AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR035540

Predicting Imminent Cyanobacterial Blooms in Lakes Using Incomplete Timely Data

Christopher M. Heggerud[1,2], Jingjing Xu[1,3], Hao Wang[1], Mark A. Lewis[1,3], Ron W. Zurawell[4], Charlie J. G. Loewen[5], Rolf D. Vinebrooke[3], and Pouria Ramazi[6]

[1]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada, [2]Department of Environmental Science and Policy, University of California, Davis, Davis, CA, USA, [3]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada, [4]Alberta Environment and Protected Areas, Edmonton, AB, Canada, [5]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA, [6]Department of Mathematics and Statistics, Brock University, St. Catharines, ON, Canada

Key Points:

- Toxic algae blooms pose threats globally, longing for prompt prediction
- Our work uses less costly, incomplete data sets to make near-future bloom predictions
- We predict blooms with area under the curve 0.83 using partial data of regional and local information

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P. Ramazi and H. Wang,
pramazi@brocku.ca;
hao8@ualberta.ca

Citation:

Heggerud, C. M., Xu, J., Wang, H., Lewis, M. A., Zurawell, R. W., Loewen, C. J. G., et al. (2024). Predicting imminent cyanobacterial blooms in lakes using incomplete timely data. *Water Resources Research*, *60*, e2023WR035540. https://doi.org/10.1029/2023WR035540

Received 16 JUN 2023
Accepted 5 JAN 2024

Author Contributions:

Conceptualization: Hao Wang, Mark A. Lewis, Pouria Ramazi
Data curation: Christopher M. Heggerud, Jingjing Xu, Hao Wang, Ron W. Zurawell, Charlie J. G. Loewen, Pouria Ramazi
Formal analysis: Christopher M. Heggerud, Jingjing Xu, Pouria Ramazi
Funding acquisition: Hao Wang
Investigation: Christopher M. Heggerud
Methodology: Jingjing Xu, Hao Wang, Pouria Ramazi
Project Administration: Hao Wang

Abstract Toxic cyanobacterial blooms (CBs) are becoming more frequent globally, posing a threat to freshwater ecosystems. While making long-range forecasts is overly challenging, predicting imminent CBs is possible from precise monitoring data of the underlying covariates. It is, however, infeasibly costly to conduct precise monitoring on a large scale, leaving most lakes unmonitored or only partially monitored. The challenge is hence to build a predictive model that can use the incomplete, partially-monitored data to make near-future CB predictions. By using 30 years of monitoring data for 78 water bodies in Alberta, Canada, combined with data of watershed characteristics (including natural land cover and anthropogenic land use) and meteorological conditions, we train a Bayesian network that predicts future 2-week CB with an area under the curve (AUC) of 0.83. The only monitoring data that the model needs to reach this level of accuracy are whether the cell count and Secchi depth are low, medium, or high, which can be estimated by advanced high-resolution imaging technology or trained local citizens. The model is robust against missing values as in the absence of any single covariate, it performs with an AUC of at least 0.78. While taking a major step toward reduced-cost, less data-intensive CB forecasting, our results identify those key covariates that are worth the monitoring investment for highly accurate predictions.

1. Introduction

Cyanobacterial blooms (CBs) are becoming increasingly common in freshwater ecosystems, causing many negative impacts, socially and financially (Huisman et al., 2018). CBs are visually and olfactorily unpleasant and can be highly toxic, leading to various adverse effects on livestock, wildlife, fish, and humans (Backer et al., 2015). Climate change has been reported to catalyze the expansion of CB (Paerl & Huisman, 2008, 2009), and the frequency, intensity, and extent of harmful bloom occurrence is increasing worldwide (Ho et al., 2019). Interestingly, a recent study reveals a potential positive feedback loop between CBs and global warming (Bižić et al., 2020). This positive feedback is inspired by the "methane paradox," that is, "the production of methane in oxygenated surface waters," given that the production of methane, one of the most common greenhouse gases, during cyanobacteria photosynthesis has been shown.

CBs can have harmful effects on human health through exposure to toxins from recreation and insufficient treatment of drinking water. To mitigate these adverse effects, environmental managers can greatly benefit from predicting the presence of blooms. Many attempts have been made using various methods to understand the drivers of CBs. Nutrient availability is most often highlighted as the key factor of CBs (Wilhelm et al., 2011). However, other studies have shown that the importance of nutrients varies with other lake factors, such as stratification and meteorological conditions (Taranu et al., 2012). Even more so, several deterministic models show the importance of CB dynamics on environmental variables related to growth and other metabolic processes (Heggerud et al., 2020, 2022; Wang et al., 2007). However, there is no comprehensive study with all potential covariates included. Moreover, despite providing valuable intuition on CB, such models typically lack predictive power, and their mechanisms and key factors are assumed by expert knowledge drawn from empirical observations. Complex machine-learning models, such as neural networks, may predict future blooms, yet add little to our understanding of the dynamics (Muttil & Chau, 2006). Moreover, they do not reveal the influence of a single

Resources: Hao Wang
Software: Jingjing Xu, Pouria Ramazi
Supervision: Hao Wang
Validation: Christopher M. Heggerud, Hao Wang, Pouria Ramazi
Visualization: Jingjing Xu
Writing – original draft: Christopher M. Heggerud, Jingjing Xu, Pouria Ramazi
Writing – review & editing: Christopher M. Heggerud, Jingjing Xu, Hao Wang, Mark A. Lewis, Ron W. Zurawell, Charlie J. G. Loewen, Rolf D. Vinebrooke, Pouria Ramazi

covariate without assuming all else to be fixed. Namely, the synergistic effects involved in CB dynamics (Paerl & Otten, 2013; Whitton, 2012) are not addressed, hindering attempts to rank potential management strategies.

One challenge in predicting CB is the inconsistent collected data over different lakes. For a management region with multiple lakes, CB monitoring efforts are not evenly distributed to each lake. There exists an extended monitoring history for lakes with higher social-economical values and scattered monitoring data for lakes in remote, less-populated, or less-bloomed areas. Hence, available covariates for predicting CB are not the same over the different lakes. A flexible model is desired that can identify, and once available, use *primary covariates*, that are most informative and sufficient for predicting CB, and if unavailable, exploit the remaining *secondary covariates* to make the prediction. Mathematically, primary covariates are those minimal covariates that if conditioned on, the target variable becomes statistically independent of other covariates. Making predictions in the absence of some covariates is impossible with non-probabilistic models, such as linear regressions and neural networks, because they require all predefined inputs to be available when making predictions.

Bayesian networks (*BNs*) allow the development of a single predictive model including all covariates and the use of any subset of the covariates in a probabilistically marginalized fashion for making predictions (Ramazi et al., 2021a). A BN visualizes the joint probability distribution of the variables of interest by a directed acyclic graph whose nodes represent the variables and links represent probabilistic dependence, and a conditional probability distribution parameterizes the relation between each node and its parents. Moe et al. manually constructed a BN on CB based on their a priori knowledge (Moe et al., 2016). Additionally, Rigosi et al. (2015) constructed a similar BN structure based on the input of expert knowledge, and learned the parameters from data. However, both the parameters and structure of a BN can be learned partially (Jiang et al., 2021) or entirely from data and without manipulation (Alameddine et al., 2011; Feki-Sahnoun et al., 2018), allowing for insights as to which covariates are essential for predictions, but may have been neglected in previous studies. Once the BN is constructed, the primary covariates will be the *Markov blanket* of the target variable, which is the set of its parents, children, and co-parents of its children, and the reminder will be secondary covariates.

Here, to facilitate timely mitigation of the effects of CBs, we use BNs to predict future 2-week blooms. Furthermore, we learn the conditional dependencies between 17 multi-scale covariates and CB occurrences to better inform longer term management of CBs. Unlike the common practice in the literature that are limited to a single lake–except for example, (Rigosi et al., 2015)–we use the data of 78 water bodies in Alberta, Canada, over a 30 year time period. The learning is entirely based on data without the artificial intervention of the network structure.

Our objectives are as follows: (a) learn, for the first time, a BN entirely from data to predict future 2-week CB of a specified lake based on 17 covariates, including lake monitoring, meteorological, lake features, and watershed features; (b) identify the primary and secondary covariates form the structure of the resulting BN; (c) reveal the aspects of the structure that match and differ from the current state of knowledge of CB and motivate further research; (d) find the most informative covariates to CB prediction based on the bloom probabilities conditioned on a single covariate; and (e) find the performance of groups/subsets of covariates that are of interest.

