Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 6, 2161 doi: 10.2166/hydro.2023.229

Predicting cyanobacteria abundance with Bayesian zero-inflated models

Yirao Zhang* and Nicolas M. Peleato

School of Engineering, Faculty of Applied Science, The University of British Columbia Okanagan, 1137 Alumni Ave, Kelowna, BC, Canada
*Corresponding author. E-mail: yirao.zhang@outlook.com

YZ, 0000-0003-3499-945X

ABSTRACT

Cyanobacterial blooms are a persistent concern to water management and treatment, with blooms potentially causing the release of toxins and degrading water quality. However, previous models have not considered the zero inflation of cyanobacteria count data. Typically, a relatively large proportion of measured count data are zeros or non-detects of cyanobacteria, representing either no cyanobacteria was present or the cell number was too low to be detected. Commonly used Poisson and negative binomial models for count data underestimate the probability of zero data, making these models less reliable. This study proposes a Bayesian approach to fit the cyanobacteria abundance data with mixture models that handle zero-inflated data. Predictor variables considered included weather and water quality measures that can easily be obtained day-to-day. The optimal model (zero-inflated negative binomial) was used to predict cyanobacteria alert levels on a separate test set. The ability to predict narrow alert levels was limited, however, 76% accuracy was achieved in predicting cyanobacteria counts above or below 1,000 cells/mL. Parameter estimates were highly variable and demonstrated that complex and uncertain factors influence cyanobacteria count predictions. The modelling approach can be applied to a wide range of environmental problems where zero-inflated data is common.

Key words: Bayesian modelling, cyanobacteria, environmental modelling, water management, zero-inflated

HIGHLIGHTS

- Bayesian mixture models were used to model zero-inflated cyanobacteria count data.
- A Bayesian variable selection method was applied to select important variables.
- A zero-inflated model achieved 76% accuracy in predicting binary alert levels.
- Bayesian framework produced probabilistic categorization of alert levels.
- The model is well suited for management of complex systems with high uncertainty.

1. INTRODUCTION

Cyanobacteria are photosynthetic microorganisms that can result in degraded freshwater quality and threaten human health. Cyanobacterial blooms can significantly increase turbidity, result in dissolved oxygen depletion due to the biological degradation of cyanobacteria biomass, and produce unpleasant taste and odour compounds (Huisman *et al.* 2018). Furthermore, some species can release toxins such as microcystins, nodularins, cylindrospermopsin, anatoxins, and saxitoxins (Catherine *et al.* 2013). It has previously been observed that a significant positive relation exists between non-alcoholic liver disease and large-scale blooms associated with toxin release (harmful algal blooms or CyanoHABs) (Zhang *et al.* 2015). Furthermore, associations between drinking surface water from cyanobacteria contaminated water bodies and a higher incidence of colorectal cancer have been noted (Lee *et al.* 2017). CyanoHABs also pose severe problems for ecological systems. Even low concentrations of microcystin-LR (5 μg/L) and microcystins (50 μg/L) have been found to impact fish growth and survival rates. At high concentrations of microcystins (>10 mg/L), morphological effects on fish have been observed (Oberemm *et al.* 1999). In addition, the accumulation of microcystins and cyanotoxins through the food web is a threat to human health (Bownik 2016). Based on analysis of lake sediments over the past 200 years, data show that cyanobacteria have increased significantly, with the most rapid growth in blooms occurring from 1945 until the present (Taranu *et al.* 2015).

CyanoHABs are caused or promoted by a combination of environmental factors, with strong associations with several anthropogenic and natural processes. Agricultural activities can increase nitrogen and phosphorus input into the water system, promoting cyanobacteria growth (O'neil *et al.* 2012). Climate change impacts are also likely to increase the occurrence of blooms in the future (Chapra *et al.* 2017). Higher water temperatures stimulate the growth of cyanobacteria, since their optimal growth rate is often reached at temperatures above 25 °C (Thomas & Litchman 2016). Cyanobacteria are carbon-fixing bacteria that rely on a CO_2 concentrating mechanism, and therefore rising concentrations of CO_2 in the atmosphere and water bodies may also promote blooms (Verspagen *et al.* 2014). Elevated pH is also known to reduce the energy cost of the CO_2 concentrating process, with higher efficiencies observed in acidic environments (Mangan *et al.* 2016).

Cyanobacteria bloom density is usually counted with a mechanical or electronic counter using an inverted microscope following sedimentation in a chamber or filtration (Chorus & Welker 2021). Cell counting is a labour-intensive, time-consuming, and expensive method that limits the extent and frequency of monitoring campaigns. As such, there is a need for methods that can enumerate or estimate cyanobacterial levels rapidly and preferably without the need for sampling. Several studies have developed models to fit count data and make predictions of day-to-day counts based on easy-to-measure parameters. Dzialowski *et al.* (2009) attempted to build a linear regression model for predicting the cyanobacteria abundance and toxins in five reservoirs in Kansas, USA. However, their results suggest that simple linear models could not accurately predict cyanobacteria counts (Dzialowski *et al.* 2009). Pyo *et al.* (2020) utilized a convolutional neural network applied to the output of a spatial fluid dynamics model of cyanobacteria abundance, which achieved good short-term prediction of microcystis. Zhao *et al.* (2019) put forward a species identification model and analysed the dominant species using canonical correspondence analysis (CCA). The model was used to identify major driving factors, including water temperature, pH, total phosphorus, ammonia nitrogen, chemical oxygen demand and dissolved oxygen, and predict the risk of algal blooms. Harris & Graham (2017) developed 12 linear and non-linear models to predict cyanobacteria abundance, microcystin and geosmin in a reservoir. Support vector machines, random forests, boosted trees, and cubist modelling approaches were observed to have the best performance. However, all models underestimated cyanobacteria abundance, and none of the models predicted peak bloom events or the highest counts.

