Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 19 | No. 5 (Suppl) | pp. 1887-1897 | 2020

Original Research Paper

https://doi.org/10.46488/NEPT.2020.v19i05.013

Open Access Journal

A New Index Contributing to an Early Warning System for Cyanobacterial Bloom Occurrence in Atlantic Canada Lakes

K. Hushchyna and T. Nguyen-Quang†
Biofluids and Biosystems Modeling Lab (BBML), Faculty of Agriculture, Dalhousie University, 39 Cox Road, B2N 5E3, Truro-Bible Hill, Nova Scotia, Canada
†Corresponding author: T. Nguyen-Quang; Tri.Nguyen-Quang@Dal.Ca

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 16-03-2020
Revised: 21-04-2020
Accepted: 15-06-2020

Key Words:
Cyanobacterial bloom
Threshold index (TRINDEX)
ROC analysis
Early warning system

ABSTRACT

Cyanobacterial harmful algal blooms (cyanoHAB) have become more frequent and prominent in Atlantic Canada freshwater bodies over the last several years, especially in Nova Scotia (NS). Inspired by the trophic index of Vollenweider, a new index was developed with modification and adaptation for freshwater systems. Our model TRINDEX shows the effectiveness of estimation for the variation of cyanobacterial dominance in phytoplankton communities. TRINDEX can assist in determining the threshold for cyanobacterial bloom onset. Combinations of nutrients and pigments under TRINDEX were tested by a binary discrimination test to find the optimal range of threshold for cyanoHAB formation in freshwater lakes.

INTRODUCTION

Many drinking waters and recreational reservoirs around the globe suffer from cyanobacterial HABs (cyanoHABs) and the used restrictions have caused negative social and economic impacts on local communities and businesses. The desire to protect valuable freshwater and marine resources have motivated extensive research on methods for predicting and mitigating algal blooms. Mitigation strategies for HABs have been divided into two categories, namely precautionary prevention (or early warning protocols) and bloom controls (Kim 2006). Precautionary prevention refers to monitoring and predicting events while bloom control involves both direct controls applied after an HAB has begun and indirect controls deal with strategies, such as management of land-derived nutrient inputs (Kim 2006).

The wish to predict the algal bloom occurrence and proliferation under a complex environmental situation has led to developing many indices for the estimation of eutrophication. There is a real need for improving the knowledge of the eco-physiological mechanisms leading to cyanoHABs; but this cannot be achieved only through reliance on the bulk indices (chlorophyll-a (chl-a), temperature, nutrients).

Most research on eutrophication are based on chemical components, i.e. nutrient characteristic of water bodies such as total phosphorus (TP) and its mineral part (PO_4-P, the dissolved and bioavailable form of phosphate easily consumed by algae) and dissolved inorganic nitrogen (DIN), which contribute significantly to algal growth. All single indices or combined parameters (Novotny & Olem 1994, Bartram et al. 1999, Brient et al. 2008, Brylinsky 2009, Chorus 2012, Ndong et al. 2014, Ahn et al. 2017), despite their usefulness, can show only the real-time conditions and do not predict adequately the cyanobacterial growth in mass and the situation of bloom occurrence as well as the thresholds of bloom onset. They were primarily developed for the determination of trophic status only. There were a few studies which focused on the bloom forecasting using simulations (Anderson et al. 2016). However, almost no work has been done to determine the thresholds that can predict exactly the HAB occurrence in freshwater ecosystems, except for the model from Downing et al. (2001) which provided statistical analysis for predicting the risk of cyanobacterial dominance.

A warning system is an essential tool, from our perspective, which should be able to adequately foresee the irregular patterns such as massive blooms and contribute to water management and decision making. Nowadays, some modern approaches use sophisticated tools and devices including remote sensing, imaging process, etc. to observe and predict blooms. However, two main issues have persisted: (1) They are costly (especially for computational cost) and are used primarily for long-term forecasting (Anderson et al. 2016); (2) The biophysical coupling effects involving bloom occurrences and the distinction between cyanoHAB and other algal blooms were not satisfactorily considered (Kudela et al. 2015, Anderson et al. 2016).

In this paper, a new index is developed to forecast the potential bloom occurrence in freshwater bodies and to assess the freshwater quality relating to the cyanoHAB presence. Our goal is to determine the bloom threshold based on the nutrient level combined with algal pigments and then estimate the appearance of bloom patterns. Specifically, the three following objectives were addressed: (1) To develop a new index, the Threshold Index (or colloquial TRINDEX) for cyanoHAB onset prediction; (2) To validate the TRINDEX using field data from two Nova Scotian lakes, Mattatall (Colchester and Cumberland counties) and Torment (Kings County). Determining the threshold TRINDEX for bloom prediction is assessed by binary discrimination test (Receiver Operating Characteristic (ROC) analyses); (3) To suggest a practical scheme for bloom prediction based on TRINDEX definition with the expectation that our approach can be applied at a larger scale for different trophic waterbodies where blooms could happen.

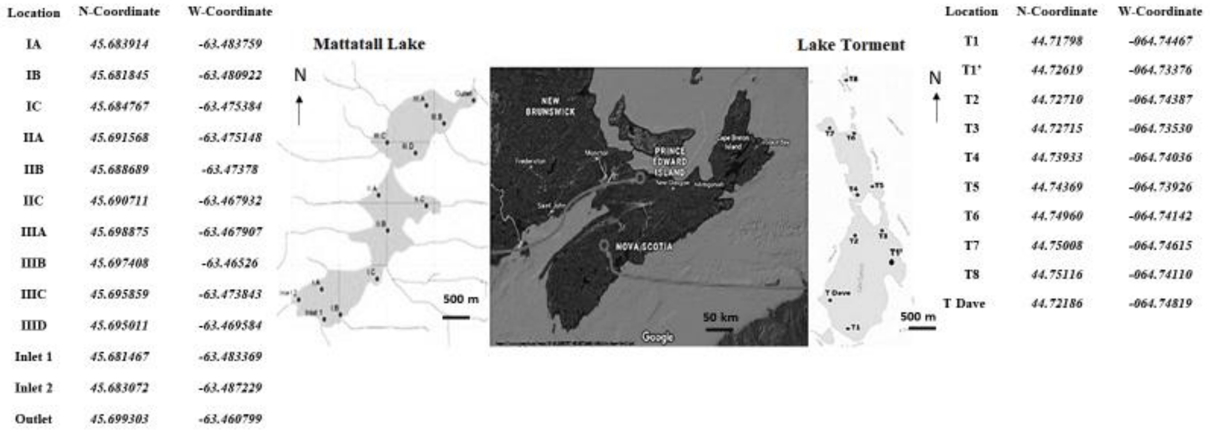
MATERIALS AND METHODS

Study Sites

Two sites in the province of Nova Scotia (NS) Canada, were our targets. Mattatall lake (ML), between Colchester and Cumberland counties, was served as the main site; and lake Torment LT (Kings County) was used for independent verification. The datasets collected from both lakes are independent as LT is located over 200 km from ML in a different geographic area. Both locations are shown in Fig. 1 with their information in Table 1.