We first show that the performance of our predictive BN model is comparable with other predictive models. We then reveal the information encoded by the learned network structure. With the learned network, we compare the predictive power of covariates by examining the informativeness of covariates on the test data. Finally, we compare the performance of the prediction for several scenarios of partially missing data. The informativeness of covariates and predictions using partially missing data will help monitor and predict bloom occurrence in Alberta.

2. Materials and Methods

2.1. Data and Variables

We used data of 78 Albertan lakes and reservoirs collected from 1986 to 2017 (Figure S1 in Supporting Information S1). The lakes and reservoirs are located within the longitudes of 49.3°–58.8°N, and latitudes of 110.0079°W - 119.21667°W. Our data set consisted of three parts: (a) the lake monitoring data collected in the open water season, May to October, with phytoplankton data collected as part of the Alberta Lake Monitoring Program, (b) the meteorological data, and (c) the lake and watershed information data. The used variables are summarized in Table 1 with histograms provided in Figure S4 of Supporting Information S1.

Table 1
List of Variables and Their Abbreviations

Variable	Description
Month	Month of the monitoring
Current cell count	Cell count of cyanobacteria sampled on the monitoring day
P	Phosphorus concentration sampled on the monitoring day
N	Nitrogen concentration sampled on the monitoring day
Secchi depth	Secchi depth of lake on the monitoring day
Stratified	Stratification state of lake on the monitoring day
Temperature	Average daily temperature from the day of monitoring to 2 weeks
	Before the next available cyanobacteria sample
Precipitation	Average daily precipitation from the day of monitoring to 2 weeks
	Before the next available Cyanobacteria sample
wind speed	Average daily wind speed from the day of monitoring to 2 weeks
	Before the next available cyanobacteria sample
Solar radiation	Average daily solar radiation from the day of monitoring to 2 weeks
	Before the next available cyanobacteria sample
Lake depth	Lake depth
Lake elevation	Lake elevation
% pastureland	% pastureland in watershed
% cropland	% cropland in watershed
% forest	% forest in watershed
% wetland	% wetland in watershed
Lake-watershed area ratio	Lake-watershed area ratio
Target	Cell count of cyanobacteria at the next available sample (about 2 weeks in the future)

The collected data set was not ideal for predicting a fixed time in the future as the time difference between two monitoring/samples of the cyanobacteria cell counts was not fixed (Figures S2 and S3 in Supporting Information S1). Having the goal of predicting future 2-week blooms, out of every two consecutive cyanobacteria cell count samples for each lake, we took the latter as the target variable, say at time T, and the former as a covariate which was sampled at anytime from about $T-44$ to $T-14$. We set $T-14$ as the time the prediction is made, so the meteorological data were averaged from $T-14$ backward to a length of at most 30 days, that is, $T-44$. This mimics the realistic situation where from "today" ($T-14$), we want to predict blooms 2 weeks in the future (T) using the previous cell count samples for the lake, which could have been collected anytime from a month ago to today ($T-44$ to $T-14$), and the meteorological data are available for all past days.

Alberta Environment and Parks (AEP) provided data other than those related to meteorology. The original lake monitoring data set contained chemical concentrations in lake water samples, Secchi depth, lake stratification, cyanobacteria cell count, etc. Depth-integrated water samples were taken for water columns in the euphotic zone at 10 locations in a lake and then mixed to obtain the composite sample to represent the overall condition of the lake. The regional watershed information for each lake was generated using ArcGIS in a previous study (Loewen et al., 2020).

We took the following steps to obtain the rest of the covariates for our data set:

1. **Total cyanobacterial cell count data.** The cell count data were extracted from a taxonomic enumeration data set of phytoplankton communities in water samples. We first extracted all the cyanobacteria cell count (see Table S1 in Supporting Information S1 for a list of 256 cyanobacteria taxa included in our study) from the phytoplankton data set of 1,306 phytoplankton taxa in sampled lake water. We only used data sampled in the open water season and collected following the standard protocols (see Figure S2 in Supporting Information S1 for the monitoring times). With respect to the definition of blooms, the guidelines vary from

country to country and by scenario. Here, we define a bloom to be an algal density over 1,530 cells/mL, which is the median CB density of sampled lake data. The 100,000 cells/mL threshold by WHO (Chorus and Bartram, 1999) was devised for near-shoreline areas where surface grabs (of the top several inches of water) were common and not directly comparable to our pelagic zone sampling, where samples were taken vertically from one to several meters below the surface. For this reason, when a bloom is present we expect our samples to show significantly lower cell counts than the WHO threshold. Furthermore, our threshold is consistent to observed HABs in literature (Trainer et al., 2020)

2. **Meteorological data.** Daily meteorological data were extracted from the software BioSIM (Régnière et al., 2014) and its model "Climatic Daily" that takes as input the geographic information including longitude, latitude, and elevation, and BioSIM generated daily mean temperature (Celsius), total precipitation (mm), wind speed (km/hr), total radiation (MJ/m^2). The software obtained the meteorological data for each lake by interpolating data measured at the local weather stations in Canada from 1900 to 2017 with some adjustment according to the lake elevation. Then, for each data instance, we took the average of meteorological data from the monitoring date to 2 weeks before the day of the prediction. For example, for lake 1 with a sample on 4 July 2016, and a target prediction time of 25 July 2016, we averaged the meteorological data between July 4 and July 11, so we could make a 2-week ahead prediction using the available meteorological data on and before 11 July 2016. We restricted the weather history-length (the number of days before the prediction date) to a maximum of 30 days, as averaging weather over too many days could reduce the prediction power.
3. **Lake features and Watershed data.** Lake elevation and lake-watershed area ratio were obtained using geographical information system (ArcGIS 10.3.1). Proportional coverage of forest and wetlands were obtained from a reclassified Agriculture and Agri-Food Canada 2010 Land Use grid (ca. 2010, see AAFC (Agriculture and Agri-Food Canada), 2015), representing natural land cover. Proportional coverage of pastureland and cropland were obtained from reclassified Alberta Biodiversity Monitoring Institute (ABMI) Human Footprint Inventory data (ca. 2016, see ABMI (Alberta Biodiversity Monitoring Institute), 2018; for reclassification details, see Table S4 in Loewen et al., 2020), representing human-footprint.

We chose those variables according to a previous multi-scale study of the CB in the study area (Loewen et al., 2020). Some factors were not included, such as buoyancy and water temperature, because of a considerable portion of missing data. However, we included air temperature, which is a representative of water temperature and easier to obtain with the help of the weather stations. A summary of variables is shown in Table S3 of Supporting Information S1. The original covariates except for the current month and stratification of the lake were continuous. We discretized the 15 covariates by frequency into three levels: low, medium, and high (with approximately one third in each level). The three levels are essentially formed based on historical observations and local knowledge of the lake. Moreover, although more levels might yield more accurate predictions, we limited them to three so that trained citizens can also easily distinguish and report the level, avoiding the need of costly monitoring. The target variable, cyanobacteria cell count on the target date, was discretized into two levels based on the threshold of 1,530 cells/mL discussed previously with "0" indicating bloom-free and "1" indicating a bloom event.

2.2. Partitioning Data Into Training and Testing

Unlike the typical machine-learning approach of randomly partitioning the data set into *training* (calibration), used for learning the structure and parameters, and *testing* (validation), used for evaluating the model, here, we used a temporal partitioning. For each lake, we included in the testing data set the data instances collected for that lake during the most recent 3 years and included the remaining older data instances in the training data set. For example, if a lake was sampled at years 2010, …, 2018, we used instances at 2018, 2017, and 2016 for the testing and 2010, …, 2015 for the training. Since in practice, we may want to apply the model to a lake that we have never sampled before, we held out six lakes for the testing data set and did not use any of the corresponding data instances in the training. This resulted in a test data set consisting of 95 data instances, that is a portion of 15% of the whole data set, and the remainder was included in the training data set. The distribution of month over the test data set was as follows: 12 June, 46 July, 64 August, 48 September, and 20 October instances.

2.3. Learning and Testing

We found the BN structure that scored the lowest Bayesian information criterion (BIC) (Schwarz, 1978) on the training data set. BIC measures the fit to the data set via the likelihood function and penalizes the number of model parameters. Thus, it can prevent overfitting, which explains its use in learning BNs in the literature

Table 2
AUC Scores of the Selected Bayesian Network (in Bold) and Other Predictive Models

Model	AUC on train	AUC on test
BN	0.86	**0.79**
GLM	0.89	0.85
NN	0.78	0.83
GBM	0.93	0.82
NB	1	0.75
KNN	0.86	0.71

Note. AUC for a random classifier is 0.5.

