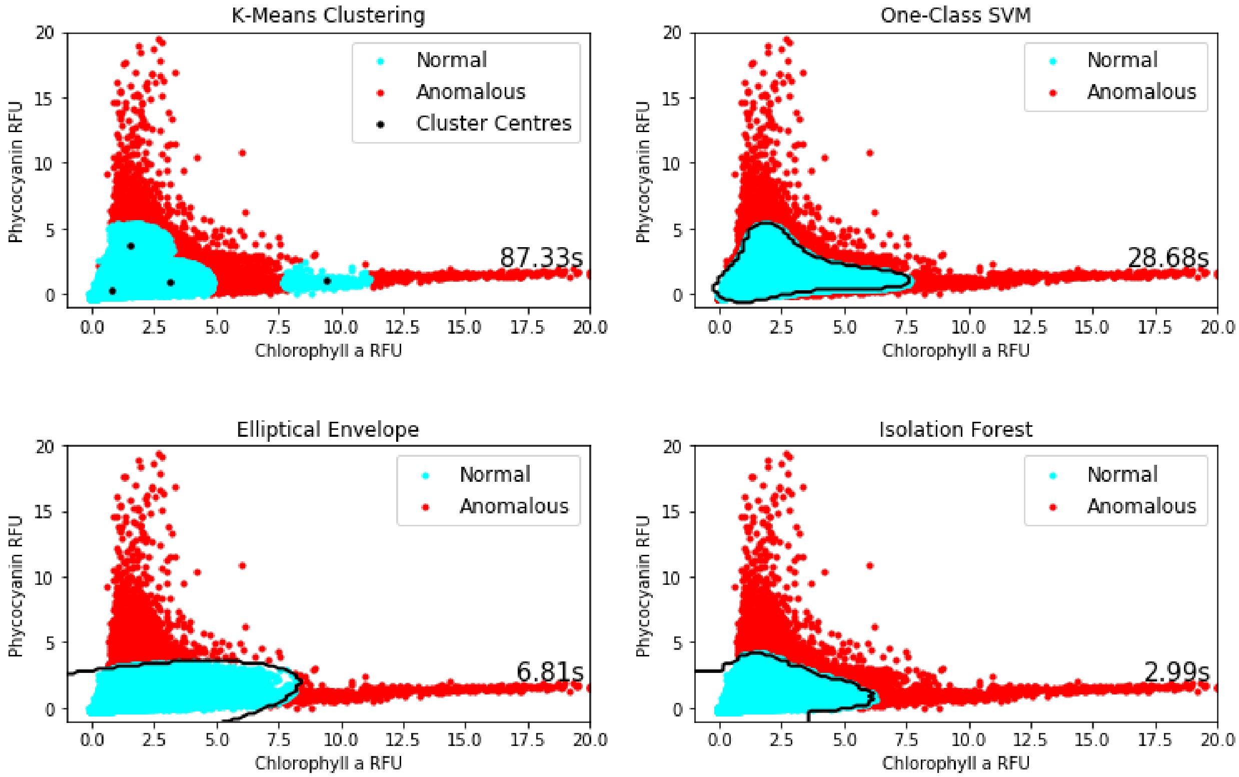


Figure S3. Algorithms trained on the 2015-2019 WE8 dataset.