A common challenge with modelling cyanobacteria abundance is the innate imbalance in monitoring datasets. A significant excess number of zero counts is typical and may have resulted from either failure to detect cyanobacteria or an actual absence of cyanobacteria. Although data balancing techniques, such as SMOTE (Synthetic Minority Oversampling Technique), have been proven to be effective in addressing imbalanced data, they also come with disadvantages, such as increase the risking of overfitting (He & Garcia 2009), and the generated data may not fully preserve the true distribution of the minority class (Japkowicz & Stephen 2002). Poisson and negative binomial distributions are commonly used for modelling count data, but they cannot account for the information contained in the excess proportion of zeros. Several mixture models have been proposed to consider better high numbers of zero counts: zero-inflated models and hurdle models. Zero-inflated models assume zeros are generated by a Bernoulli distribution with probability P and negative binomial (or Poisson) distribution with probability $1 - P$ (Lambert 1992). In hurdle models, the zeros and non-zero values are generated separately by a Bernoulli distribution and negative binomial (or Poisson) distribution (Min & Agresti 2005). Both hurdle and zero-inflated models have been used in environmental and ecological fields of study. Wenger & Freeman (2008) showed an improved fit of zero-inflated models to duck species abundance and stream fish abundance. Cha *et al.* (2014) developed a Bayesian hurdle Poisson model for predicting cyanobacteria abundance in Lake Paldang, Korea. However, comparisons between hurdle models and zero-inflated models were not made, and Poisson models cannot accommodate a common challenge of greater variability in observed data than expected for a given model (overdispersion) (Cha *et al.* 2014).

This study presents a Bayesian approach to fit cyanobacteria data with a negative binomial model, zero-inflated negative binomial (ZINB) model, and hurdle negative binomial model to address challenges with inflated zero counts. It is hypothesized that through the novel use of zero-inflated models for this application, the elevated zero counts inherent in the majority of cyanobacteria abundance data can be accounted for and model fit will be improved. Additionally, a Bayesian framework was used to present abundance predictions as distributions rather than point estimates, allowing for a more direct interpretation of uncertainty. Through these two key aspects of the presented models, the aim is to improve the integratability of models in water management by accounting for expected data distributions and emphasizing the need for knowledge of uncertainty in predictions of environmental systems. The fit of each model is compared to select an optimal model. Predictions from the optimal model are then classified according to Australian Management Strategies for Cyanobacteria (Newcombe *et al.* 2010) to assess the capabilities of the presented approach to identify cyanobacteria levels used in

water management. The application of the selected model integrated into a Bayesian framework and utilizing the predictive distribution of each prediction obtained from MCMC sampling to assign the prediction into predefined categories are novel and can achieve higher accuracy compared to the regression method. The established model was also used to assess the importance and impact of environmental variables on the probability of cyanobacteria blooms. We employed the state-of-the-art projection predictive variable selection for generalized linear models which has shown superior performance to competing variable selection methods (Catalina *et al.* 2020), and validate the selected model through the posterior predictive checks (PPCs), which are useful tools to inspect the discrepancies between real and predicted/simulated data. The proposed approach for predicting cyanobacterial blooms offers clear advancements over existing approaches: unlike other methods that rely on traditional variable selection approaches, our approach incorporates a Bayesian variable selection method to the framework, which allows for more effective identification of key site-specific predictors under Bayesian framework. The approach also account for the distinctive zero-inflation characteristics of such imbalanced datasets and significantly improves the predictive performance, which is especially important when large or extensive datasets are not accessible. Finally, the Bayesian framework enables us to incorporate prior beliefs and expert knowledge into the analysis and capture the uncertainty of the prediction, providing a more informed and robust understanding and decision making. The whole process has not been used to resolve water quality issues, and the developed framework is appropriate to resolve a wide range of problems of predicting the classification of imbalance data in environmental and ecological fields. The experimental results validate the effectiveness of our methods, highlighting their potential for practical applications and further research in this field.

2. METHODS

2.1. Study site and data source

Data used in this study was collected from a eutrophic lake, Cheney Reservoir (37°45′35″N, 97°50′06″W), the main water supply for Wichita, Kansas, USA (Christensen *et al.* 2006). The reservoir has experienced frequent cyanobacterial blooms, presence of microcystin, and taste-and-odor problems. In part, this could be due to the shallow depth (average depth = 6.1 m) and persistent winds that cause maximal turbulence and a resulting turbid environment. Among the 185 samples in the dataset, 34 samples indicate zero counts of cyanobacteria (18.4% of the data).

The dataset that included water quality variables was obtained from the United States Geological Survey (USGS) from 2002 to 2015 (US Geological Survey 2015). The station code is 07144790. Precipitation, solar radiation, and wind speed were obtained from NASA Power project. The original dataset included nine variables: temperature, pH, total phosphorous, total nitrogen, chlorophyll *a* (Chl *a*), all sky insolation incident on a horizontal surface (i.e. solar radiation), wind, turbidity, and precipitation (Table 1). Although Chl *a* and turbidity are not the cause of cyanobacterial blooms, they are parameters that are expected to be correlated with the presence of cyanobacteria, and therefore can be used as predictors. Wind, temperature, and precipitation were also considered to explore the relationship between weather conditions and cyanobacterial blooms. Prior to modelling, variables were further selected by projection predictive inference, a Bayesian approach for model selection and decision making.

Table 1 | Selected variables used to build initial models

Variable	Abbreviation	Units
Total phosphorous	TP	mg/L as P
pH	pH	NA
Temperature	Temp	°C
Chlorophyll *a*	Chl *a*	μg/L
All sky insolation incident on a horizontal surface	Solar radiation	Wh/m^2
Wind	Wind	m/s
Total nitrogen	TN	mg/L as N
Precipitation	Precipitation	mm
Turbidity	Turb	FNU

2.2. Mixture models for zero-inflated count data

2.2.1. Zero-inflated negative binomial model

The ZINB model (Lambert 1992) is a mixture model consisting of a Bernoulli distribution and an untruncated negative binomial distribution. In a ZINB model, zeros are generated in two processes: the first binomial process accounts for the count being zero or some value, and the second negative binomial distribution generates counts, in which zeros are included. By combining these two processes, the ZINB model accounts for both the real zeros (the first binomial process) and measurement zeros. For some random variable Y, the ZINB model can be written as:

$$p(Y = y_i)\begin{cases} \pi + (1-\pi)f(y_i = 0), & \text{if } y_i = 0 \\ (1-\pi)f(y_i), & \text{if } y_i > 0 \end{cases}$$

where π is the parameter denoting the probability of zeros in a binomial distribution. $f(y_i)$ is the probability density function of the negative binomial distribution.

We use an alternative negative binomial distribution function:

$$p(y|\ \mu,\ \ \phi) = \binom{y+\phi-1}{y}\left(\frac{y}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}$$

where $\mu \in R^{+}$, $\phi \in\ R^{+}$, and $y \in\ N$. The inverse of parameter ϕ controls the overdispersion, which is scaled by the square of μ.