(Alameddine et al., 2011; Feki-Sahnoun et al., 2018; Ramazi et al., 2021a). Additionally, we assumed no manipulation, or forced structure in construction of the BN, as to allow the entire structure to be learned from the training data set only. We then learned the BN parameters, and compared the prediction accuracy of the resulting BN with four predictive models: generalized linear model (GLM), naive Bayes (NB), boosted decision tree (or, gradient boosting machine (GBM)), k nearest neighbor (KNN), and neural network (NN) (Ramazi et al., 2021b). We used the area under the curve (AUC) (Hajian-Tilaki, 2013) as the performance measure over the test data set. We tested whether there is a significant difference between the performance of BN and the other models.

Then we tested the performance of the BN when none or one of the primary covariates were missing and reported both the AUC and the area under the precision-recall curve (AUPR) (Raghavan et al., 1989), which is appropriate for unbalanced data.

The code was implemented in the programming language `R`. The package `bnstruct` (Franzin et al., 2017) was used to find the optimal BN from data. Analysis of the obtained BN, including the computation of the conditional probabilities, was performed using the `bnlearn` package (Scutari, 2009). The NN model was implemented using the `deepnet` package (Rong, 2014), the NB model was implemented using the `bnlearn` package, and the other tested models, that is, GLM, KNN, and GBM, were implemented using the `caret` package (Kuhn, 2008). Based on our prior experience, we set the value of k to 5 for the KNN model and the number of trees to 100 for training the GBM. The number in the best iteration was used during the test. The NN had one hidden layer with the number of neurons equal to half of the number of the input variables and the sigmoid function as the activation function. The AUC and AUPR were calculated using the `pROC` (Robin et al., 2011) and `PRROC` (Grau et al., 2015) packages.

3. Results

3.1. Prediction Accuracy

The final trained BN scores the best BIC on the training data set and predicts future 2-week CB bloom with 0.79 AUC on the testing data set and 0.86 AUC on the training data set. The AUC scores of the competitive models range from 0.71 to 0.85 on the testing data set (Table 2, See Figure S6 in Supporting Information S1 for the ROC and AUPR curves and Table S6 in Supporting Information S1 for the AUC of each lake). Using the DeLong test (DeLong et al., 1988), we find that the AUC score of BN is not significantly different from the AUC of the other six models. The prediction accuracy of our model is comparable to other competitive models.

Table 3
AUC and AUPR Scores of the Selected Bayesian Network When the Value of a Single Covariate Is Missing

Missing covariate	AUC	AUPR
Nothing missing	0.79	0.82
A secondary covariate	0.79	0.82
Current month	0.81	0.85
Current cell count	0.79	0.81
P	0.80	0.83
N	0.82	0.84
% pastureland	0.78	0.81

Note. The results are on the test data set. AUC and AUPR for a random classifier on the test data set are 0.5 and 0.55.

3.2. Network Structure

The BN suggests the following covariates as primary, which form the *Markov blanket* of the response variable bloom and are sufficient for predicting it (Koller & Friedman, 2009): the current cyanobacterial cell count, phosphorous, nitrogen, % pastureland in the watershed, and the current month. These covariates are sufficient for the BN to achieve an AUC of 0.79. The remaining covariates are secondary, which do not contribute to the above prediction but are used when the value of one or more of the primary covariates is missing. Namely, the prediction accuracy does not change when any secondary covariates are missing, given the five primary covariates. Should the value of any of the primary covariates be missing, the prediction performance would not get lower than 0.78 AUC (Table 3).

The structure of the BN (Figure 1) proposes a number of probabilistic conditional independencies among the variables (Table S4 in Supporting Information S1). One observation is the presence of watershed covariates, in particular, % pastureland, in the Markov blanket. Furthermore, it is not surprising that the covariate, current month, is present as primary because of the high latitude of Alberta.

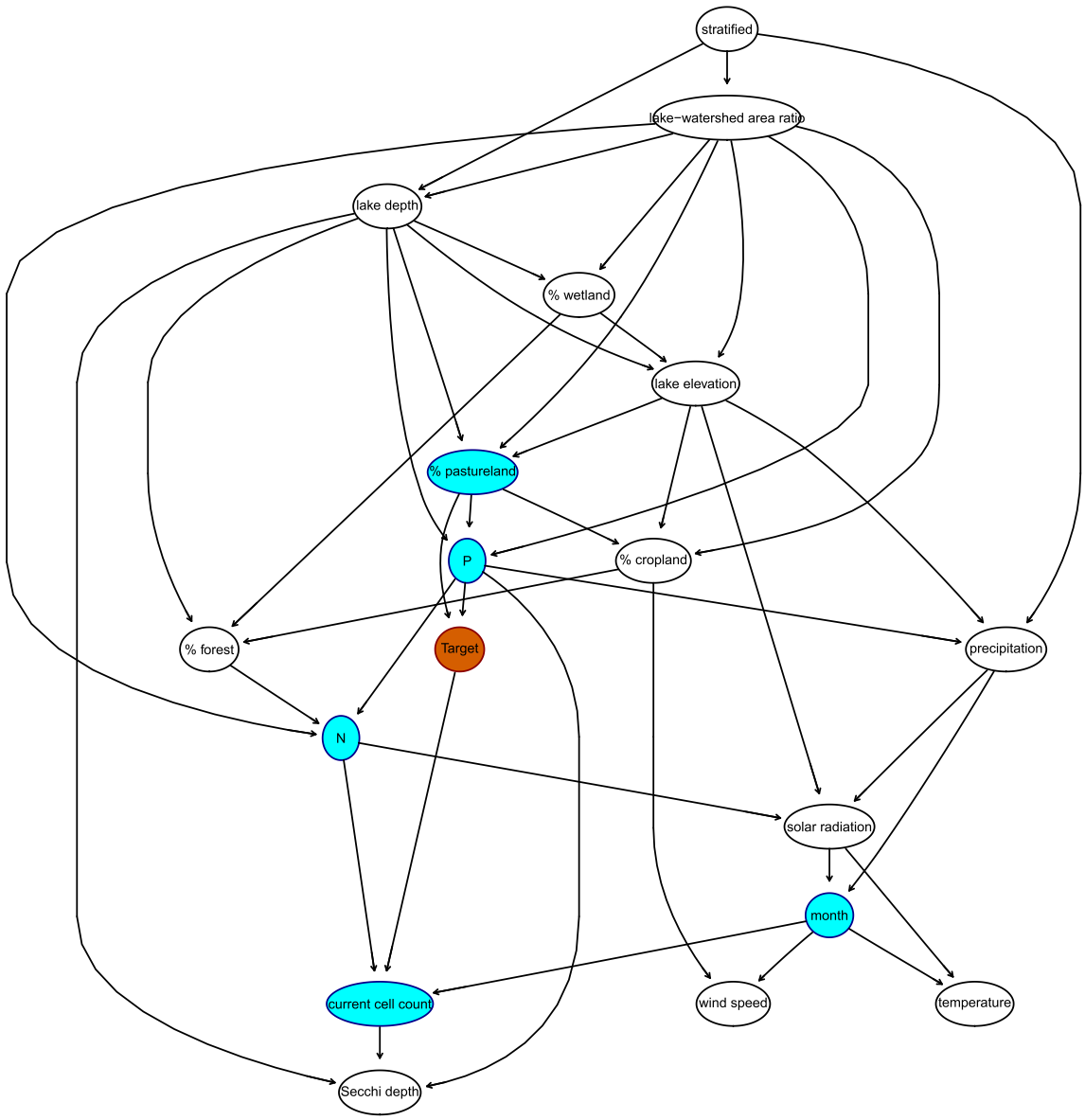

Figure 1. The structure of the selected Bayesian network. This network has the lowest Bayesian information criterion score on the training data set. The red node is the response variable (*future 2-week CB*), the other nodes represent the current covariate levels. Blue nodes are in the Markov blanket of the response variable, and hence, primary covariates, and white ones are the secondary covariates. The links do not represent causal relationships but conditional probabilistic dependencies. Note that all the continuous predictor variables were discretized into three levels, and the response variable was discretized into two levels, that is, bloom-free and bloom.

We further note that the meteorological covariates are statistically dependent on month, whereas the nutrients within the lake are independent of the month.

The existence of links between watershed variables and nutrient measurements are particularly interesting because they indicate a connection between land use and eutrophication that highlights potential anthropogenic nutrient sources. Additionally, the existence of a link between two variables implies that the two do not become independent conditioned on any other (subset of) variables. It does not imply the ability of one node to fully predict another.