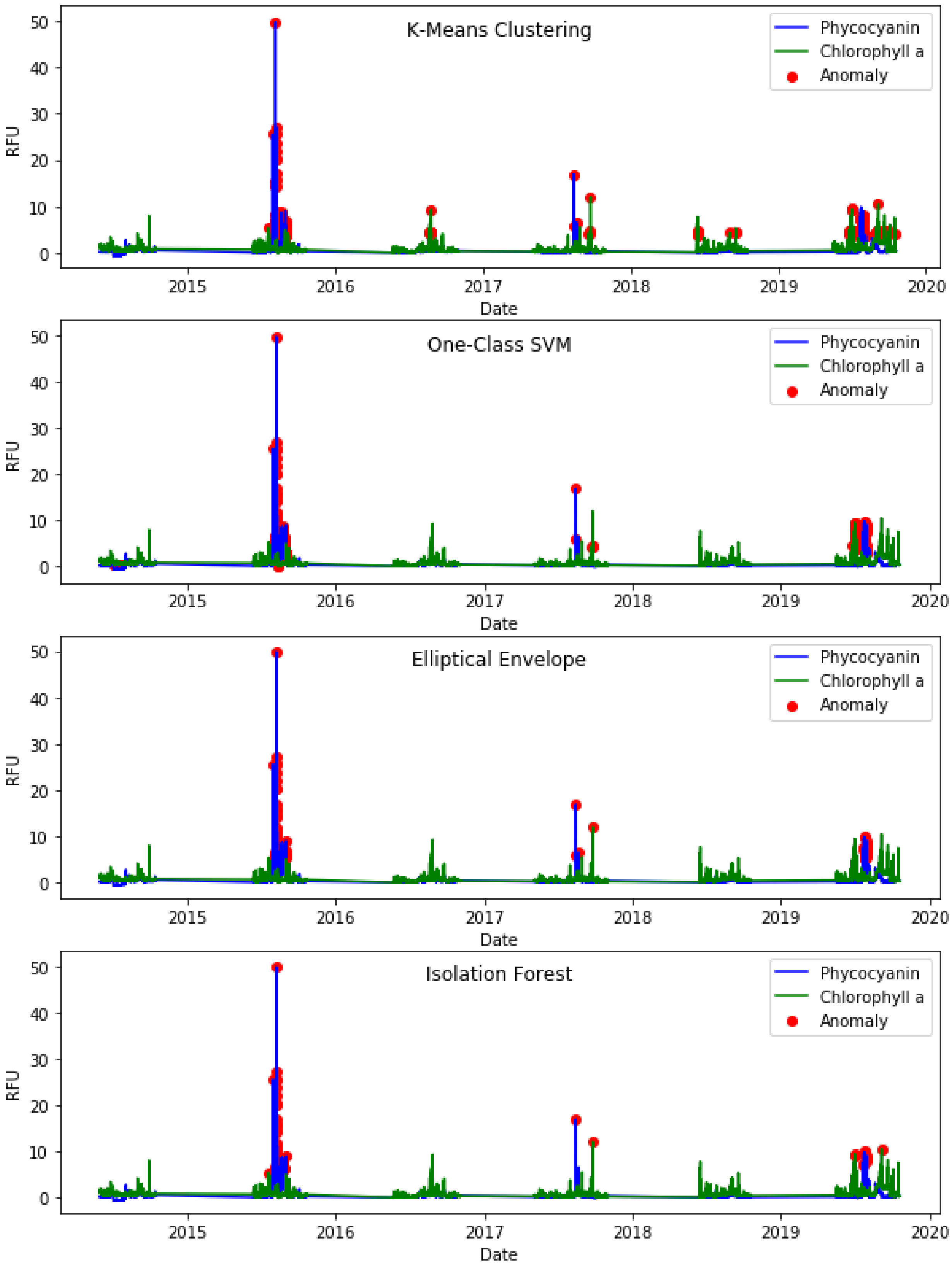


Figure S4. Tested on WE4 2014-2019 with contamination rate of 1%

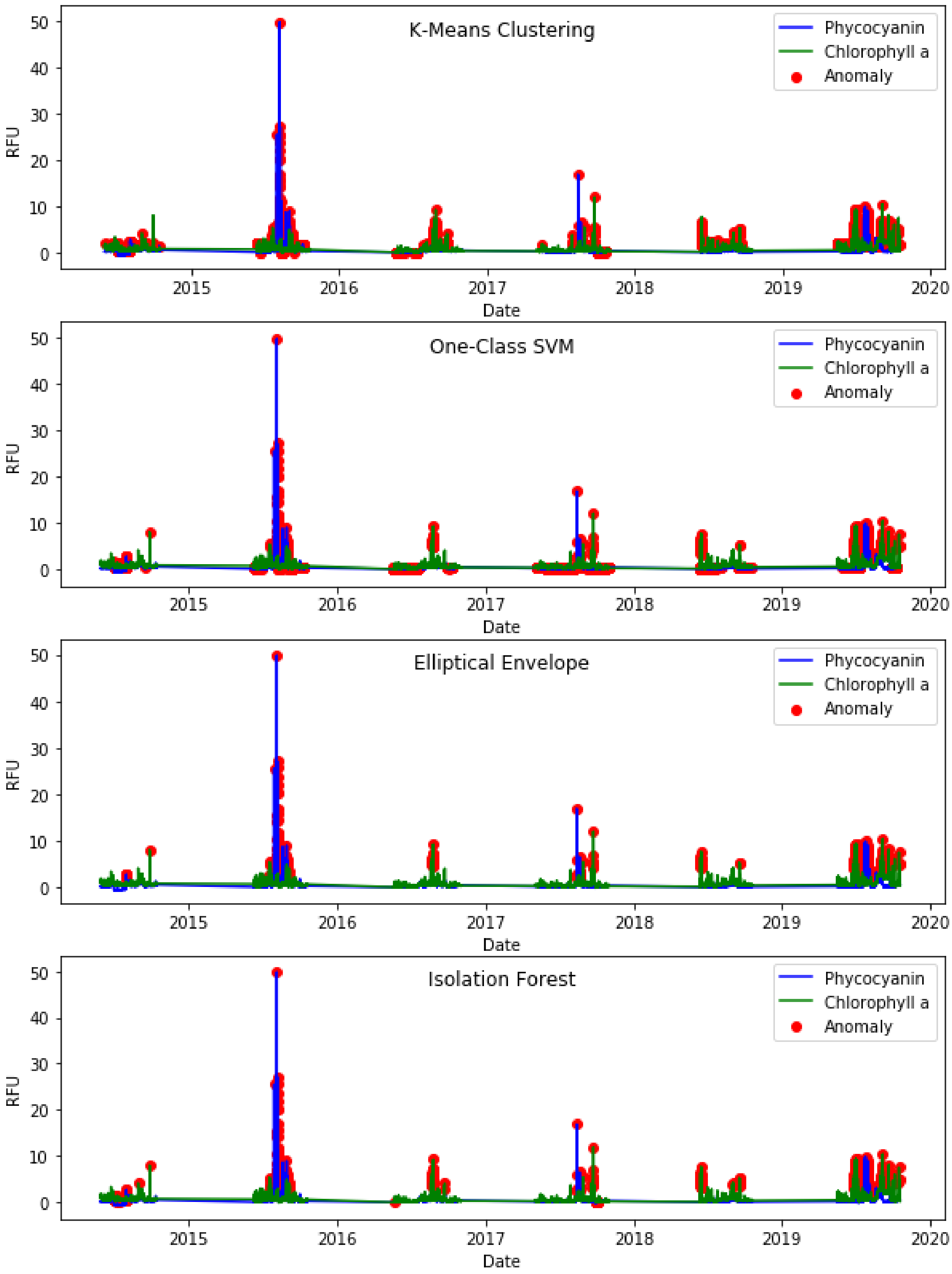