The mean and variance of the variable $y \sim NB(y \mid \mu,\ \phi)$ are:

$$E[Y] = \mu$$

$$Var(Y) = \mu + \frac{\mu^2}{\phi}$$

In a generalized linear model, the canonical link functions of the binomial and negative binomial models are:

$$\log(\mu) = BX$$

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = AX'$$

$$B = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_n \end{bmatrix}$$

$$A = \begin{bmatrix} \alpha_1 \\ . \\ . \\ . \\ \alpha_n \end{bmatrix}$$

X, X' are the selected variables. In a zero-inflated model, X and X' can be different sets of variables. Here, X and X' are the same, which are the selected predictor variables, Y is the cyanobacteria counts.

2.2.2. Hurdle negative binomial model

In a hurdle model, there are two parts in control of count generation (Welsh *et al.* 1996). The first part decides the presence of a positive count, which is typically accomplished through logistic modelling. The second part, a truncated negative binomial

model, models the number of observations (non-zero value). The hurdle NB model can be written as:

$$p(Y = 0) = \pi$$

$$p(Y = y_i) = \frac{(1 - \pi)f(y_i)}{1 - f(0)}, \quad y_i \neq 0$$

The parameter and link functions of the alternative negative binomial distribution are the same as the above ZINB model.

2.3. Bayesian approach

Bayesian framework is an approach to model data and estimate parameters based on Bayes' theorem:

$$P(A, B|X) = \frac{P(X|A, B)P(A, B)}{P(X)}$$

In a Bayesian approach, parameter estimation workflow consists of three processes: first, a prior distribution of the parameters A, B in sections 2.2.1 and 2.2.2. $P(A, B)$ is determined based on available experience and knowledge. Second, the likelihood $P(X|A, B)$ of observed data is calculated using the parameters A, B. Finally, the likelihood and prior are combined to determine the posterior distribution $P(A, B|X)$, reflecting an updated representation of knowledge (van de Schoot *et al.* 2021).

Priors used typically fall into three categories: informative, weakly informative, and diffuse (Depaoli *et al.* 2020). For our generalized linear models, the prior distributions for the parameters were specified as weakly informative priors with a large spread. The specific information of prior parameters and prior sensitivity analysis can be found in Supplementary Table S1.

Once the posterior distribution of parameters is determined, sample observations can be drawn. However, the parameter distribution is high-dimensional and usually not a probability distribution we are familiar with, making exact inference intractable (Bishop 2006). Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970), are used to generate random samples from the target distribution.

Stan is a probabilistic programming language for Bayesian statistical inference written in C++. It provides a No-U-Turn sampler (NUTS) to obtain simulations from the user-specified posterior distribution (Carpenter *et al.* 2017). In this study, we have used the R package *rstan*, which provides an interface to Stan using R. Through *rstan,* we implemented mixture models such as zero-inflated and hurdle models for discrete distributions.

Convergence of MCMC chains can be diagnosed with trace plots and Gelman–Rubin diagnostic $\hat{R}$ (Brooks & Gelman 1998). Trace plots are helpful when identifying the burn-in process and the convergence of Markov chains. Gelman–Rubin statistic compares the total-within and between-chain variation to analyse the difference between multiple Markov chains. $\hat{R} = 1$ indicates good convergence. Practically, a 0.975 quantile for $\hat{R} \leq 1.2$ denotes convergence.

2.4. Model development, selection, and validation

A summary of the fitting and testing process is presented in Figure 1. Initially, the data were split into training and test sets, where the test set was only used to assess predictive performance. We took a 5-fold cross-validation with stratified random sampling to prevent an imbalance between training and test data and reduce the randomness in results. In our data, the common attribute is zero or non-zero cyanobacteria counts (18.4% of data was zero counts). As such, we stratified the data into two subgroups: zero and non-zero. In each subgroup, the data were randomly split into five equal folds and then one fold from each group were combined to form testing set with an equal proportion of zero and non-zero cyanobacteria counts, and the rest were combined to form training set. The test set contained 44 samples, and train data contained 141 samples. Data split and cross-validation procedure is illustrated in Figure 2.

After splitting the data into training/test sets, the most representative variables were selected by projection predictive inference. The selected variables were then used to build a Bayesian negative binomial model, a Bayesian zero-inflated model and a Bayesian hurdle negative binomial model. Model comparison and selection were achieved by leaving-one-out cross-validation (LOO-CV) and model validation (PPCs) for the best model.

When using generalized linear models to solve regression problems (e.g. binary and multinomial logistic regression), a threshold is commonly chosen as the decision rule. For example, in binary logistic regression, it is a general practice to

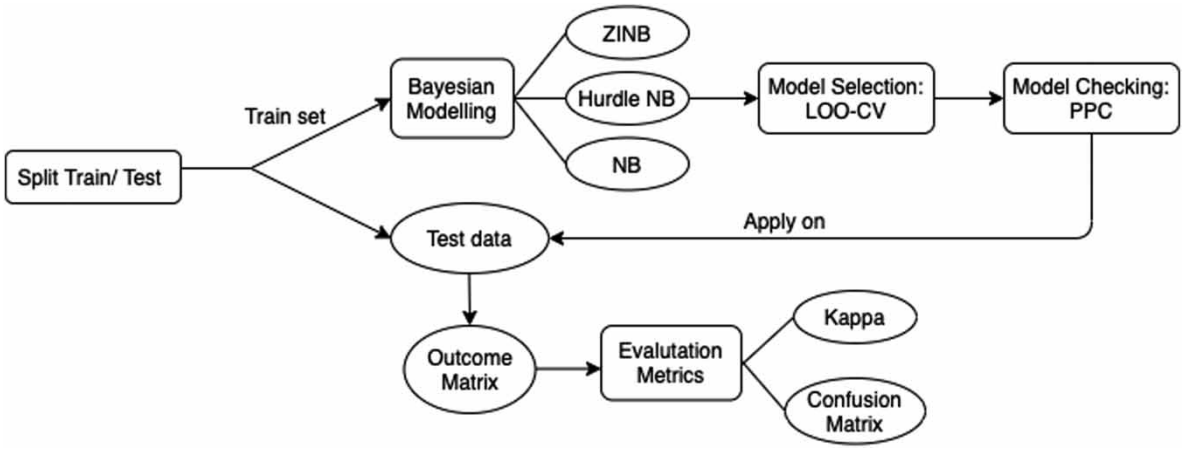

Figure 1 | Flowchart of modelling and application on cyanobacteria abundance prediction. ZINB, zero-inflated model; Hurdle NB, hurdle negative binomial model; NB, negative binomial model; LOO-CV, leave-one-out cross-validation; PPC, posterior predictive check.