ML is mainly spring-fed with some brooks. In terms of human activities, there are blueberry fields and forestry on the west side of the lake. There are approximately 60 residences (both seasonal and year-round) with varying lot sizes and ages. With the data from three years (2015-2017), ML showed a moderately eutrophic level based on chlorophyll-a and TP measurements and contained potentially toxic cyanobacterial species (Hushchyna & Nguyen-Quang 2017). There was a bloom of green algae (*Mougeotia* sp.) in the middle of summer (2015, 2016) following by a cyanobacterial bloom of *Dolichospermum planctonicum* in late summer-autumn.

LT is in East Dalhousie, Kings County. The lake is used for residential and recreational purposes. It covers 261 hectares. There are 250 cottages and homes around the lake. It is surrounded by a forest with no significant agricultural activity on the watershed. The lake is dystrophic with brown water (colour changes 70-145 mg.L^{-1} Pt), low pH (5-6), and high organic content (DOC was 6.5-8.5 mg.L^{-1}) (Marty & Reardon 2016, Nguyen-Quang et al. 2017). The frequency of HAB has increased every year, i.e. since summer 2016. The cyanobacterial blooms were dominated by other cyano-



Location	N-Coordinate	W-Coordinate
IA	45.683914	-63.483759
IB	45.681845	-63.480922
IC	45.684767	-63.475384
IIA	45.691568	-63.475148
IIB	45.688689	-63.47378
IIC	45.690711	-63.467932
IIIA	45.698875	-63.467907
IIIB	45.697408	-63.46526
IIIC	45.695859	-63.473843
IIID	45.695011	-63.469584
Inlet 1	45.681467	-63.483369
Inlet 2	45.683072	-63.487229
Outlet	45.699303	-63.460799

Location	N-Coordinate	W-Coordinate
T1	44.71798	-064.74467
T1'	44.72619	-064.73376
T2	44.72710	-064.74387
T3	44.72715	-064.73530
T4	44.73933	-064.74036
T5	44.74369	-064.73926
T6	44.74960	-064.74142
T7	44.75008	-064.74615
T8	44.75116	-064.74110
T Dave	44.72186	-064.74819

Fig. 1: The location of Mattatall lake (left) and Torment lake (right).

Table 1: Background information about the studied lakes.

Lake	Elevation, m	Surface area, ha	Length, km	Max depth, m
Mattatall	49	120	5	9.3
Torment	172	261	4	6.5

bacteria species (*Dolichospermum flos-aquae*) in LT and appeared randomly from June to November (Nguyen-Quang et al. 2017).

Field Sampling Process and Lab Analysis

Samples were taken bi-weekly or every month, depending on the weather conditions, starting in May through to November, at the surface and bottom levels. Sampling locations are presented in Fig. 1. DO was measured by YSI probe (Professional Plus, Hoskin Scientific LTD, USA). The data on phosphate (PO_4), nitrate (NO_3), chlorophyll-a (chl-a) and phycocyanin (PC) were analysed by our Laboratory.

To determine the concentration of pigments, water samples were filtered through the GF/A Whatman filters. The filters were extracted after that in 90% acetone for chlorophyll-a or in phosphate buffer saline for phycocyanin, then sonicated (50% amplitude for 30 seconds) and centrifuged twice (first centrifugation at room temperature with 3500 *g* for 10 mins; second centrifugation at 4ºC, 13000 *g* for 1.5 hours). The pigment concentrations (chl-a and PC) were measured in $\mu g.L^{-1}$ unit by using the Turner 10AU Fluorometer (Turner Designs, USA) based on the calibration standard curve for both pigments. A dissolved fraction of phosphate and nitrate were measured after filtration through GF-A filters by a photometer using a tablet reagent system (YSI 2010).

Mathematical Formulations

We believe the TRIX concept suggested by Vollenweider et al. (1998) for coastal marine zones is reasonable to be employed for freshwater resources, as it uses the combination of key biological and hydrochemical parameters in a logarithmic relationship without specific characteristics of the marine environment. However, this conception was not well known in freshwater literature. To deal with the non-normal distribution of most of the environmental data, the logarithmic transformation is an appropriate way to 'transform' random data into the normal distribution form. Inspired by the logarithmic transformation of Vollenweider et al. (1998), we suggest herein our Threshold Index (hereafter named TRINDEX) formula as follows.

$$\text{TRINDEX}=\frac{k}{n}\sum_{1}^{i=n}[\frac{(logM_i-logL_i)}{(logU_i-logL_i)}], \qquad \ldots(1)$$

Where,

TRINDEX – Threshold Index to be considered

M_i – measured parameter *i*

L_i – lower limit (concentration) of the considered parameter *i*

U_i – upper limit (concentration) of the considered parameter *i*

k – factor standing for the maximum value of considered range (0,10), so k=10 by default

n – total of parameters M_i we expect to consider

It is our view that chl-a is not a perfect parameter to represent the cyanobacterial bloom detection, because chl-a can be produced by all algal species including microalgae. We propose that the pigment PC, therefore, needs to be introduced into the index as an alternative parameter to reflect the cyanobacterial presence in all phytoplankton communities. There will be hence two scenarios of TRINDEX to be considered by our study.