3.3. Informativeness of the Covariates

Comparing the informativeness of covariates helps to identify the covariates that are important for CB prediction as a complement to the primary covariates. The obtained BN allows us to determine the probability of future

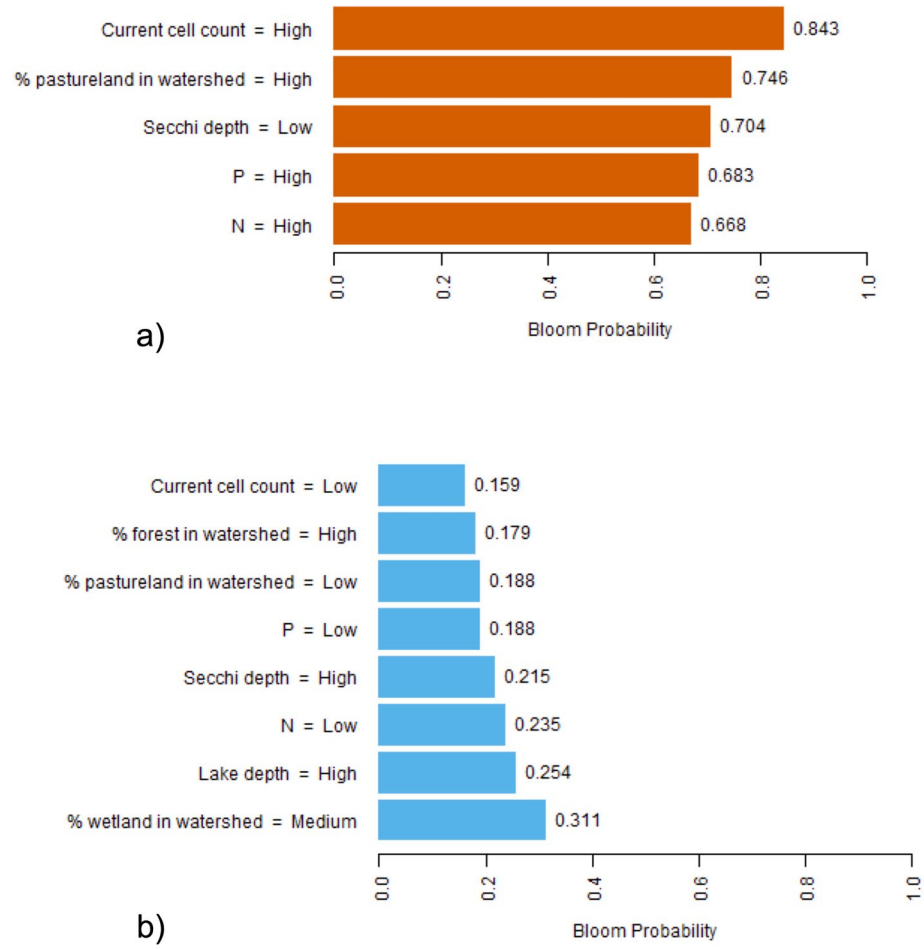

Figure 2. Predicting risk of bloom with the obtained Bayesian network (BN) given the state of a single covariate. Predictions are made by inputting the value of a single covariate into the BN. (a) The probability of a high bloom level in 2 weeks (*x*-axis) based on the current covariate levels (*y*-axis). As an example, the top bar should be interpreted as follows: There is a bloom probability of 0.843 in the future 2 weeks, if the current cell count is High. (b) The probability of a low bloom level in 2 weeks (*x*-axis) based on the current covariate levels (*y*-axis). See Figure S7 in Supporting Information S1 for the bloom probabilities of all covariates.

2-week bloom conditional on each of the discretized covariates, for example, Prob(bloom |nitrogen = low). We find five covariates that predicts high risk of bloom (Prob(bloom = 1 |covariate = *X*) > 2/3, for covariate level *X*) on the test data set: high level of current cell count predicts the bloom in 2 weeks with a probability of 0.843, and high % pastureland in watershed predicts a bloom of probability 0.746, followed by low level of Secchi depth (0.704), high level of phosphorus (0.683), and high level of nitrogen (0.668) (Figure 2a). There are 8 scenarios with low risk bloom (Prob(bloom = 1 |covariate = *X*) < 1/3, for covariate level *X*) (Figure 2b): high levels of % forest in watershed, Secchi depth and lake depth, low levels of current cell count, % pastureland in watershed, phosphorus and nitrogen, and medium level of % wetland in watershed. Those high-risk scenarios and low-risk scenarios are useful to quickly estimate the risk of bloom.

Another way to study the relative informativeness of the covariates is to compare the level of certainty that a single covariate can provide before obtaining the state of that covariate. We use Shannon's entropy to measure the level of uncertainty for a probability distribution, where "0" means the lowest uncertainty level, or most certain, and "1" is the highest uncertainty level, for example, probability distribution of tossing a fair coin. We rank the uncertainty (Table S5 in Supporting Information S1) of the conditional probability tables (CPT) using the entropy (Shannon, 1948). We find that the CPT of the current cell count level has the lowest uncertainty (0.75), meaning that knowing the current cell count level is more likely to avoid ambiguous prediction for bloom occurrences, if one were to choose only one covariate for prediction. There are five covariate CPTs with an uncertainty level

Table 4
Groups/Subsets of Covariates and Their Corresponding Prediction Performance Over the Testing Data Set

Group/subset	Covariates	AUC
Lake Monitoring	Current cyanobacterial cell count, P, N, Secchi depth, stratification status (obtained on the currently most recent sampling date)	0.78
Meteorology	Temperature, precipitation, wind speed, solar radiation (the average daily meteorological condition from the sampling date to today)	0.53
Lake feature	Lake depth, lake elevation	0.73
Watershed feature	% pastureland, % cropland, % forest, % wetland, lake-watershed area ratio	0.77
Regional covariates[a]	All the covariates in the groups of meteorology, lake feature and watershed feature	0.77
Regional covariates + local[b]	In addition to the regional covariates we add covariates that can be estimated locally: the level of total cyanobacterial cell count and the Secchi depth	0.83

[a]Regional covariates refer to the meteorological covariates and those that are fixed for a given lake. [b]Local covariates here only refer to the level of total cyanobacterial cell count and the Secchi depth.

between 0.84 and 0.9, including % pastureland in the watershed, phosphorus, Secchi depth, % forest in the watershed, and nitrogen.

3.4. Performance of Groups of Covariates

We categorize the covariates except current month into four groups: lake monitoring, meteorology, lake features and watershed features (Table 4) and investigate the prediction performance of a single group, using our obtained BN and values of the chosen group in the test data set. We find that the lake monitoring covariates performs the best, with an AUC of 0.78, almost the same accuracy as when all covariates are available in the BN. The next best is the group of watershed features (0.77 AUC). Just the two lake feature covariates predict bloom with a 0.73 AUC. The meteorological covariates predict almost as poorly as a random classifier, that is, 0.53 AUC.

We hypothesize that using an educated estimate might help boost the prediction accuracy, when we need timely predictions of CBs. With limited ability of getting lab-based output from monitoring samples promptly, the total cyanobacteria cell count and Secchi depth may be estimated by a well-trained local by identifying the level of cyanobacteria amount and the water clarity using their own experience and some scientific guideline. Moreover, with recent advances in high-resolution imaging, automated cyanobacterial cell counting methods (Graham et al., 2018) could be used to generate data of sufficient taxonomic resolution (i.e., total cyanobacterial cell count) in a timely manner for use in the predictive model. We compare two subsets of the covariates (Table 4). One subset of regional covariates includes the covariates associated with meteorology, lake features and watershed features. Using this subset and our obtained BN, we predict the bloom in 2 weeks with an AUC of 0.77. We complement the set of regional covariates with the estimates of the level of cyanobacteria growth (corresponding to cell count) and the water clarity (corresponding to Secchi depth) by a well-trained local or the high-resolution imaging technology. With the additional data—estimated level of cell count and Secchi depth, the prediction of bloom in 2 weeks outperforms the prediction using only the regional data, and ends up with an AUC of 0.83, which we take as our final predictive model.

4. Discussion

We obtained the following three main findings: (a) the purely data-based BN structure is consistent with most results from previous studies, identifying five primary covariates—P, N, and cyanobacteria cell count from the most recent sample, the pastureland percentage in the corresponding watershed, and the current month; (b) the BN proposes the most recent cell count, the pastureland percentage, and the Secchi depth to be the three major covariates that are the most informative as a single predictor of CB; (c) the BN predicts future 2-week CB with AUC 0.79 in the testing data set; however, if we select regional and local covariates as the only input variables, the AUC increases to 0.83.