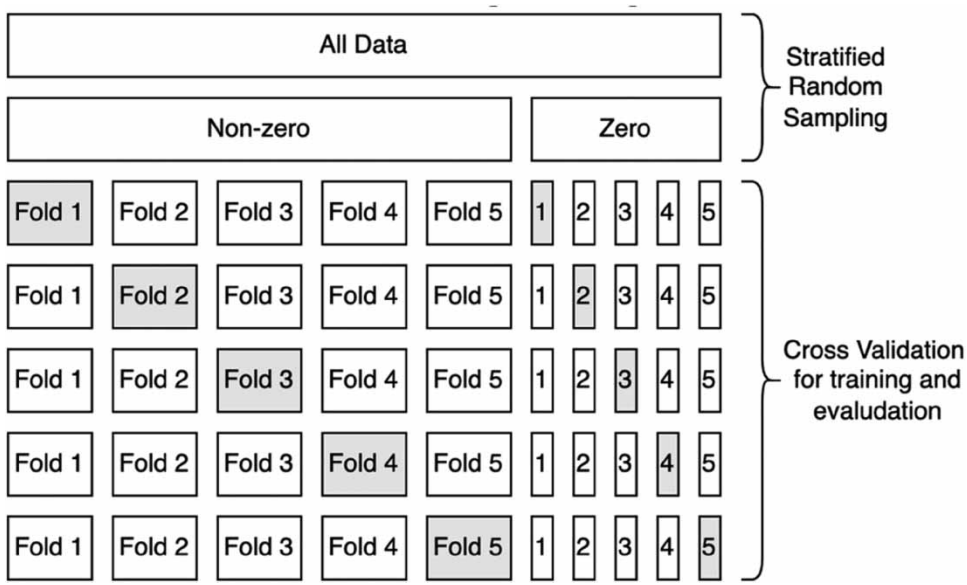

Figure 2 | Dataset split into non-zero and zero subgroups and 5-fold cross-validation. In both non-zero and zero subgroups, the data were split into five equally sized folds through random sampling. One fold from each subgroup was combined to form the testing sets for evaluation and the rest were used for training.

choose 0.5 as the threshold, but in practice, different thresholds can be manually selected for specific situations. If high discriminative accuracy is required for positive cases, a larger threshold can be chosen (Kuk *et al.* 2014). Traditional multinomial logistic regression is subject to large bias when dealing with imbalanced data and does not take the distribution of the data into account. Thus, in order to make the result more indicative, we approximated the probability distribution of the prediction points by the density distribution obtained by MCMC sampling, and assigned the predictions to alert levels according to the management strategies for cyanobacteria by Water Quality Research Australia (WQRA) (Table 2). The categorization process is analog to assigning the predicted class according to the posterior distribution and the probability threshold we set in advance. Finally, we applied the fitted model to our test set to generate predictions and classified the results.

2.4.1. Projection predictive inference

Projection predictive inference (Piironen *et al.* 2020) is a Bayesian variable selection method. Variable selection was carried out using the *projpred* package in R.

Table 2 | Alert levels for management of toxic cyanobacteria (WQRA)

Alert Level	Definition	Description
Safe	< 500 cell/mL	Safe for drinking water
Low	≥ 500 and < 2000 cell/mL	Detected at low levels
Medium	≥ 2000 and < 6500 cell/mL	Potential toxin to $1/3 \sim 1/2$ to guideline concentration
High	≥ 6500 and < 65000 cell/mL	Potential toxin greater than guideline concentration
Very high	≥ 65000 cell/mL	Potential toxin $10\times$ greater than guideline concentration

Initially, a model with all predictor variables was fitted and considered as the reference model. Sub-models are then fitted, initially with one variable, and then sequentially more variables are added. A model with the smallest subset of variables with an approximately similar fit to a full model was selected. In the forward search process, where variables are sequentially added, each step determines the variable that would result in the largest decrease in the discrepancy between the reference model and the sub-model. The sub-models were compared with the reference model by cross-validation prediction accuracy using LOO-CV to prevent overfitting. Fit or performance of all models was based on expected log predictive density (elpd) and RMSE. A model with a higher elpd value and smaller RMSE value is considered to be good.

$$\text{elpd} = E_f(\log[f(y_{\text{new}}|y)]) = \int \log[f(y_{\text{new}}|y)]f(y_{\text{new}})dy_{\text{new}}$$

2.4.2. Leave-one-out cross-validation

Several measures have been developed to compare the fit of different models, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), deviance information criterion (DIC), Watanabe–Akaike information criterion (WAIC), and LOO-CV. To measure the wider applicability of a statistical model, out-of-sample data are commonly used to evaluate its predictive power. However, usually we do not have access to extra data (if available, we would include that part as part of training data). As such, we estimated the predictive capability of any given model using the expected log point-wise predictive density (elppd) with a penalty term (Gelman *et al.* 2013).

$$\text{elppd} = \sum_{i=1}^{n} E_f(\log[p(y_{\text{new}}|y)])$$

Measures including AIC, BIC, DIC, and WAIC methods utilize all data to determine fit and therefore can be biased in assessment. Therefore, a LOO-CV (*loo* package in R) approach was used in order to determine model fit based on out-of-sample data. This can be extremely computationally expensive, especially if the dataset is large. However, with less than 200 samples, computation time can be ignored. In LOO-CV, a single sample from the dataset is removed to test the model and the remaining samples are used to train the model. The process is repeated n times (where n is the size of the dataset) so that each sample is considered.

From each iteration, the log predictive density (lpd) is evaluated by:

$$\text{lpd} = \log[p(y_i|y_{-i})]$$

where y_i denotes the ith data point, and y_{-i} denotes the rest data. After n times, the elppd can be estimated by:

$$\widehat{\text{elppd}} = \sum_{i=1}^{n} \log[p(y_i|y_{-i})]$$

2.4.3. Posterior predictive checks

Posterior predictive check is a classical approach to compare the test statistics $T(y)$ (arbitrary function of data) of the actual observed data and the data generated from the model with parameters sampled from the posterior predictive distribution (Berkhof *et al.* 2000).

The posterior predictive distribution can be written as:

$$\Pr(y^{\text{rep}}|y^{\text{obs}}) = \int P(y^{\text{rep}}|\theta)P(\theta|y^{\text{obs}})d\theta$$

where y^{rep} denotes the replicated data, and y^{obs} denotes the observed data.