Figure S5. Tested on the WE4 2014-2019 dataset with a contamination rate of 10%

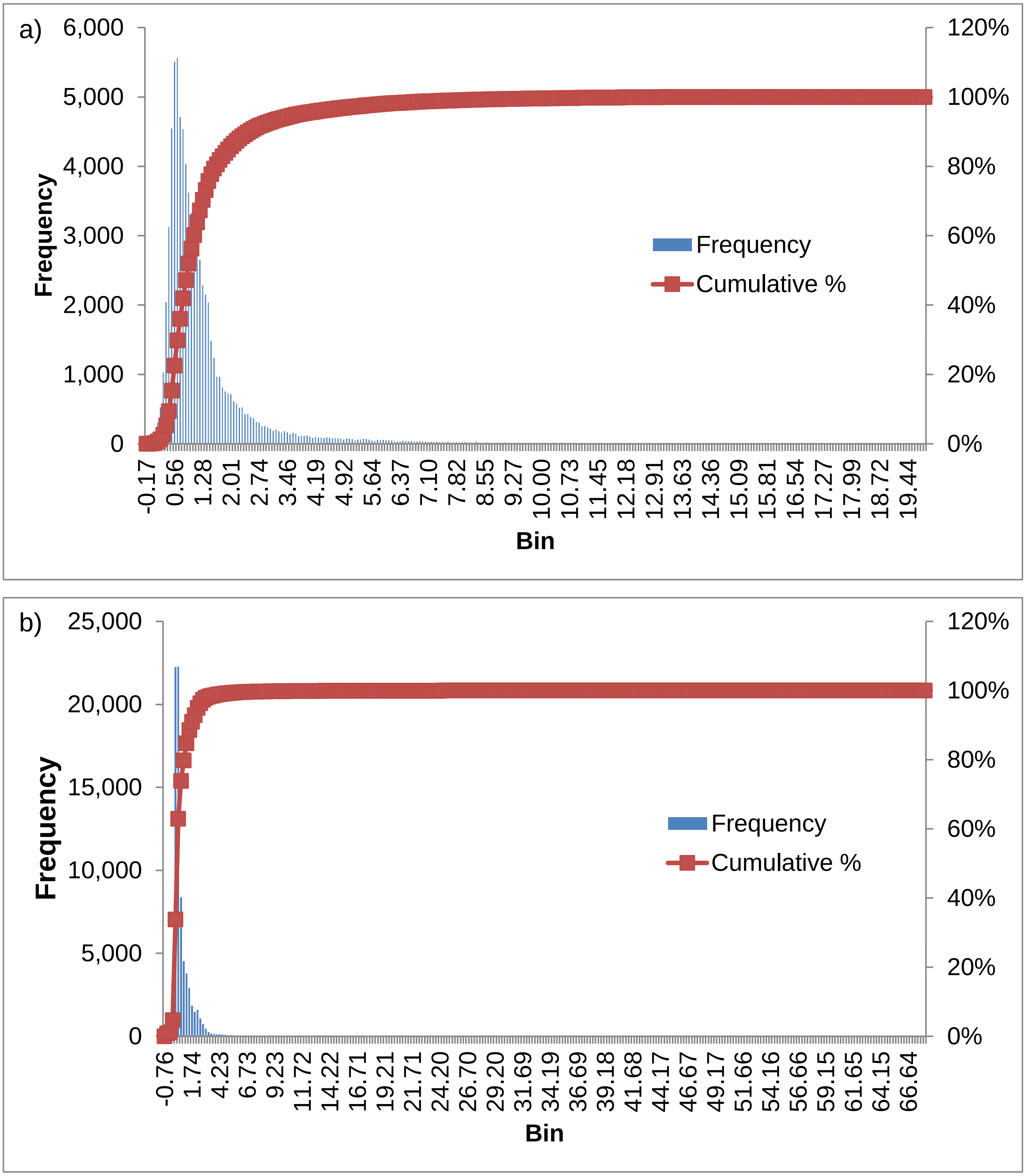


Figure S6. Histograms of the WE2 a) chlorophyll a and b) phycocyanin RFU data from 2014-2019 showing that most of the data is concentrated in the lower range.