4.1. Prediction Accuracy

Our approach to predicting future 2-week CB provides statistically similar accuracy as six other competing approaches. As for the performance of other studies (see references in Rousso et al., 2020), only a few had a

forecast horizon of at least 2 weeks, and we could not make a conclusive comparison because most of them use R^2, that is, the coefficient of determination, instead of AUC scores. Nevertheless, our obtained BN is fully data based, which relaxes the stated limitation of requiring expert knowledge to develop BN models in previous work (Rousso et al., 2020).

Our prediction framework differs from others in that we do not require the lake monitoring data to be collected exactly "today" to make predictions for 2 weeks later. Instead, the monitoring data could be collected several days ($0 \sim 30$) before "today" (see Figure S3 in Supporting Information S1), and the past meteorological data were used to complement the monitoring data gathered earlier. Although we had to choose such a framework due to the irregular gaps between consecutive monitoring dates, our framework better represents the reality that monitoring is not very frequent, and there is an expected time delay in reporting the monitoring data, especially the cell count. The time lag can be variable—perhaps anywhere from 48 hr for an emergency sample to weeks or months for non-priority sampling (Graham et al., 2018).

While the obtained BN has the advantage of handling missing values, compared to other models, it could fall short of making accurate predictions. Nevertheless, there was no significant difference between the AUC of the BN and other models, supporting the choice of BNs. The AUC of the BN was slightly better than NB, which is consistent with predicting *Karenia selliformis* blooms in Feki-Sahnoun et al. (2018). This supports the dependence between the covariates even if conditioned on the target variable. While the parameters of GLM and NB are obtained using standard methods such as the maximum likelihood estimation (Koller & Friedman, 2009), obtaining the optimal configurations of some of the models may not be possible. In particular, a NN may take an arbitrary number of layers and neurons as well as an arbitrary activation function, and finding the "best" architecture is an open problem. Nevertheless, one may "hope" to obtain a competitive configuration by following some rules of thumbs such as those in Panchal et al. (2011) as we did in this paper.

The reason why omitting some of the covariates increases the accuracy (Table 3) is as follows. Our aim is to find the model that best describes the relationships between the variables and does not overfit the data set. Hence, we optimized the BN based on BIC (although one could also use AIC). This is a *generative* task (Jebara, 2001) with the goal of best estimating the joint distribution $\text{Prob}(B, \mathcal{X})$ where B is the target variable, bloom, and $\mathcal{X}$ is the set of the covariates. Now, when we use the model for predicting bloom, we are performing a *discriminative* task, that is, estimating the conditional distribution $\text{Prob}(B \mid \mathcal{X})$. The log of these likelihoods are being maximized during the training phase, and the log of the first likelihood has an extra term compared to the second: $\log \text{Prob}(B, \mathcal{X}) = \log \text{Prob}(B \mid \mathcal{X}) + \log \text{Prob}(\mathcal{X})$. Thus, the generative task comes with the cost of having perhaps a lower prediction accuracy of the target variable since it is also optimizing the extra term $\log \text{Prob}(\mathcal{X})$ which stands for the log-likelihood of the covariates. So it is possible to obtain another model, such as NB, that is more accurate at predicting bloom but less accurate at matching the likelihood of the data, or equivalently the relationships between the covariates.

4.2. Network Structure

To exploit the full potential of the data set, the network structure was learned entirely from the training data set and had no prescribed connections or parameters (Ramazi et al., 2021a), which could have resulted in a different structure. Although previous studies suggest that such manipulations based on expert knowledge can decrease the complexity of the network (Alameddine et al., 2011), one goal of this work was to compare an entirely learned structure to the state of current knowledge.

The inclusion of P and N concentration in the Markov blanket of the target variable reinforces the idea that they both play an essential role in bloom formation (Dolman et al., 2012). Moreover, the link between N and P in the BN implies that both nutrients are likely to contribute to nutrient pollution, as opposed to the excess of just a single nutrient (Lang et al., 2013). The implied correlation between N and P from the network structure also leads to the unsurprising link between N and current bloom status. There is an overwhelming body of literature supporting that nutrient concentrations drive CB dynamics, and our model is one such study that supports this assertion. Furthermore, the temperature is believed to play an important role in bloom formation. Keeping in mind that Alberta has large average temperature fluctuations between seasons, literature would suggest that the season, or month, would be an important factor for evaluating the probability of CB. In our model, we note that the current month is linked to the current cyanobacterial density. However, since we are making 2-week predictions,

it is not surprising that the month is not directly linked to the target but is still contained in the Markov blanket. Lastly, we note that our network structure shows a link between land type/use within the watershed of the focal lake and nutrient concentrations in the lake water. Although this link is sparsely discussed in the literature, there is an intuitive understanding of the link between the surrounding land type and nutrient concentrations. That is, lakes surrounded by farmlands such as pastureland and cropland may be more eutrophic than lakes surrounded by forests or wetlands due to an increase in anthropogenic nutrient sources (Carpenter, 2005; Keatley et al., 2011; Rodríguez-Gallego et al., 2017).

The conditional independencies imposed by the network suggests that certain measurements may not be required. For example, our model shows that knowing Secchi depth is not necessary for making a prediction, provided we know the covariates in the Markov blanket. This result is informative, as it suggests that even though Secchi depth is commonly used to predict algal abundance, it may be unnecessary (Canfield & Hodgson, 1983; Krause-Jensen et al., 2009). Also, certain lake characteristics such as lake depth and stratification are not deemed valuable predictors compared to other covariates. It is commonly referred to in the literature that CBs are often successful in stratified conditions due to their ability to regulate their buoyancy (Paerl & Huisman, 2009; Reynolds et al., 1987; Wang et al., 2007), and that deeper lakes generally have lower CB abundance (Berger et al., 2006; Wang et al., 2007). The covariates related to meteorology, wind speed, precipitation, air temperature, and solar radiation are not directly required for our prediction either. There are many reasons to expect that meteorology plays a crucial role in bloom formation. Several CB species growth rates are shown to have optimal growth near a temperature of 25°C (Butterwick et al., 2005), and higher winds encourage water column mixing, which in turn alters light and nutrient availability (Paerl, 2014).

Although these results may stray from the common beliefs found in the literature, they are not contradictory nor necessarily based upon error. When considering a more extensive breadth of environmental covariates, new and previously unconsidered covariates can cover the information provided by the typical covariates. For example, Secchi depth, in part, is a way of measuring the turbidity of a lake. However, one can reasonably postulate that the turbidity is also directly related to P and N concentrations, or perhaps even the surrounding land use. Hence, if we know the P and N concentrations and know the lake is in an agricultural watershed, Secchi depth may not provide any further information. Additionally, the absence of meteorological covariates required for the prediction can be explained by the relatively predictable monthly meteorological patterns of our study system of Alberta, Canada. That is, the average meteorological covariates are described mainly by the month from which we are predicting.

Moreover, BIC is a *consistent score* for learning BN structures; that is, as the data size approaches infinity, the structure that correctly represents the conditional independencies between the variables minimizes BIC and no other structure does so (Koller & Friedman, 2009). So for a large enough data set, the Markov blanket is expected to be correct. While this supports our approach to find the primary covariates, the validity of the results is limited to the size of the dataset–an issue with almost all data-driven results. One may consider re-training the model once more data is collected.

Lastly, our network shows several logically standard links, that is, the meteorological and land type covariates are interconnected. Of course, the surrounding land-use variables should be anti-correlated in the sense that the sum of the proportions of surrounding wetland (%), pastureland (%), cropland (%), and forest (%) should not exceed 100. The meteorological covariates and month should also be correlated as the average meteorological conditions in Alberta are highly cyclic and regular annually, implying a direct link to season or month. Finally, we expect a grouping of the covariates pertaining to nutrient concentrations, surrounding land use, current bloom status, and target. When a lake is considered eutrophic, it is often eutrophic in both P and N, although P is generally the main driver of bloom dynamics (as we see in this network (Figure 1): P is directly but N is indirectly linked to the target variable). We further expect N and P to be influenced by the type of surrounding land, that is, it is argued that pasture and croplands will contribute to eutrophication more than the other types of land in the watershed. However, for a given lake, whether a bloom will occur also depends on the local conditions prior to the target date.