The principle behind PPCs is that if a model provides a good fit to the data, the generated data would have a similar pattern (test statistics) with the observed data. Bayesian *p*-value (Posterior *p*-value) is a quantitative measurement of the goodness of fit. The *p*-value-like measure represents the probability that the test statistic (such as mean, maximum, minimum, and zero proportion) in the replicated (or predicted new observations) dataset exceeds that in the original data (or new observations).

$$\Pr(T(y^{\text{rep}}) \geq T(y^{\text{obs}})) = \int I(T(y^{\text{rep}}) \geq T(y^{\text{obs}})|y) \cdot p(y^{\text{rep}}|y^{\text{obs}})dy^{\text{rep}}$$

If the model provides a good fit, the Bayesian *p*-value should be around 0.5. A value close to 0 or 1 indicates that the model is a poor fit (Meng 1994).

For each simulation ($s = 1, \ldots, S$) of parameters from the posterior distribution, we have a n-dimensional vector of n predicted outcomes of y. Thus, the result is an $S \times N$ sized matrix of predicted outcomes from all simulations.

In doing PPCs, either the same predictors X or new observations of predictors could be used when building the model. In the latter case, the test statistics of predicted values y^{pre} and test statistics of the actual observed values y^{new} are compared. The *bayesplot* package in R was used for plotting PPCs.

3. RESULTS AND DISCUSSION

3.1. Variable selection

Initially, all nine variables were used to develop a generalized linear model to serve as a reference model. Variable selection was carried out by sequentially adding additional predictor variables. In each step, one additional variable is included (starting with no variables or only an intercept), and the elpd and RMSE of each model was calculated (Figure 3). The change in elpd or RMSE indicates the impact of that variable addition. The order of variables added is based on the maximizing fit and is therefore indicative of variable importance. The selection order decided by the algorithm was Chl *a*, temperature, turbidity, total phosphorus, solar radiation, wind, pH, total nitrogen, and precipitation (Figure 3).

Figure 3 shows that the first five variables were sufficient predictors as they result in a similar elpd and RMSE to the reference model (the final point includes all variables). The selected variables are consistent with prior knowledge of factors that can be used to determine cyanobacteria counts. Chl *a* is produced by cyanobacteria and, therefore, a strong indicator of abundance. Temperature promotes cyanobacterial growth (Thomas & Litchman 2016) and is expected to be a significant driver of blooms. Nutrients (nitrogen and phosphorous) stimulate the growth of cyanobacteria (O'neil *et al.* 2012). However, it is worth noting that only total phosphorous was identified as a variable of importance and total nitrogen had no impact on model fit. Previous studies indicate that the optimal mass-based ratio of total nitrogen to total phosphorus is 16:1 (Davidson *et al.* 2012). The variable selection indicates the reservoir was phosphorus-dominated, and the nitrogen concentration was either sufficient or very stable. However, the average mass-based ratio of total nitrogen to total phosphorus was 11:1 in the reservoir, suggesting that nitrogen should be limited. The unselected variables, wind and precipitation, have been previously reported to have effects on cyanobacterial blooms: strong winds can disrupt the water layers, leading to mixing and reducing the stability of the water column, which can change the turbidity and limit the availability of light and nutrients (Anderson *et al.* 2002). Precipitation, particularly rainfall, also has effects on cyanobacterial blooms: heavy rainfalls can result in nutrients runoff from land into water bodies (Paerl & Otten 2013), which can promote the growth of cyanobacteria and lead to blooms. It is therefore possible that the effects of wind and precipitation can be better represented by the selected variables of turbidity, radiation, and nutrients, and are not directly needed for modelling.

3.2. Model selection

Weakly informative priors were used for parameters, and four Markov chains were run for each model for 1,000 iterations, discarding the first 500 iterations as a burn-in process. Supplementary Figures S1–S3 present traceplots for parameters in NB,

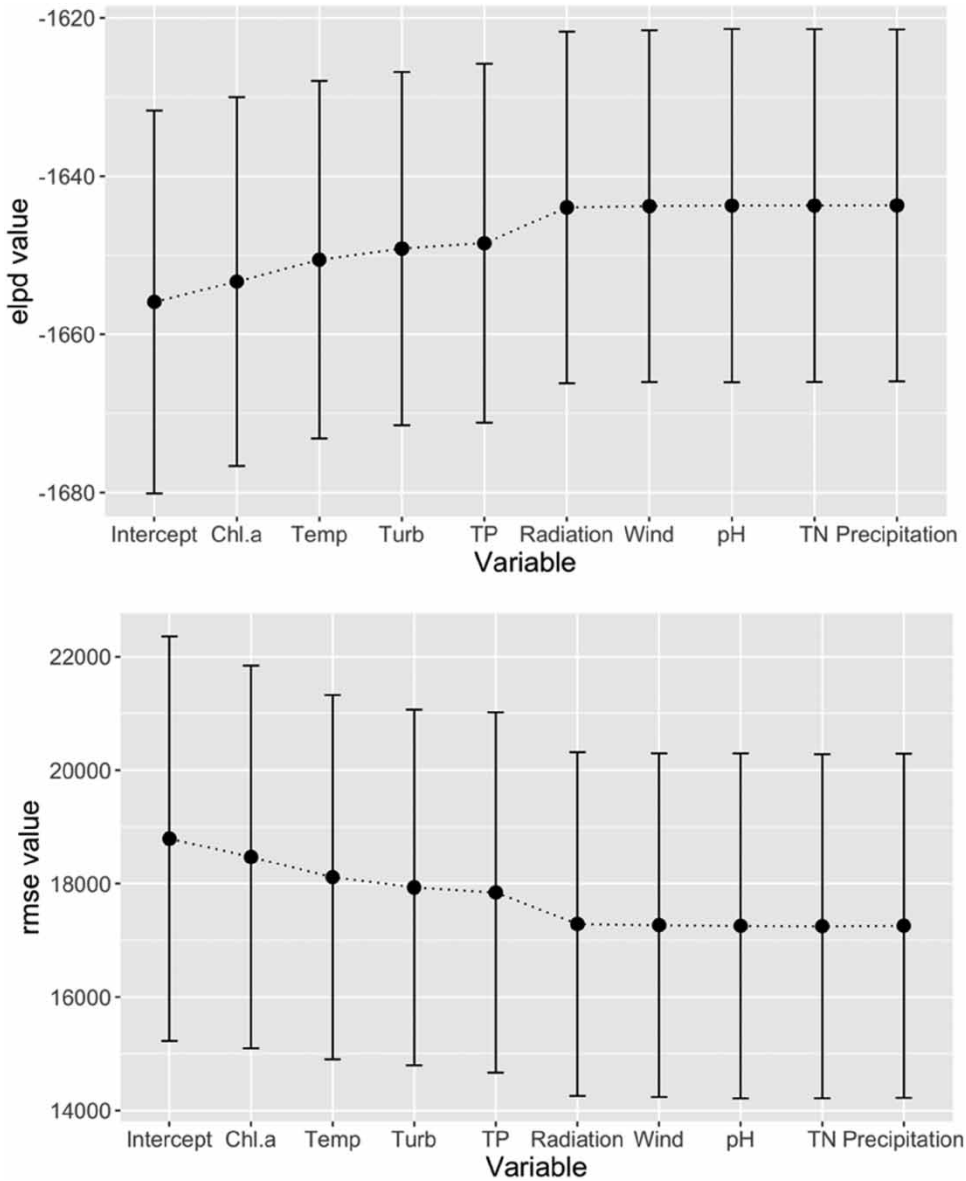

Figure 3 | Model elpd and RMSE from LOO-CV plotted as a function of stepwise addition of variables. Each point represents the performance of a model including all previous variables and the new variable indicated on the *y*-axis.