Scenario 1: Four parameters M_i: PC, D%O, NO_3 and PO_4 will be used (*n = 4*) …(2)

Scenario 2: Five parameters M_i: Chl-a, PC, D%O, NO_3 and PO_4 will be used (*n = 5*) …(3)

The absolute deviation of oxygen from 100% (D%O) shows the main processes of phytoplankton growth which can be used for the detection of bloom onset; nitrogen and phosphorus were chosen in the form of nitrate (NO_3) and phosphate (PO_4) as the main sources of nutrients for cyanobacteria growth. These components can be easily measured daily. The DO fluctuation could be high depending on each period of the day in eutrophic waters. Our observations on various Nova Scotian mesotrophic lakes showed that the period between 8 AM to 2 PM was the optimal time for the development and accumulation of phytoplankton. This period was, therefore, suggested to be used in our monitoring purposes.

The quantity ($logU_i$- $logL_i$) is defined by the difference between upper and lower limits. When these limits are determined, all values being out of this range should be excluded. Therefore, to have an appropriate range to cover different trophic conditions, we used limits of detection (LOD) as the lower limit and maximum value obtained in measurements of the considered variable for the upper limit.

Data Analysis and Discrimination Test for the Thresholds

Definition of onset of blooms: The onset of a bloom can be defined as the start or beginning of any visible signs of blooms, i.e. the first visible appearance of signs or symptoms of some surface scums of a waterbody. However, our definition of bloom onset herein is not only associated with visible signs of algal appearance, but also with scenarios where there are no visible algal signs (but certain amounts of phycocyanin present). Therefore, we suggest that when PC concentration is over $0.03\ mg.L^{-1} \pm 0.002$, these cases can be considered as onset of bloom (PC criteria based on Brient et al. 2008), equivalent to the cell count 20,000 cells per mL of cyanobacteria.

The onset could be a visible bloom or scum situation, but this may not be stable. The surface bloom at onset status can be observed appearing and disappearing unstably in a short period (critical phase) while the supercritical phase of blooms can show a stable situation where blooms or scums can last visibly for long periods (many hours or many days). The onset status can lead to the 'stable blooms' if ambient conditions allow them to develop, or completely vanish, also due to the ambient conditions.

Discrimination Test for Threshold: Receiver Operating Characteristic (ROC) Curve

As in the clinical practice (Carter et al. 2016), a 'yes or no' decision is usually required for 'diseased or non-diseased' situation, herein two states for the bloom: 'yes - bloom occurrence and no - no bloom' are also defined. The bloom threshold T is based on the variable TRINDEX that will drive the outcomes of the decision, as positive (yes - bloom) or negative (no - no bloom) as follows:

$$Decision(TRINDEX) = \begin{cases} + \ \ positive, or\ bloom\ occurence\ if\ TRINDEX > T \\ - \ \ negative, or\ no\ bloom \qquad\quad if\ TRINDEX < T' \end{cases} \quad \ldots(4)$$

If TRINDEX does have some ability to adequately discriminate between positive and negative situations, there are an infinite number of possible decision thresholds. These include three following possibilities (Fig. 2a): (1) threshold T_1, calling all patterns positive with TRINDEX $\geq T_1$, would correctly identify nearly every positive pattern, although a large proportion of negative patterns would inappropriately be called positive; (2) threshold T_2 more of a balance is struck, as both positive and negative events are missed, and finally, (3) T_3 most negative events are correctly identified, but a large proportion of the positive patterns are incorrectly deemed negative.

Four possible outcomes can result for each trial: correctly positive, correctly negative, incorrectly positive and incorrectly negative. At this point, the cut-off area will be introduced as the area which measures the discrimination, i.e. the ability of the TRINDEX test to correctly classify those with, or without the 'disease', as a binary variable. That is equivalent to bloom occurrence (yes) or no bloom (no) respectively.

The ROC analysis is a binary discriminator test which assesses the predictive power of a binary classification system to evaluate a model in a decision-making process and it helps to identify the threshold T. This test is recognized as a useful tool for interpreting medical test results and in many other fields as a method for evaluating the accuracy of analyses (***Lerman et al. 2010***). For more details of ROC curves and related metrics, refer to Brown & Davis (2006).

A curve illustrating the model performance can then be determined by plotting CPF (correct positive fraction or sensitivity) on the vertical axis and (1 - CNF) (CNF is correct negative fraction or specificity) on the horizontal axis (Fig. 2a). The sensitivity is the probability that case X was classified correctly as above the threshold while specificity is the inverse, namely probability that X classified correctly as below the threshold.

The perfect model (Fig. 2b) corresponds to a point in the top left-hand corner of the Y-axis (i.e. CNF = CPF = 1), the top right (CPF = 1, CNF = 0) and bottom left (CPF = 0 and CNF = 1) of the diagram correspond to the extremes of the decision process where every trial is always deemed either positive or negative. A random predictor (CP = IP and CN = IN) gives a straight line CPF = 1 – CNF (X = Y, line of equality or random change). This can be explained

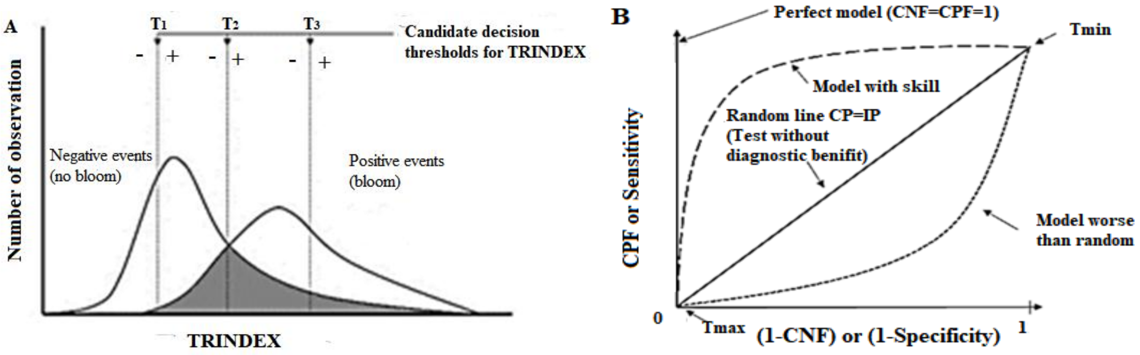


Fig. 2: (a) Illustration of the occurrence of positive and negative observations *versus* TRINDEX on the left (adapted from Brown & David 2006); (b) Binary discrimination skill assessment curves on the right (Adapted from Stow et al. 2009).

by a value of index reaching equalling numbers of true and false positives occur. This value is considered as critical or threshold. The definition of the area under the ROC curve (AUC) was introduced as a criterion to evaluate the overall performance of the discrimination test. This is the percentage of randomly drawn pairs for which this is true. AUC may take values ranging from 0.5 (no discrimination) to 1 (perfect discrimination). A rough practical guide for evaluating the accuracy of a discrimination test with the AUC criteria described as in Table 2 (Carter et al. 2016).