4.3. Insights for CB Management and Mitigation

The BN structure and prediction results also present useful information in terms of management and mitigation strategies. Current mitigation strategies can be categorized as either within waterbody or within watershed/airshed controls (Paerl et al., 2016). Various strategies apply to each of these categories, although both within-lake and within-watershed interventions may have unintended consequences. Some within-lake strategies involve

significant alteration of the ecosystem, such as algaecides, sediment capping or dredging, and food web manipulations. There may be no direct link between such mitigation strategies and the alteration of covariates considered in our model thus, our work is limited in understanding the efficacy of such strategies. For example, in-lake remediation strategies targeting nutrient levels may also indirectly alter turbidity, in which case multiple covariates are altered in a way that is difficult to determine. However, certain within-watershed mitigation strategies may directly target covariates in our model, in which case we can offer insight into the potential effectiveness of that strategy.

First, it is readily known that excess nutrient inputs of nitrogen and phosphorus increase CB abundance. It is contested in the literature whether P, N, or a combination of these variables drive freshwater CB abundance (it is also worth noting that relationships differ in marine environments). However, due to the ability of several genera to fix nitrogen, phosphorus reduction has been thought to be more feasible in managing blooms. Many avenues can be pursued to reduce anthropogenic nutrient inputs into the watershed, a few examples that could potentially apply to Alberta lakes include; reduction of urban and industrial waste, reduction of agricultural pollutants such as fertilizers, manure and disturbance of legacy nutrients in soils, and riparian habitat restoration or preservation to sequester mobile surface nutrient. All the above strategies would directly target a reduction in our model's P and N covariates. Interestingly, P and N are contained within the Markov blanket, and Figure 2 shows that low P or N concentrations give a low probability of bloom occurring. It is worth noting that our model shows a low P concentration has a lower probability of a bloom occurring than N (19% vs. 24%). At the same time, a previous work in the province (Loewen et al., 2021) implies that the ratio of N:P may be important in Alberta, with higher N to P favoring non-cyanobacterial species which are unable to fix atmospheric N.

Additionally, altering the land type surrounding a lake and within the watershed has been shown to be a useful mitigation strategy. For example, thoughtful placement of riparian buffers within the watershed can increase the overall water quality and decrease nutrient pollution (Anbumozhi et al., 2005). Also, wetlands, both constructed and natural, can act as supplemental wastewater treatment processes to remove and trap even more nutrients (Paerl et al., 2016; Vymazal, 2010). To this end, our model supports the claim that land type within the watershed is an important indicator for predicting CB blooms. Although the watershed land-type variables (% pastureland, % cropland, % wetland, and % forest) are all correlated variables, we see certain trends in our model. As seen in Table 4 watershed features, in general, can be helpful in predicting CB blooms, however, the results do not single out an increase in percent wetlands as a potential management strategy as suggested by many studies (Paerl et al., 2016; Vymazal, 2010). Instead, as seen in Figure 2, our model suggests that reducing the percentage of pastureland may result in a lower probability of CB blooms.

4.4. Informativeness of the Covariates

Our findings that the level of cyanobacterial cell count, Secchi depth, nutrients, and most watershed covariates are among the most informative covariates are consistent with previous studies in other water systems (Rigosi et al., 2015; Steffen et al., 2017; Zhao et al., 2019). Contrary to the conventional findings on the crucial effect of wetland on nutrient retention and water quality (Baron et al., 2013; Verhoeven et al., 2006), we found that a high level of % wetland predicts a higher risk of bloom compared with a medium level of % wetland. We speculate that in addition to the total amount of wetland in a watershed the number of wetlands may also be important, as the accumulative effect of small wetlands in controlling CBs is reported to outperform one large wetland (Cheng & Basu, 2017). While wetlands frequently intercept and retain nutrients, they maintain stronger groundwater connections and nutrient supplies to lakes that may create eutrophic lake conditions that support blooms in undeveloped watersheds, especially later in the season (Loewen et al., 2020). We did not expect that % pastureland in the corresponding watershed of a focal lake ranks higher than the level of phosphorus and nitrogen within the lake, even though the % pastureland of a lake remains stationary. In combination with our finding that the covariate group of watershed feature has an AUC of 0.77 (Table 4), one of our important future directions is to develop a preliminary appraisal using our model framework and those stationary covariates, including lake and watershed features, for the less frequently monitored and currently bloom-free lakes in our study region. This will enable the environmental managers to make proactive management strategies (Qu et al., 2014) so that the limited monitoring and management resources can be efficiently allocated to the lakes with higher pre-assessed bloom risk before detrimental outcomes occur (Michalak et al., 2013). Taking proactive measures could possibly reduce the cost of management, because cyanobacteria are reported to be resistant to current mitigation efforts and management measures are notoriously costly (Cooke et al., 2016). To our surprise, unlike studies in Kosten et al. (2012) and

Taranu et al. (2012), none of the meteorological covariates here were shown to be very informative. The reason behind this is unclear to us, even though we acknowledge the correlation between the current month and meteorological covariates. One explanation for the lack of importance of meteorological factors is that they may have limited direct influence, but influence blooms indirectly. For instance, warm or low wind conditions might push an agriculturally-driven watershed with high nutrient concentrations into a bloom, where the bloom might not otherwise have occurred if not for the combination of factors.

Our model suggests that the connections described above are existent; however, when a covariate has a high uncertainly level with respect to its probability distribution of predicting blooms, (Table S5 in Supporting Information S1) it may not be a strong predictor. Consequently, the connection may be spurious. This highlights the importance of determining the confidence over the BN connections, which can be done by *Bayesian averaging* (Koller & Friedman, 2009) and is left as future work. In this light, the interpretations provided are merely a potential explanation of the model results, regardless of the confidence of the model.

4.5. Performance of Groups of Covariates

A valuable feature of the BN is its ability to make predictions with missing covariates. This feature makes its range of applicability much broader. For example, a subset of the regional covariates could be used to make predictions if prompt monitoring data is not available. In certain instances, only a subset of covariates may be measurable due to technical issues or varied interests at a given time and region. Furthermore, our model can help understand which group of covariates is most useful for predictions, potentially cutting down on data collection costs and sampling. We considered six logical groups/subsets of covariates to show the applicably of our model for varying availability of data. All groups/subsets of covariates except the meteorological group perform reasonably well. In particular, considering the timely availability of data, the model can make good 2-week predictions without relying on detailed analysis of chemicals. The poor performance of the meteorological covariates may be explained by the invariability in the data, lack of important covariates (e.g., those in the Markov blanket), or the structure of the conditional dependencies of the covariates being used. In any sense, the ability of our approach to utilize subsets of covariates to make predictions makes it a potentially powerful tool for timely predictions to be made by health officials, managers, and scientists in general.

4.6. Limitations and Future Direction

Our data set includes a few frequently sampled lakes and many less frequently sampled lakes (see Figure S2 in Supporting Information S1). We obtain a representative set for testing, as is described in Section 2. However, the imbalanced nature of our data set might have lead to an unexpected prediction of a higher bloom level when the % wetland level is high (see Figure S6 in Supporting Information S1 for the conditional probability for % wetland). It is also possible that this unique result is due to the effect of wetland sizes (Cheng & Basu, 2017) which we do not consider. Obtaining more data could help to uncover the underlying cause for this result. When there is enough data, a balanced data set can be constructed to generate a BN. From that BN, we could verify the effect of % wetland in predicting CB.

The possible reason for meteorological covariates not showing a critical contribution to CB prediction as much as other groups of covariates is that we restrict our study to lakes in Alberta, where there are not distinct meteorological variations among lakes. The meteorological covariates might be shown to have a larger impact in CB prediction for a larger scale, for example, continental or global scale.

Overall, our approach of predicting CB performs as well as the competing methods and proposes the importance of watershed features in the CB study for regional water body systems. The proposed framework allows us to achieve more realistic goals when data are partially missing. Limited by the size of the data set, we did not include the effect of buoyancy because its effect is not as important as shown in an earlier study (Loewen et al., 2020). In the future, it would be interesting to justify the role of meteorology for CBs in the study area. Extending the methodology to obtain, rather than a discrete, a continuous BN (Jackson-Blake et al., 2022) allows predicting the Cyanobacteria level in a finer scale. The informativeness of lake and watershed features suggests that our approach could help prediction when monitor data is not available for some lakes.