ZINB, and hurdle NB models. The overlapping of different chains indicates convergence. Furthermore, parameters from all three models have $\hat{R} < 1.003$, further suggesting convergence of each chain (Brooks & Gelman 1998).

After confirming the convergence of all MCMC chains, LOO-CV was applied to assess the strength of each modelling approach. Assessment of model strength was based on both elpd and standard error (SE) (Table 3). The difference in $\widehat{\text{elpd}}$ relative to the model with the largest $\widehat{\text{elpd}}$ (i.e. the ZINB model) can be used to consider the magnitude of difference between models. The significance of observed differences in elpd was determined by calculating *z*-scores and corresponding *p*-values of paired comparisons (Lambert 2018). Results indicate zero-truncated models (zero-inflated and hurdle models) were better than a negative binomial model ($p = 0.002$); however, the performance of ZINB and hurdle NB were comparable ($p = 0.14$).

While the fit between ZINB and hurdle NB were comparable, it should be considered that the mechanism of zero generation is different between them. In a zero-inflated model, zero counts may come from two sources: (a) the cell number is too

Table 3 | LOO-CV results to compare strength of model fits

Model	elpd difference	SE difference
ZINB	0.0	0.0
Hurdle NB	−1.2	2.7
NB	−25.3	8.9

Differences in elpd and standard error (SE) were calculated using the highest performing model (ZINB).

low to be by the enumeration method used and (b) the cell number was truly zero. Zero counts are assumed only to be caused by cell numbers below the detection limit in a hurdle model. Cyanobacteria are likely not present in the reservoir at some times, and therefore, the ZINB was chosen as the best model based on goodness of fit and the ability to consider true zero counts.

3.3. Model checking

3.3.1. Posterior predictive checks

PPCs are used to evaluate if the model fit is reasonable and identify potential differences between observed data and the fitted model. PPCs were initially run using the training set of data. We chose zero proportion as a test statistics for y and y^{rep}, which represents the proportion of zero values in the real observed data and the replicated data (predicted data for the same dataset) and calculated the Bayesian p-value. Bayesian p-values in this context indicate the probability that replicated data are not more extreme than the observed distribution (Gelman 2005). A Bayesian p-value close to 0.5 indicates a good fit, values approaching 0 indicate lack of fit, and values close to 1 indicate overfitting (Korner-Nievergelt *et al.* 2015).

The top left and right panels of Figure 4 show the density plot of the original training data (dark blue) and the density plot of the replicated data (light blue). The overlapping of observations distribution and replications distributions showing the model represents a good model fit. However, Figure 4 (top right) shows that the model tends to underestimate the zero proportion. The computed Bayesian p-value is 0.2, indicating that the model tends to underestimate the zero proportions. It is possible that the zero-inflated generalized linear models still cannot account for all zeros in the data due to not capturing non-linear relationships between cyanobacteria and predictor variables.

A 5-fold PPC cross-validation was applied to evaluate the model using out-of-bag samples. PPCs were repeated for each validation set and compared the test statistics for y^{new} and y^{pre}. The Bayesian p-value of the five validation sets were 0.43, 0.56, 0.32, 0.42, and 0.53 with an average of 0.45. One validation set is shown as an example in Figure 4 (Bayesian p-value $= 0.32$). The predicted y^{pre} and the actual new observations y^{new} overlap (Figure 4, bottom left), although there is a slight underestimation of zero proportion (Figure 4, bottom right). The difference in estimated zero proportions and p-values is likely due to the varying proportion of zero counts in each of the five validation sets. The Bayesian p-values of both replicated data and new data close to zero suggest that the linear model may be inadequate for the cyanobacteria growth model. Considering non-linear models, such as the dynamic phytoplankton model proposed by Malve *et al.* (2007) would add complexity but may also increase model fit.

3.4. Cyanobacteria alert level prediction

Predictions are produced by first sampling regression parameters from their respective distributions, followed by calculating cyanobacteria counts. Since 4 MCMC chains of 1,000 iterations were generated, and the first 500 iterations of each chain (burn-in) were discarded, the number of replicates for each prediction was 2,000. The advantage of Bayesian models is that instead of predicting a single value, the model presents a predictive probability distribution based on MCMC iterations. For example, the predictive distribution based on MCMC results of two data points in the test set are shown in Figure 5.

From Figure 5, it can be observed that even if the peaks of the predictive density do not fall precisely on the true observed value, the maximum predictive density may be approximately adjacent to the true value and the overall predictive density shifts. It is also of note that despite density being highest immediately adjacent to true predictions, there is non-zero probability of elevated cyanobacteria abundance. The Bayesian modelling approach allows for direct interpretation of this uncertainty and the uncertain nature of factors influencing cyanobacterial population dynamics to carry through to predictions.