Another factor to estimate the effectiveness of our test is the Youden index *J*. The Youden index *J* (Youden 1950) is defined as:

$$J = \max \{ \text{sensitivity}_c + \text{specificity}_c - 1 \}, \qquad \ldots(5)$$

Where *c* ranges over all possible criterion values.

The Youden index *J,* ranging between 0 and 1, is commonly used to measure overall diagnostic effectiveness (Schisterman et al. 2005). When *J* values are close to 1, it indicates that the effectiveness is relatively good, while values close to 0 indicate limited effectiveness.

We use the dataset from ML (2015-2017) for TRINDEX development and data from LT (2015-2018) to validate our approach. In the following calculations, our parameters *sensitivity* (correct positive fraction) and *specificity* (correct negative fraction) are displayed in the percentage (%) instead of the fraction (see Fig.2b).

In our model, the real sample size of two lakes is different (170 samples of LT compared to 266 ones of ML, greater than the required minimum number 132), hence it is statistically significant.

Our experimental data related to HAB for both lakes (Mattatall and Torment) are not normally distributed. Using log transformation as above mentioned is to convert them into the *'normal distribution'* and TRINDEX can be then processed. The statistical software R combined with Excel and MedCalc is used to carry out all steps.

RESULTS AND DISCUSSION

TRINDEX Calculations and the ROC Curve for Performance of Bloom Prediction

Data used for determining lower and upper limits are given in Table 3.

Based on Table 3, formulas (2) and (3) for TRINDEX will become:

Table 2: The AUC criteria to evaluate the accuracy of the diagnostic test.

AUC value	0.9-1.0	0.8-0.9	0.7-0.8	0.6-0.7	0.5-0.6
Evaluation	excellent (A)	good (B)	fair (C)	poor (D)	fail (F)

Table 3: Limits and ranges - Mattatall lake data.

Limits and ranges	L_i	U_i	$LogL_i$	$LogU_i$	$LogU_i$-$LogL_i$
Phycocyanin, $mg.L^{-1}$	4×10^{-5}	1.855	-4.4	0.3	4.7
Chlorophyll-a, $mg.L^{-1}$	5×10^{-5}	1.692	-4.3	0.2	4.5
Oxygen |100–%O|	0.01	1000	-2	2.0	4.0
Phosphate, $mg.L^{-1}$	0.01	1.52	-2	0.2	2.2
Nitrate, $mg.L^{-1}$	0.01	1.72	-2	0.2	2.2

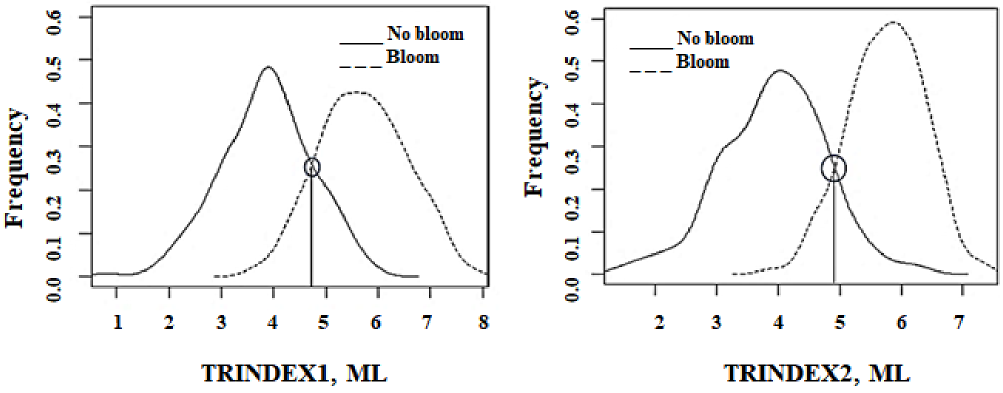


Fig. 3: Distribution of no bloom and bloom cases for TRINDEX1 and TRINDEX2, Mattatall lake.

$$TRINDEX1 = 2.5 \times [\frac{logPC+4.4}{4.7} + \frac{logD\%O+2}{4.0} + \frac{logPO4+2}{2.2} + \frac{logNO3+2}{2.2}], \quad \ldots(6)$$

$$TRINDEX2 = 2 \times [\frac{logPC+4.4}{4.7} + \frac{logChl-a+ 4.3}{4.5} + \frac{logD\%O+2}{4.0} + \frac{logPO4+2}{2.2} + \frac{logNO3+2}{2.2}], \quad \ldots(7)$$

Data were divided into two groups: (i) bloom occurrence and (ii) no bloom. The distinct scenario for both bloom and no bloom conditions for TRINDEX1 and TRINDEX2 using *rnorm* in R software is graphically represented in Fig. 3.

The cut-off point 5.0 was estimated from Fig. 3. However, this cut-off point should be validated by field observations via ROC curve analysis to precisely determine the threshold value for cyanobacterial bloom. This discrimination test was processed with field observation data (Fig. 4 and Table 5).

Sensitivity (true positive cases) was calculated by assuming that every TRINDEX value can lead to bloom. Inversely, specificity of false positive was done by assuming that every TRINDEX cannot lead to blooms. All calculations of TRINDEX were rounded at 0.2 unit. Formulas for false positive and false negative are as follows.

True positive = Sensitivity = ∑Number of TRINDEX with bloom/ Total of bloom case, …(8a)

False positive = (100 - Specificity) = ∑Number of no bloom TRINDEX /Total of no bloom cases, …(8b)

The dataset of 266 values of TRINDEX1 during 2015-2017 was used for TRINDEX in ML, among them 74 cases with bloom and 192 cases without bloom. Single cases of bloom were detected when TRINDEX1 started from value 4 (Table 5). The higher TRINDEX1 (greater than 5.0), more frequent bloom cases were recorded than no bloom cases; and maximum bloom cases (14 cases) happened when TRINDEX1 = 6.2. Therefore, it can be said that the TRINDEX1 range from 4.0 to 5.0 is the marginal situation, where there is likely to be no sign of a visible bloom but just small disturbances of the environmental conditions (leading to a higher TRINDEX1) could trigger the cyanobacterial bloom.