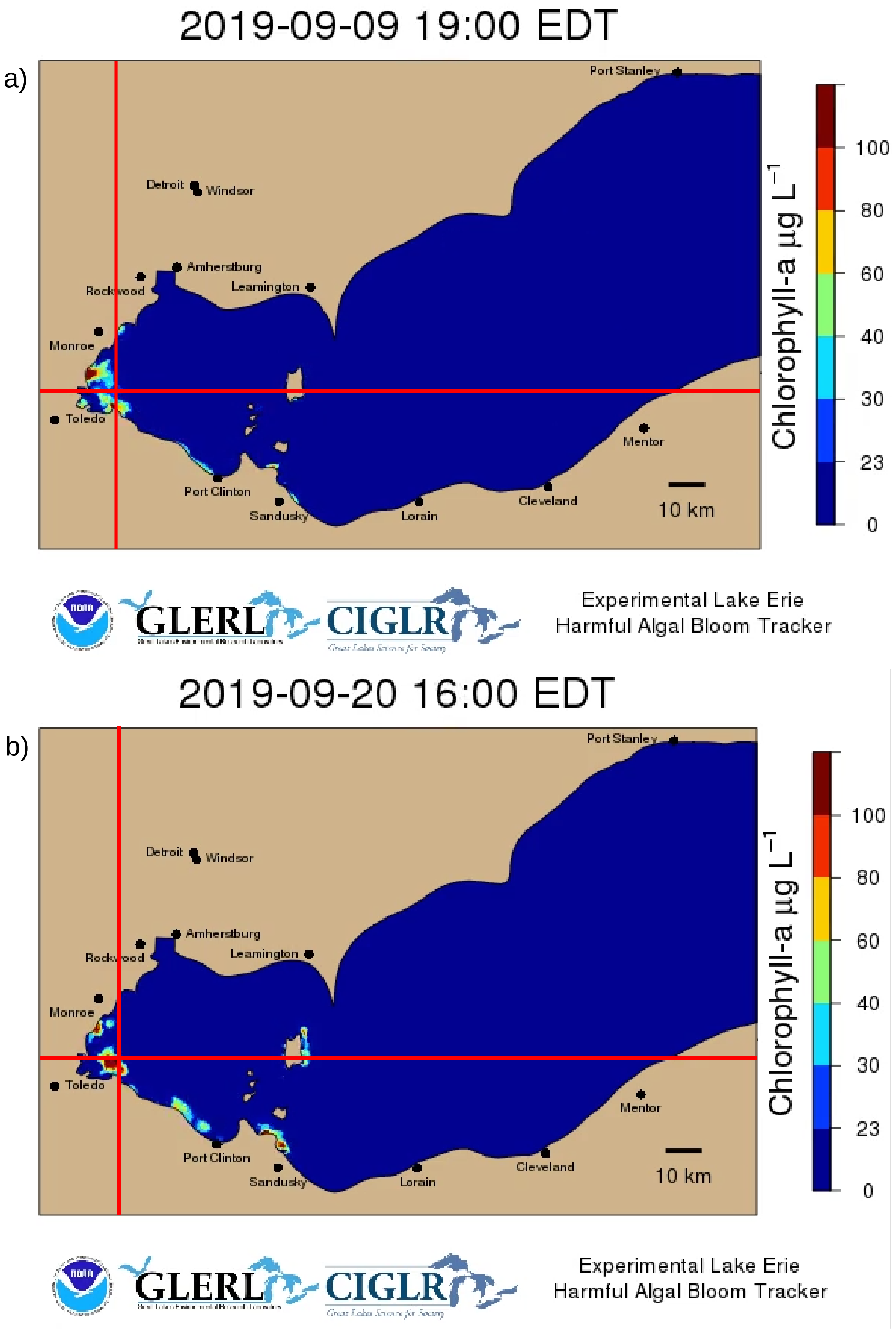


Figure S7. The intersection of the red lines indicates the approximate position of the WE2 buoy. The bloom patches, although present, may not have been measured by the sensors on the buoy from a) September 6 to 10 and b) September 19 to 21, 2019.