5. Conclusion

We developed a purely data-driven BN that uses watershed characteristics, meteorological conditions, Secchi depth, and the (low, medium, high) level of cyanobacterial cell-count to predict future 2-week CBs in lakes with

0.83 AUC. Some predictive models, such as a GLM, perform up to 0.02 AUC higher; however, they cannot make predictions when one or more of the covariates are randomly or systematically missing in the data as in large-scale monitoring. However, being a probabilistic model, the BN can marginalize the missing covariates and may perform only 0.05 AUC worse in the absence of a covariate, making it a suitable fit for the case with incomplete, partially-monitored data. On the other hand, unlike most machine-learning models that bear a black-box design with little to no insight, BNs are probabilistic and graphical, revealing the dependence between the variables. In particular, the obtained BN identifies five primary covariates: P, N, percentage of pastureland in the watershed, current cyanobacterial cell count, and current month, and uses the remaining secondary covariates to predict CB in the absence of one or more of the primary. Moreover, the probabilistic analysis of the BN proposes cell count, pastureland percentage, and Secchi depth as the three most informative covariates in predicting CB. We further discussed known and new insights on the relationship between the covariates and their "contribution" to the predictions and implications on CB management and mitigation. Overall, the BN modeling approach has a high potential for predicting CBs in lakes using incomplete, timely data and identifies those informative groups of covariates that are worth future monitoring investments for further improving the prediction accuracy.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data set used in this work was compiled by Alberta Environment and Parks. Both the data set and codes for this work are available at https://zenodo.org/doi/10.5281/zenodo.10109224 (Heggerud et al., 2023).

Acknowledgments
This research was primarily supported by a Grant from Alberta Conservation Association. Hao Wang was partially supported by an NSERC Individual Discovery Grant RGPIN-2020-03911 and an NSERC Discovery Accelerator Supplement Award RGPAS-2020-00090 as well as a Canada Research Chair. Mark Lewis was supported by an NSERC Discovery RGPIN-2018-05210 and the Gilbert and Betty Kennedy Chair. Pouria Ramazi was partly supported by an NSERC Discovery Grant RGPIN-2022-05199.

References

AAFC (Agriculture and Agri-Food Canada). (2015). *Land use 2010*. Agriculture and Agri-Food Canada. Retrieved from https://open.canada.ca/data/en/dataset/fa84a70f-03ad-4946-b0f8-a3b481dd5248

ABMI (Alberta Biodiversity Monitoring Institute). (2018). *Human Footprint Inventory 2016*. Alberta Biodiversity Monitoring Institute. Retrieved from https://ftp-public.abmi.ca/GISData/HumanFootprint/2016/HFI2016_Metadata.pdf

Alameddine, I., Cha, Y., & Reckhow, K. H. (2011). An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics. *Environmental Modelling & Software*, *26*(2), 163–172. https://doi.org/10.1016/j.envsoft.2010.08.007

Anbumozhi, V., Radhakrishnan, J., & Yamaji, E. (2005). Impact of riparian buffer zones on water quality and associated management considerations. *Ecological Engineering*, *24*(5), 517–523. https://doi.org/10.1016/j.ecoleng.2004.01.007

Backer, L. C., Manassaram-Baptiste, D., LePrell, R., & Bolton, B. (2015). Cyanobacteria and algae blooms: Review of health and environmental data from the harmful algal bloom-related illness surveillance system (HABISS) 2007–2011. *Toxins*, *7*(4), 1048–1064. https://doi.org/10.3390/toxins7041048

Baron, J. S., Hall, E., Nolan, B., Finlay, J., Bernhardt, E., Harrison, J., et al. (2013). The interactive effects of excess reactive nitrogen and climate change on aquatic ecosystems and water resources of the United States. *Biogeochemistry*, *114*(1–3), 71–92. https://doi.org/10.1007/s10533-012-9788-y

Berger, S. A., Diehl, S., Kunz, T. J., Albrecht, D., Oucible, A. M., & Ritzer, S. (2006). Light supply, plankton biomass, and seston stoichiometry in a gradient of lake mixing depths. *Limnology and Oceanography*, *51*(4), 1898–1905. https://doi.org/10.4319/lo.2006.51.4.1898

Bižić, M., Klintzsch, T., Ionescu, D., Hindiyeh, M., Günthel, M., Muro-Pastor, A. M., et al. (2020). Aquatic and terrestrial cyanobacteria produce methane. *Science Advances*, *6*(3), eaax5343. https://doi.org/10.1126/sciadv.aax5343

Butterwick, C., Heaney, S., & Talling, J. (2005). Diversity in the influence of temperature on the growth rates of freshwater algae, and its ecological relevance. *Freshwater Biology*, *50*(2), 291–300. https://doi.org/10.1111/j.1365-2427.2004.01317.x

Canfield, D. E., & Hodgson, L. M. (1983). Prediction of Secchi disc depths in Florida lakes: Impact of algal biomass and organic color. *Hydrobiologia*, *99*(1), 51–60. https://doi.org/10.1007/bf00013717

Carpenter, S. R. (2005). Eutrophication of aquatic ecosystems: Bistability and soil phosphorus. *Proceedings of the National Academy of Sciences*, *102*(29), 10002–10005. https://doi.org/10.1073/pnas.0503959102

Cheng, F. Y., & Basu, N. B. (2017). Biogeochemical hotspots: Role of small water bodies in landscape nutrient processing. *Water Resources Research*, *53*(6), 5038–5056. https://doi.org/10.1002/2016wr020102

Chorus, I. & Bartram, J. (Eds.). (1999). *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring, and management*. WHO.

Cooke, G. D., Welch, E. B., Peterson, S., & Nichols, S. A. (2016). *Restoration and management of lakes and reservoirs*. CRC Press.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845. https://doi.org/10.2307/2531595

Dolman, A. M., Rucker, J., Pick, F. R., Fastner, J., Rohrlack, T., Mischke, U., & Wiedner, C. (2012). Cyanobacteria and cyanotoxins: The influence of nitrogen versus phosphorus. *PLoS One*, *7*(6), e38757. https://doi.org/10.1371/journal.pone.0038757

Feki-Sahnoun, W., Njah, H., Hamza, A., Barraj, N., Mahfoudi, M., Rebai, A., & Hassen, M. B. (2018). Using general linear model, Bayesian networks and Naive Bayes classifier for prediction of Karenia selliformis occurrences and blooms. *Ecological Informatics*, *43*, 12–23. https://doi.org/10.1016/j.ecoinf.2017.10.017

Franzin, A., Sambo, F., & Di Camillo, B. (2017). bnstruct: An R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics*, *33*(8), 1250–1252. https://doi.org/10.1093/bioinformatics/btw807

Graham, M., Cook, J., Graydon, J., Kinniburgh, D., Nelson, H., Pilieci, S., & Vinebrooke, R. (2018). High-resolution imaging particle analysis of freshwater cyanobacterial blooms. *Limnology and Oceanography: Methods*, *16*(10), 669–679. https://doi.org/10.1002/lom3.10274

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, *31*(15), 2595–2597. https://doi.org/10.1093/bioinformatics/btv153

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*, 627.

Heggerud, C. M., Wang, H., & Lewis, M. A. (2020). Transient dynamics of a stoichiometric cyanobacteria model via multiple-scale analysis. *SIAM Journal on Applied Mathematics*, *80*(3), 1223–1246. https://doi.org/10.1137/19m1251217

Heggerud, C. M., Wang, H., & Lewis, M. A. (2022). Coupling the socio-economic and ecological dynamics of cyanobacteria: Single lake and network dynamics. *Ecological Economics*, *194*, 107324. https://doi.org/10.1016/j.ecolecon.2021.107324

Heggerud, C. M., Xu, J., Wang, H., Lewis, M. A., Zurawell, R. W., Loewen, C. J., et al. (2023). CB-prediction [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10109224

Ho, J. C., Michalak, A. M., & Pahlevan, N. (2019). Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, *574*(7780), 667–670. https://doi.org/10.1038/s41586-019-1648-7

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M., & Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology*, *16*(8), 471–483. https://doi.org/10.1038/s41579-018-0040-1

Jackson-Blake, L. A., Clayer, F., Haande, S., Sample, J. E., & Moe, S. J. (2022). Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network. *Hydrology and Earth System Sciences*, *26*(12), 3103–3124. https://doi.org/10.5194/hess-26-3103-2022

Jebara, T. (2001). *Discriminative, generative and imitative learning* (Ph.D. thesis). Media Laboratory.