There were 249 calculated values of TRINDEX2 (Table 5) with 75 bloom cases and 174 no bloom cases. The lowest TRINDEX2 showing bloom was 4.4, but when TRINDEX2 = 5.2 the number of bloom cases was more prevalent than no bloom cases. The maximum number of no bloom cases was noticed when TRINDEX2 = 4.0 and the maximum bloom cases were when TRINDEX2 = 6.2. From the above analyses, the proposition of a transition range for TRINDEX2 was from 4.4 to 5.2 and the suggested threshold value for bloom occurrence suggested was 5.2.

From ROC curves (Fig. 4), the appropriate threshold for bloom onset can be chosen. It should have the maximum sensitivity and at the same time the minimum false positive cases. As two axes of our ROC curve are determined by the *sensitivity* 100% (the probability of true positive results) and (*100% - specificity*) (the probability of false positive results). As such, the false positive cases show TRINDEX are high but no blooms are occurring, while the false negative ones show the opposite scenario: TRINDEX are low but blooms are observed.

ROC curve for TRINDEX1 has the best combination of high sensitivity (81%) and low false positive (12%) (Fig. 4, left side), equivalent to the point 5.0 in Table 5. So, all results of TRINDEX1 equal or greater than 5.0 must be resulting in cyanobacterial blooms. TRINDEX2 (Fig.4 right) has the best combination of high sensitivity (83%) and low false positive (6%), equivalent to 5.2 in Table 5.

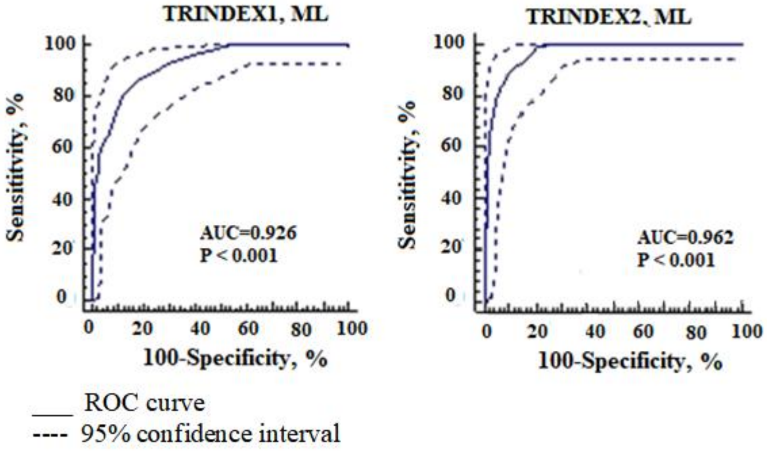


Fig. 4: ROC curves for TRINDEX1 and TRINDEX2 in Mattatall lake.

The significant level represented by p-value stands for the probability that the observed sample AUC (area under the curve) is found when the true (population) AUC is 0.5. When p is small ($p < 0.05$) then it can be concluded that AUC differs significantly from 0.5. Carter et al. (2016) have mentioned that a ROC curve test has (at least) some discriminatory power if the 95% confidence interval of AUC does not include 0.50. In our case of ML, the AUC is 0.926 (95% CI: 0.887 to 0.954; $p < 0.0001$) for TRINDEX1 of ML and AUC is 0.961 (95% CI: 0.929 to 0.981; $p < 0.0001$) for TRINDEX2. This confirms the good fit of our threshold 5.0 for ML as the AUC = 0.926 and 0.961, the discrimination test was then excellent (Table 2).

An AUC over 0.9 (0.926 and 0.961 for TRINDEX1 and TRINDEX2, respectively) implies that in a hypothetical experiment in which we randomly select pairs of positive cases (no bloom) a false negative result is deemed comparable to that of a false positive result. With the environmental factors that can affect a lake system, the random excitation can cause a change of stability around the equilibrium point and beyond this equilibrium point, blooms occur, i.e. instability will cause the HAB.

The range of values of TRINDEX1 from 4.0 to 5.0 can be classified as the transition phase, i.e. potential for a bloom occurrence in the near future. Considering TRINDEX as 'predictor' for bloom, the Youden index *J* is significant in our tests: 0.69 for TRINDEX1 and 0.78 for TRINDEX 2 (Table 4c). Hence, it is concluded that for ML, two following cut-off points are considered as thresholds: 5.0 for TRINDEX1 while 5.2 for TRINDEX2 with a goodness of fit of discrimination test.

Independent Verification by Lake Torment Data

The same procedure was followed by using data from lake Torment (LT) and cut-off point was found approximately 4.6 for TRINDEX1 and TRINDEX2 (Fig.5). Fig. 6 shows the ROC curve analyses for LT.

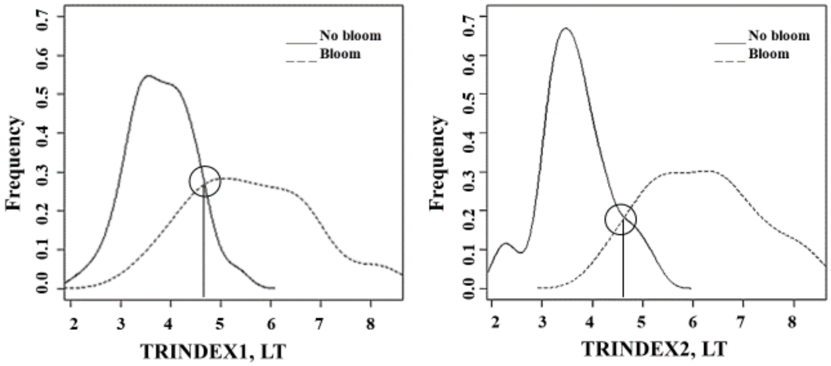


Fig. 5: Distribution of no bloom and bloom cases for TRINDEX1 (Left) and TRINDEX2 (Right), lake Torment.