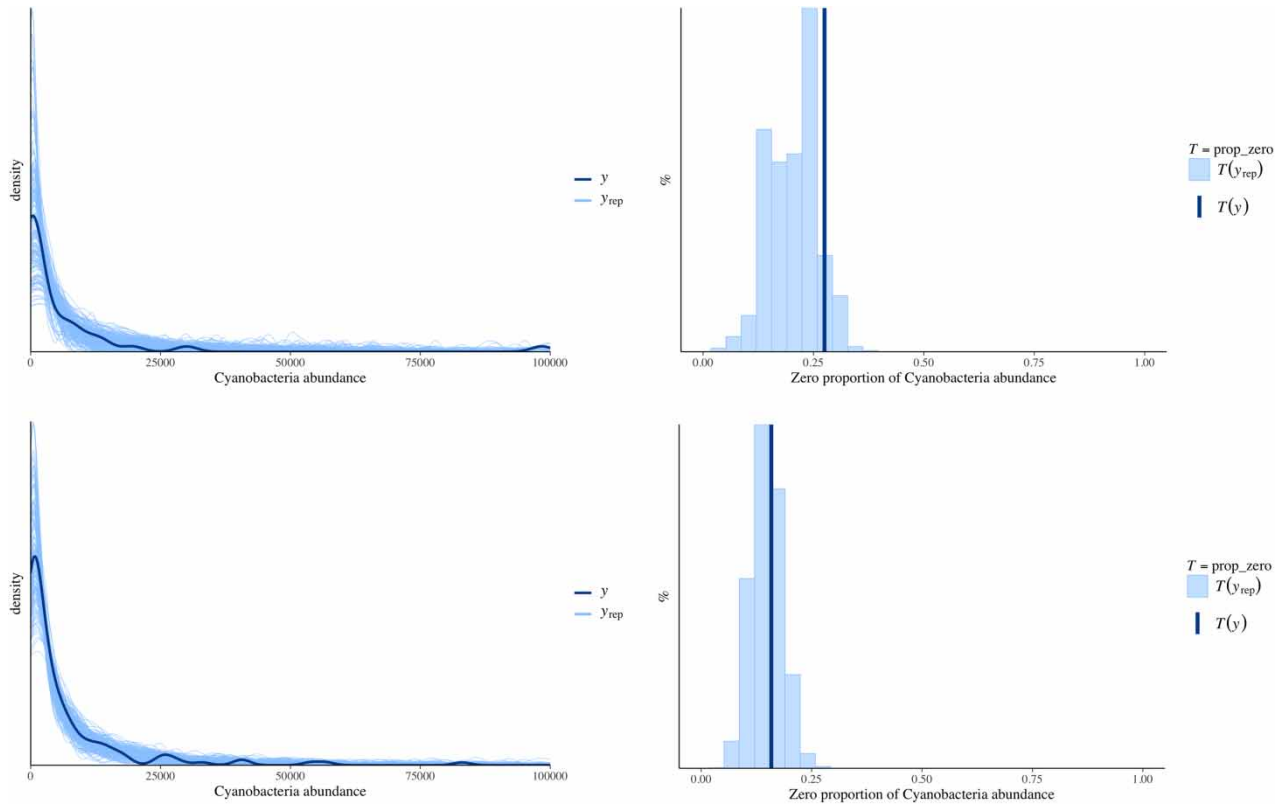

Figure 4 | Top left: Kernel density estimate of observations in the training set y (dark line) and replications y^{rep} (light line). Top right: Zero proportion as test statistics $T(y)$. The dark line is the zero proportion of observations in the training set. Light lines are the distribution of zero proportions of replicated data. Bottom left: Kernel density estimate of new observations y^{new} (dark line) and predictions y^{pre} (light lines). Bottom right: Zero proportion as test statistics $T(y)$. The dark line is the zero proportion of new observations. Light lines are the distribution of zero proportions of predicted data.

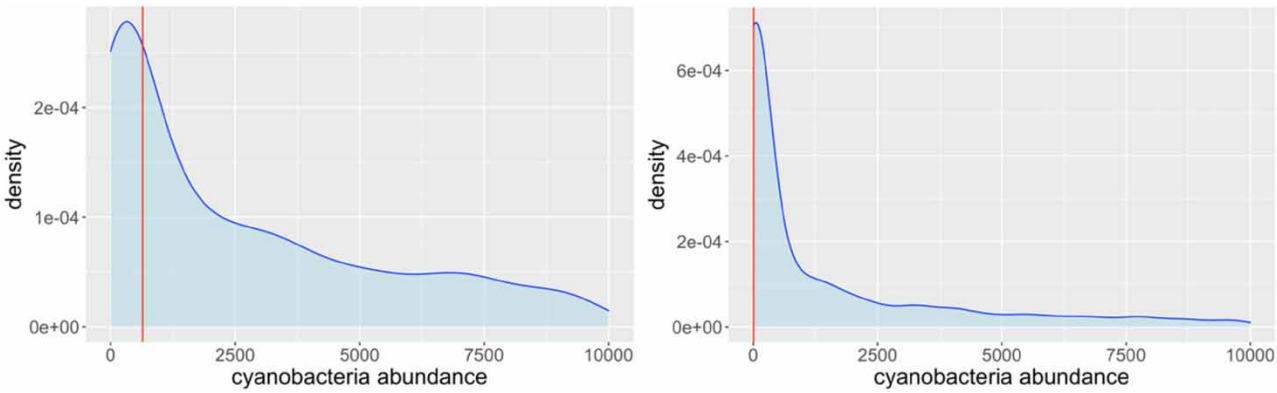

Figure 5 | Predictive density plot of two new observations. The red line indicates the true observed values, and the density is determined based on 2,000 MCMC replicates.

Predictive density was used to categorize predictions according to WQRA alert levels. By taking probability density in bins rather than point estimates, we account for the high levels of uncertainty in both the impacts of influencing factors and how to interpret risk from cyanobacteria abundance. Not all species will release toxins (Lee *et al.* 2015) and environmental

conditions such as temperature impacts toxin release (Walls *et al.* 2018). As such, management of surface waters often is in response to categorized levels of cell counts or other water quality parameters (Ibelings *et al.* 2014).

The predicted class was determined by the mode or most common predicted class based on probability density. The accuracies in each fold were 0.50, 0.32, 0.36, 0.32, and 0.45. The overall confusion matrix for multiclass prediction (all WQRA alert levels) is shown in Table 4a. The average accuracy was found to be 0.40, generally indicating poor performance. In particular, it was noted that the model performed poorly in predicting low or medium alert levels and predictions of safe levels dominated. The dominant safe level probabilities are evident from the figure inset in Table 4a.

Based on poor performance with narrow alert level bands, and generally better separation of 'high' vs. 'safe' levels, it was considered to reduce classification to a binary decision of potential toxin presence or not. We set the threshold to 1,000 cells/mL, corresponding to the middle of the low alert level in WQRA and associated with a level where toxin release may be possible. For this binary decision, the precision and recall were found to be 0.62 and 0.99, respectively. As such, on a more coarse level, the model performance improved and has potential for distinguishing conditions that could result in toxin presence (Table 4b). In particular, the binary decision approach did not under-predict alert (false negatives), and performance was high for correctly predicting counts greater than 1,000 cells/mL.