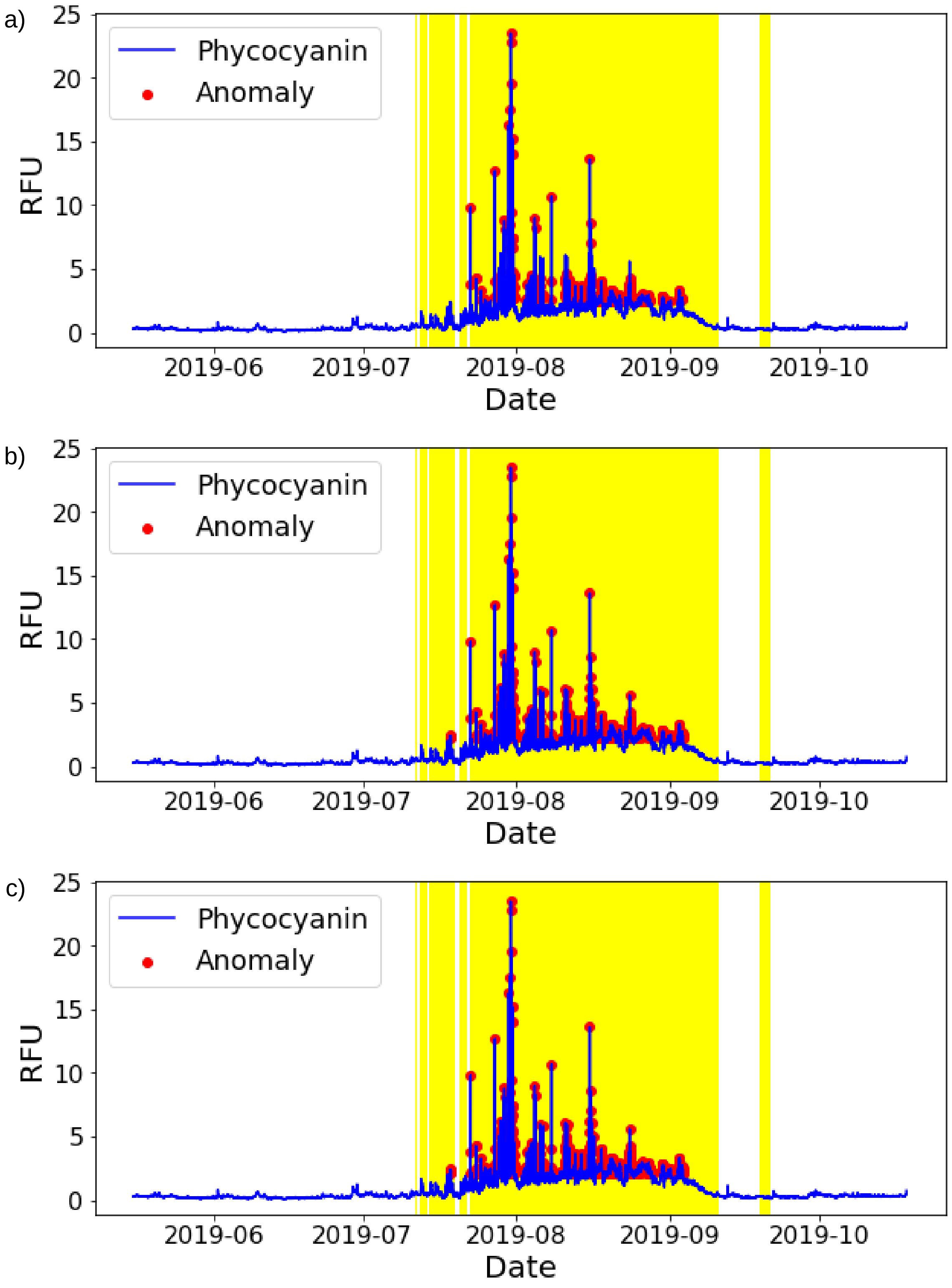


Figure S8. Outlier detection algorithms trained on WE2 2014-2018 data and tested on WE2 2019 data. a) k-means clustering; b) one class support vector machine; c) elliptical envelope.