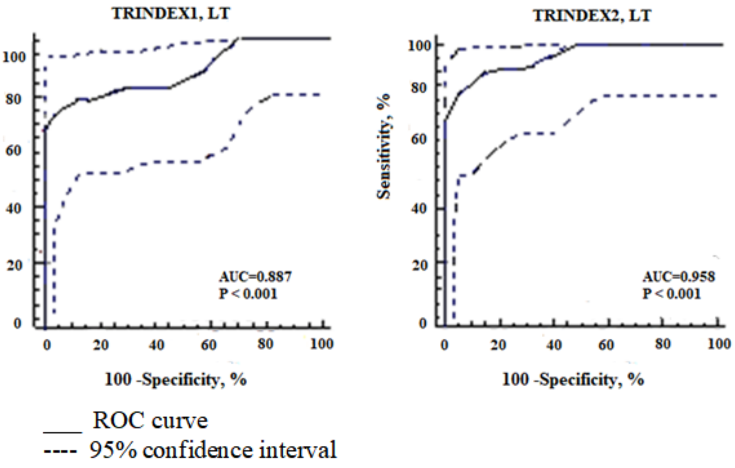


Fig. 6: ROC curves for TRINDEX1 and TRINDEX2 in lake Torment.

TRINDEX1 (Fig. 6 left) had the best combination of high sensitivity (79%) and low false positive (6%) at value 4.8, while TRINDEX2 (Fig. 6 right) had the best combination of high sensitivity (79%) and false positive (2%), at value 5.2.

The transition phase of TRINDEX1 for LT data was 3.4-4.8 and the cut-off point was 4.8. The AUC = 0.887 confirmed the discrimination test was excellent (95%CI: 0.830 to 0.930; $p < 0.0001$). For TRINDEX2, the transition range is 4.0-5.2 and cut-off point at 5.2, and AUC = 0.956 also showing that the discrimination test is excellent (95%CI: 0.914 to 0.982; $p < 0.0001$). Youden index J is also significant: 0.73 for TRINDEX1 and 0.78 for TRINDEX2.

Tables 4 a,b,c show comparison between LT and ML data in term of threshold and ROC curve analyses.

As TRINDEX2 combines both pigments PC and chl-a, it seems inaccurate for the prediction of cyanobacterial bloom thresholds due to the increase of chl-a by other phytoplankton rather than just cyanobacteria, hence increasing TRINDEX2 above the real threshold. Therefore, TRINDEX1 based only on PC seems the better indicator to estimate the threshold for cyanobacterial bloom than TRINDEX2. The lowest value of TRINDEX1 when blooms appear in the 2 lakes was chosen for the transition phase and the cut-off point is the threshold for bloom onset. TRINDEX1 thresholds for the prediction of cyanobacterial blooms can be defined:

- TRINDEX1 < 3.4: no bloom happens as the system is stable.
- TRINDEX1 is between 3.4 and 4.8: there will be a high risk of cyanobacterial bloom development; this range is called ‘transition phase’. Other environmental components such as weather conditions should be triggering factors to predict bloom development. In this transitional phase, the situation tends to the onset tendency, that means blooms are happening but can be unstable (appearing and then disappearing in a short period) or becoming stable, depending on ambient conditions.
- TRINDEX1 > 4.8: cyanobacterial blooms could happen and become stable during a certain period (hours or even days).

The performance of our model was evaluated via Accuracy, Precision, Recall and F1 score metrics. Precisely, among these 40 observations, we have 4 false positive cases (10%); 22 true positive (TP) cases; 1 false negative (FN) case (5%); and 13 true negative (TN) cases. It is important to note when we have a large number of true negative cases, it can influence the accuracy of our predictions.

Table 4a: Statistical analyses of TRINDEX1 and 2 for both lakes (by using MedCalc).

	TRINDEX1		TRINDEX2	
	Mattatall	Torment	Mattatall	Torment
Sample size	266	170	249	170
Positive group[a]	74 (27.82%)	24 (14.12%)	74 (29.72%)	24 (14.12%)
Negative group[b]	192 (72.18%)	146 (85.88%)	175 (70.28%)	146 (85.88%)

[a]results = 1 (having bloom); [b]results = 0 (no bloom)

Table 4b: Area under the ROC curve (AUC) for two lake data.

	TRINDEX1		TRINDEX2	
	Mattatall	Torment	Mattatall	Torment
Area under the ROC curve	0.926	0.887	0.961	0.956
Standard Error[a]	0.0162	0.0465	0.0106	0.0226
95% confidence interval[b]	0.887 to 0.954	0.830 to 0.930	0.929 to 0.981	0.914 to 0.982
z statistic	26.237	8.324	43.505	20.172
Significance level p (Area=0.5)	< 0.0001	<0.0001	< 0.0001	<0.0001

[a]DeLong et al., 1988 (the method recommended by MedCalc to calculate standard error and CI95%); [b]Binomial exact

Table 4c: Youden index for data from two lakes.

	TRINDEX1		TRINDEX2	
	Mattatall	Torment	Mattatall	Torment
Youden index J	0.69	0.73	0.78	0.78

Table 5: Data for ROC curve of Mattatall lake.

TRINDEX	TRINDEX1					TRINDEX2				
	Bloom	No bloom	Total	Sensitivity	(100-Specificity)	Bloom	No bloom	Total	Sensitivity	(100-Specificity)
1.60		1	1	100	100					
1.80		1	1	100	99		1	1	100	100
2.00		5	5	100	99		1	1	100	99
2.20		3	3	100	96		3	3	100	99
2.40		3	3	100	95		4	4	100	97
2.60		4	4	100	93		4	4	100	95
2.80		12	12	100	91		4	4	100	93
3.00		12	12	100	85		18	18	100	90
3.20		5	5	100	79		6	6	100	80
3.40		6	6	100	76		10	10	100	76
3.60		17	17	100	73		18	18	100	71
3.80		21	21	100	64		17	17	100	60
4.00	2	21	23	100	53		26	26	100	51
4.20	3	24	27	97	42		13	13	100	36
4.40	3	13	16	93	30	1	13	14	100	28
4.60	2	10	12	89	23	5	11	16	99	21
4.80	4	11	15	87	18	2	7	9	92	14
5.00	5	6	11	81	12	5	7	12	89	10
5.20	7	4	11	75	9	4	3	7	83	6
5.40	5	7	12	65	7	7	3	10	77	5
5.60	8	2	10	59	3	10	2	12	68	3
5.80	3	3	6	48	2	8	1	9	55	2
6.00	4		4	44	1	5		5	44	1
6.20	14		14	39	1	15	1	16	37	1
6.40	5	1	6	20	1	7	1	8	17	1
6.60	6		6	13	0	2		2	8	0
6.80						1		1	5	0
7.00	2		2	5	0	1		1	4	0
7.20	1		1	3	0	1		1	3	0
7.40						1		1	1	0
Total	74	192	266			75	174	249		