Jiang, P., Liu, X., Zhang, J., Te, S. H., Gin, K. Y. H., Van Fan, Y., et al. (2021). Cyanobacterial risk prevention under global warming using an extended Bayesian network. *Journal of Cleaner Production*, *312*, 127729. https://doi.org/10.1016/j.jclepro.2021.127729

Keatley, B. E., Bennett, E. M., MacDonald, G. K., Taranu, Z. E., & Gregory-Eaves, I. (2011). Land-use legacies are important determinants of lake eutrophication in the Anthropocene. *PLoS One*, *6*(1), e15913. https://doi.org/10.1371/journal.pone.0015913

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Kosten, S., Huszar, V. L., Bécares, E., Costa, L. S., van Donk, E., Hansson, L. A., et al. (2012). Warmer climates boost cyanobacterial dominance in shallow lakes. *Global Change Biology*, *18*(1), 118–126. https://doi.org/10.1111/j.1365-2486.2011.02488.x

Krause-Jensen, D., Carstensen, J., Dahl, K., Bäck, S., & Neuvonen, S. (2009). Testing relationships between macroalgal cover and Secchi depth in the Baltic Sea. *Ecological Indicators*, *9*(6), 1284–1287. https://doi.org/10.1016/j.ecolind.2009.02.010

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lang, M., Li, P., & Yan, X. (2013). Runoff concentration and load of nitrogen and phosphorus from a residential area in an intensive agricultural watershed. *Science of the total Environment*, *458*, 238–245. https://doi.org/10.1016/j.scitotenv.2013.04.044

Loewen, C. J., Vinebrooke, R. D., & Zurawell, R. W. (2021). Quantifying seasonal succession of phytoplankton trait-environment associations in human-altered landscapes. *Limnology and Oceanography*, *66*(4), 1409–1423. https://doi.org/10.1002/lno.11694

Loewen, C. J., Wyatt, F. R., Mortimer, C. A., Vinebrooke, R. D., & Zurawell, R. W. (2020). Multiscale drivers of phytoplankton communities in north-temperate lakes. *Ecological Applications*, *30*(5), e02102. https://doi.org/10.1002/eap.2102

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., et al. (2013). Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences*, *110*(16), 6448–6452. https://doi.org/10.1073/pnas.1216006110

Moe, S. J., Haande, S., & Couture, R. M. (2016). Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. *Ecological Modelling*, *337*, 330–347. https://doi.org/10.1016/j.ecolmodel.2016.07.004

Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, *28*(3/4), 223–238. https://doi.org/10.1504/ijep.2006.011208

Paerl, H. W. (2014). Mitigating harmful cyanobacterial blooms in a human- and climatically-impacted world. *Life*, *4*, 988–1012. https://doi.org/10.3390/life4040988

Paerl, H. W., Gardner, W. S., Havens, K. E., Joyner, A. R., McCarthy, M. J., Newell, S. E., et al. (2016). Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients. *Harmful Algae*, *54*, 213–222. https://doi.org/10.1016/j.hal.2015.09.009

Paerl, H. W., & Huisman, J. (2008). Blooms like it hot. *Science*, *320*(5872), 57–58. https://doi.org/10.1126/science.1155398

Paerl, H. W., & Huisman, J. (2009). Climate change: A catalyst for global expansion of harmful cyanobacterial blooms. *Environmental Microbiology Reports*, *1*, 27–37. https://doi.org/10.1111/j.1758-2229.2008.00004.x

Paerl, H. W., & Otten, T. G. (2013). Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microbial Ecology*, *65*(4), 995–1010. https://doi.org/10.1007/s00248-012-0159-y

Panchal, G., Ganatra, A., Kosta, Y., & Panchal, D. (2011). Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, *3*, 332–337. https://doi.org/10.7763/ijcte.2011.v3.328

Qu, M., Lefebvre, D. D., Wang, Y., Qu, Y., Zhu, D., & Ren, W. (2014). Algal blooms: Proactive strategy. *Science*, *346*(6206), 175–176. https://doi.org/10.1126/science.346.6206.175-b

Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, *7*(3), 205–229. https://doi.org/10.1145/65943.65945

Ramazi, P., Kunegel-Lion, M., Greiner, R., & Lewis, M. A. (2021a). Exploiting the full potential of Bayesian networks in predictive ecology. *Methods in Ecology and Evolution*, *12*(1), 135–149. https://doi.org/10.1111/2041-210x.13509

Ramazi, P., Kunegel-Lion, M., Greiner, R., & Lewis, M. A. (2021b). Predicting insect outbreaks using machine learning: A mountain pine beetle case study. *Ecology and Evolution*, *11*(19), 13014–13028. https://doi.org/10.1002/ece3.7921

Régnière, J., Saint-Amant, R., Béchard, A., & Moutaoufik, A. (2014). *BioSIM 10: User's manual*. Laurentian Forestry Centre Québec.

Reynolds, C. S., Oliver, R. L., & Walsby, A. E. (1987). Cyanobacterial dominance: The role of buoyancy regulation in dynamic lake environments. *New Zealand Journal of Marine & Freshwater Research*, *21*(3), 379–390. https://doi.org/10.1080/00288330.1987.9516234

Rigosi, A., Hanson, P., Hamilton, D. P., Hipsey, M., Rusak, J. A., Bois, J., et al. (2015). Determining the probability of cyanobacterial blooms: The application of Bayesian networks in multiple lake systems. *Ecological Applications*, *25*(1), 186–199. https://doi.org/10.1890/13-1677.1

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 1–8. https://doi.org/10.1186/1471-2105-12-77

Rodríguez-Gallego, L., Achkar, M., Defeo, O., Vidal, L., Meerhoff, E., & Conde, D. (2017). Effects of land use changes on eutrophication indicators in five coastal lagoons of the southwestern atlantic ocean. *Estuarine, Coastal and Shelf Science*, *188*, 116–126. https://doi.org/10.1016/j.ecss.2017.02.010

Rong, X. (2014). Package 'deepnet'. Retrieved from https://cran.microsoft.com/snapshot/2015-01-15/web/packages/deepnet/deepnet.pdf

Rousso, B. Z., Bertone, E., Stewart, R., & Hamilton, D. P. (2020). A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research*, *182*, 115959. https://doi.org/10.1016/j.watres.2020.115959

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(4), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Steffen, M. M., Davis, T. W., McKay, R. M. L., Bullerjahn, G. S., Krausfeldt, L. E., Stough, J. M., et al. (2017). Ecophysiological examination of the Lake Erie *Microcystis* bloom in 2014: Linkages between biology and the water supply shutdown of Toledo, OH. *Environmental Science & Technology*, *51*(12), 6745–6755. https://doi.org/10.1021/acs.est.7b00856

Taranu, Z. E., Zurawell, R. W., Pick, F., & Gregory-Eaves, I. (2012). Predicting cyanobacterial dynamics in the face of global change: The importance of scale and environmental context. *Global Change Biology*, *18*(12), 3477–3490. https://doi.org/10.1111/gcb.12015

Trainer, V. L., Moore, S. K., Hallegraeff, G., Kudela, R. M., Clement, A., Mardones, J. I., & Cochlan, W. P. (2020). Pelagic harmful algal blooms and climate change: Lessons from nature's experiments with extremes. *Harmful Algae*, *91*, 101591. https://doi.org/10.1016/j.hal.2019.03.009

Verhoeven, J. T., Arheimer, B., Yin, C., & Hefting, M. M. (2006). Regional and global concerns over wetlands and water quality. *Trends in Ecology & Evolution*, *21*(2), 96–103. https://doi.org/10.1016/j.tree.2005.11.015

Vymazal, J. (2010). Constructed wetlands for wastewater treatment. *Water*, *2*(3), 530–549. https://doi.org/10.3390/w2030530

Wang, H., Smith, H. L., Kuang, Y., & Elser, J. J. (2007). Dynamics of stoichiometric bacteria-algae interactions in the epilimnion. *SIAM Journal on Applied Mathematics*, *68*(2), 503–522. https://doi.org/10.1137/060665919

Whitton, B. A. (Ed.). (2012). *Ecology of cyanobacteria II: Their diversity in space and time*. Springer Science & Business Media.

Wilhelm, S. W., Farnsley, S. E., LeCleir, G. R., Layton, A. C., Satchwell, M. F., DeBruyn, J. M., et al. (2011). The relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu (Lake Tai), China. *Harmful Algae*, *10*(2), 207–215. https://doi.org/10.1016/j.hal.2010.10.001

Zhao, C., Shao, N., Yang, S., Ren, H., Ge, Y., Feng, P., et al. (2019). Predicting cyanobacteria bloom occurrence in lakes and reservoirs before blooms occur. *Science of the Total Environment*, *670*, 837–848. https://doi.org/10.1016/j.scitotenv.2019.03.161