In summary, the TRINDEX1 model which is applied to the real observation data from LT for 3 summers is 87.5% accurate, with a recall 0.95 (which is excellent as far above 0.5), a precision of 84.6%, and a F_1 score near 0.9 (which is also very good as F_1 defined in the range from 0 (bad test) to 1 (excellent test)).

Practical Scheme for Application

As indicated, TRINDEX1 is a more appropriate indicator to predict the cyanobacterial blooms. The transition phase can point out the need and significance of a frequent monitoring program for the waterbody. For the possibility of bloom occurrence: the closer TRINDEX1 to threshold values, the higher probability of cyanobacterial growth. The scheme in Fig. 7 is suggested as a practical tool for bloom onset prediction and management based on TRINDEX1.

From this scheme, three scenarios of risk could lead to the management decision for waterbody dealing with algal bloom issue: (1) When no bloom is observed and TRINDEX1<3.4, the monitoring plan for waterbody should follow its established routine; (2) When TRINDEX1 of the lake goes between 3.4-4.8, the risk for a cyanoHAB growth increases. In this case, there might not have any visible sign of bloom in a waterbody, but a more frequent sampling plan with all nutrient parameters, plus taxonomy and toxin analyses should be initiated. Also, the early warning signs could be placed in all accesses to the lake to inform residents about the algal growth concerns. (3) If TRINDEX1 is calculated greater than 4.8, the risk of blooming issues is high and stable blooms could be either observed (on the surface) or not (blooms dissipate in the water column). In this last scenario, any activities of people and pets must be restricted in the

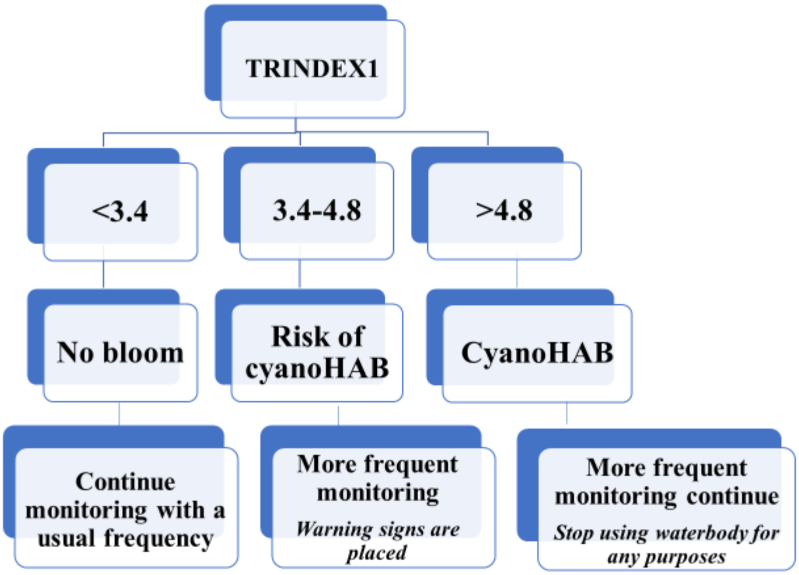

TRINDEX1
<3.4
3.4-4.8
>4.8
No bloom
Risk of cyanoHAB
CyanoHAB
Continue monitoring with a usual frequency
More frequent monitoring
Warning signs are placed
More frequent monitoring continue
Stop using waterbody for any purposes

Fig. 7: Practical scheme of TRINDEX application for bloom onset prediction.

concerned waterbody. The monitoring plan should be more intensive during the bloom episodes.

TRINDEX1 is our recommendation for the bloom prediction indicator for monitoring purposes.

Advantages and Limitations of TRINDEX

A "one-size-fits-all" approach (a term used by Anderson et al. 2015) for HAB modelling is not practical and even utopic. Whether forecasting the potentially harmful bloom occurrence or tracking its path, models should always be linked to the local chemistry, physics, and biology of the waterbody and based on the in-situ data. Alert systems and mitigation strategies will be dictated by the history of human resource use in the region and will hinge on local to federal government mandates for protecting those resources (Anderson et al. 2015). From this perspective, we would underline here the advantages and limitations that could lead to conceive TRINDEX for monitoring purposes in each waterbody.

(1) TRINDEX, especially TRINDEX1, is suggested as an indicator for the cyanobacterial bloom onset. This can tell about cyanobacteria presence and bloom occurrence in the waterbody.

(2) The range called 'transition phase' should be understood as a potential risk, or 'a situation involving exposure to bloom' possibility, i.e. that can lead to (i) stable bloom situation (supercritical); or (ii) unstable blooming immediately (critical); or (iii) nothing happening (subcritical), depending on many other factors (such as light, wind, temperature etc.). The transition phase must be considered as an important step of the bloom onset and need to be carefully studied due to its high sensitivity for both false positive and false negative cases.

(3) Temperature was not yet considered in our TRINDEX model herein, because different potentially toxic cyanobacterial species grow with various temperature ranges. We suggest our TRINDEX1 could be used when the temperature is greater than 15^{o}C (which is the lowest temperature range observed in Atlantic Canada for cyanobacterial blooms development).

(4) The inaccuracy of the model may be caused by community metabolism, which was not considered yet. For the potential users for other lakes (oligotrophic to medium eutrophic), our TRINDEX thresholds suggested herein can be appropriately applicable; while for high eutrophic ones, we should recommend that the users would define their own upper and lower limits and would go through all necessary steps to adjust correctly their own lake thresholds.

(5) Also, the dominant species generating blooms can be a significant factor that could intervene in the accuracy of TRINDEX. Both lakes in our consideration (Mattatall and Torment) have the same genus *Dolichospermum*. Further investigation needs to be undertaken with other waterbodies containing different species generating blooms and community metabolism.

CONCLUSIONS

The prediction of cyanobacterial bloom occurrence has always been a challenging subject in both marine and freshwater environments for many decades, and the emphasis on the determination of thresholds for bloom onset, especially in the freshwater ecosystem was not strong. An ideal alert system should quantitatively predict cyanoHAB likelihood, intensity, and potential blooming. The number of approaches for monitoring, detecting, predicting, and forecasting the onset, fate, and demise of algal blooms is arguably comparable to the diversity of species being studied.

Here we focus our work on the prediction of cyanoHABs using index capable to show the thresholds determining the transitional phase to blooming aspects, above which, cyanobacterial blooms in freshwater bodies could happen. All our efforts rely on a close relationship between observations and a simple model leading to developing a forecasting capability. TRINDEX could be practically developed and used in the potential application of smart systems for water management.

ACKNOWLEDGEMENTS

TNQ acknowledges the NSERC via Discovery Grant NSERC RGPIN 03796, CFI and Nova Scotia Innovation Trust via Grant No 31188 for equipment. We acknowledge the ML and LT residents for their collaboration. We thank Prof. Raymond Côté (Dalhousie University) and Dr. Daniel Beach (National Research Council of Canada) who carefully reviewed and greatly improved the quality of this contribution. We also acknowledge BBML students (Nadeem, TD, Kayla) for their contributions.

REFERENCES

Ahn, C.Y., Joung, S.H., Yoon, S.K. and Oh, H.M. 2007. Alternative alert system for cyanobacterial bloom, using phycocyanin as a level determinant. Journal of Microbiology, 45(2): 98-104.

Anderson, C.R., Kudela, R.M., Kahru, M., Chao, Y., Rosenfeld, L.K., Bahr, F.L., Anderson, D.M.T. and Norris, A. 2016. Initial skill assessment of the California Harmful Algae Risk Mapping (C-HARM) system. Harmful Algae, 59: 1-18.

Anderson, C.R., Moore, K.S., Tomlinson, M.C., Silke, J. and Cusack, C.K. 2015. Living with harmful algal blooms in a changing world: strategies for modeling and mitigating their effects in coastal marine ecosystems. In: Shroder, J. F., Ellis, J.T. and Sherman, D. J. (eds.) Coastal and Marine Hazards, Risks, and Disasters. Chapter 17.

Bartram, J., Burch, M., Falconer, R., Jones, G. and Kuiper-Goodman, T. 1999. Situation assessment, planning and management. In: Chorus, I. and Bartram, J. (eds) Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management, WHO.

Brient, L., Lengronne, M., Bertrand, E., Rolland, D., Sipel, A., Steinmann, D., Baudin, I., Legeas, M., Le Rouzic, B. and Bormans, M. 2008. A phycocyanin probe as a tool for monitoring cyanobacteria in freshwater bodies. Journal of Environmental Monitoring, 10: 248-255.

Brown, C.D. and Davis, H.T. 2006. Receiving operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems, 80: 24-38.

Brylinsky, M. 2009. Lake Utopia Water Quality Assessment. Final report, Dept. Environment of New Brunswick, Canada, pp.92.

Carter, J.V., Pan, J., Rai, S.N. and Galandiuk, S. 2016. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery, 59(6): 1638-1645.

Chorus, I. 2012. Current approaches to cyanotoxin risk assessment, risk management and regulations in different countries. Federal Environ Agency, Germany, pp.151.

Downing, J.A., Watson, S.B. and McCauley, E. 2001. Predicting cyanobacteria dominance in lakes. Canadian Journal of Fisheries and Aquatic Sciences, 58(10): 1905-1908.

Hushchyna, K. and Nguyen-Quang, T. 2017. Using the modified Redfield ratio to estimate harmful algae blooms. Environmental Problems, 2(2): 101-108.

Kim, H.G. 2006. Mitigation and controls of HABs. In: Granéli, E. and Turner, J.T. (eds) Ecology of Harmful Algae, Springer.

Kudela, R.M., Palacios, S.L., Austerberry, D.C., Accorsi, E.K., Guild, L.S. and Torres-Perez, J. 2015. Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters. Remote Sensing of Environment, 167: 196-205.

Lerman, D.C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., Mahmood, A., Strobel, M., Mullen, S., Keyl, A. et al. 2010. Applying signal-detection theory to the study of observer accuracy and bias in behavioral assessment. Journal of Applied Behavior Analysis, 43(2): 195-213.

Marty, J. and Reardon, M. 2016. King county lake monitoring program 2015 season. King County Lake Monitoring Program 2016 Season. Report. Municipality of the County of Kings.

Ndong, M., Bird, D., Nguyen-Quang, T., Boutray, M., Zamyadi, A.,Vincon-Leite, B., Lemaire, J.B., Prevost, M. and Dorner, S. 2014. Estimating the risk of cyanobacterial occurrence using an index integrating meteorological factors: Application to drinking water production. Water Research, 56: 98-108.

Nguyen-Quang, T., McLellan, K., Hushchyna, K. and Murdock, A. 2017. Harmful Algal Bloom (HAB) Monitoring for Lake Torment and Armstrong Lake. A systematic investigation in 2016-2017 versus 2014 Kings County water quality results. Technical report.

Novotny, V. and Olem, H. 1994. Water quality: prevention, identification, and management of diffuse pollution. Technology & Engineering, Wiley, pp. 1072.

Schisterman, E.F., Perkins, N.J., Liu, A. and Bondell, H. 2005. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology, 16(1): 73-81.

Stow, C.A., Jolliff, J., Mc Gillicuddy, D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A.M., Rose, K.A. and Wallhead, P. 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, 76: 4-15.

Vollenweider, R.A., Giovanardi, F., Montanari, G. and Rinaldi, A. 1998. Characterization of the trophic conditions of marine coastal waters, with special reference to the NW Adriatic Sea / Proposal for a trophic scale, turbidity and generalized water quality index. Environmetrics, 9: 329-357.

Youden, W.J. 1950. An index for rating diagnostic tests. Cancer, 3: 32-35.

YSI 9300 and 9500 direct-read photometer. 2010. User manual. YSI Incorporated.