3.5. Influence of weather and water quality factors on cyanobacteria counts

The influence of various factors can also be observed from the kernel density estimates posterior distributions of the variable-specific coefficients (Figure 6). Chl *a* was found to have the largest positive coefficient, indicating a strong positive relationship with cyanobacteria counts. This was expected since cyanobacteria will produce Chl *a*, and this measure is often used as a surrogate for cell counts (Chaffin *et al.* 2018). The temperature coefficient is distributed above zero, implying a positive impact on the probability of a bloom. A positive relationship between temperature was anticipated based on a significant amount of literature highlighting increased growth with increasing temperature (Thomas & Litchman 2016; Rousso *et al.* 2020).

The coefficient of solar radiation is mainly distributed below zero, indicating a negative correlation with cyanobacteria levels. A negative relationship between radiation and cell counts could be explained by photobleaching of pigments in cyanobacteria, such as phycobiliproteins (Sinha *et al.* 2005) or by relative competitive advantages of cyanobacteria compared to other algal taxa under limited light conditions (LeBlanc Renaud *et al.* 2011). Long-term exposure to increasing light intensity and UV-B light in particular has resulted in decreased Chl *a* content and decreased cyanobacteria populations (Xue *et al.* 2005; Cirés *et al.* 2011). At high radiation levels (340 $\mu E\ m^{-2}\ s^{-1}$), the cyanobacteria growth rate was previously found to be 30% lower than at moderate radiation (60 $\mu E\ m^{-2}\ s^{-1}$) or low radiation levels (Cirés *et al.* 2011). However, it should be noted that radiation intensity and temperature are strongly correlated, and increasing solar radiation was expected to result in increased cyanobacteria levels due to a corresponding increase in temperature (Jöhnk *et al.* 2008).

Table 4 | (a) Confusion matrix for all WQRA levels along with figure depicting probability of each class for a given prediction, and (b) reduced confusion matrix for binary decisions > or <1,000 cells/mL

		Predicted				
(a) All WQRA levels		**Safe**	**Low**	**Medium**	**High**	**Very high**
Actual	Safe	21	1	24	14	3
	Low	9	0	9	14	0
	Medium	5	1	12	20	2
	High	1	0	4	40	1
	Very high	0	0	0	4	0
		Predicted				
(b) Binary decision		**Safe (<1,000 cells/mL)**		**Potential toxin presence (>= 1,000 cells/mL)**		
Actual	Safe (<1,000 cells/mL)	14		65		
	Potential toxin presence (>= 1,000 cells/mL)	1		105		

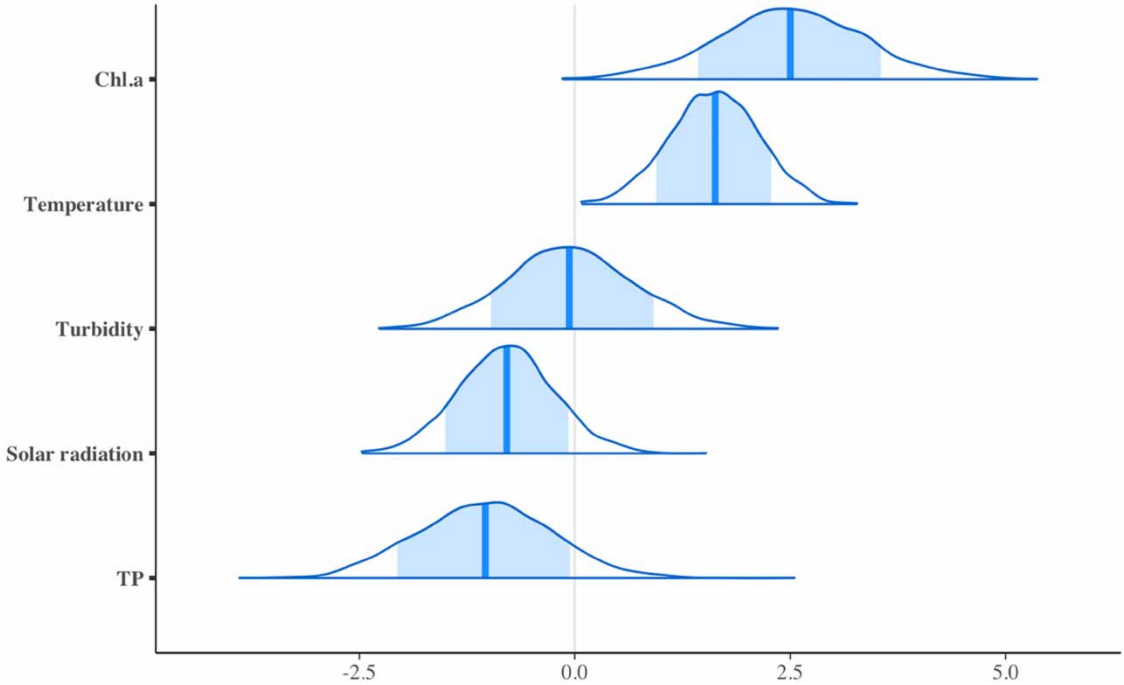

Figure 6 | Kernel density estimate of posterior distributions for parameters based on MCMC sampling with median and 80% intervals.

The turbidity coefficient was distributed on both sides of zero, indicating the possibility of either positive or negative correlations with cyanobacteria abundance. Turbidity is a general measure and does not distinguish types of matter, including no distinction between cyanobacteria and non-algal matter that would contribute to turbidity. Previously, cyanobacteria abundance of Kansas reservoirs was reported to be negatively correlated to non-algal turbidity (Dzialowski *et al.* 2011). As the non-algal turbidity increases, light penetration is reduced, and less cyanobacteria biomass is expected. Alternatively, cyanobacteria presence would lead to a measured increase in turbidity (Klemer & Konopka 1989). As such, the role of turbidity cannot be easily identified, and the parameter distribution appears to represent the uncertain relationship between turbidity and cyanobacteria counts accurately.

The coefficient distribution for total phosphorus was primarily distributed below zero, implying a negative correlation with cyanobacteria abundance. This result contradicts the expectation of phosphorous levels being positively associated with cell counts, given the substantial evidence that nutrient reduction strategies reduce blooms (Hamilton *et al.* 2016). It should be considered that there were relatively elevated levels of phosphorous in the reservoir (mean value of 0.1 mg/L), and nutrients may generally not have been a limiting factor for growth in this system. The recommended limit of total phosphorus in lakes is 0.05 mg/L (Litke 1999), and 92% of the recorded phosphorous levels in this dataset would imply the reservoir being studied is eutrophic or hypertrophic (Carlson & Simpson 1996). Relatively flat biomass responses with increasing phosphorous above a limiting threshold have also been previously reported (Dolman *et al.* 2012).

4. CONCLUSIONS

Bayesian mixture models were applied to model cyanobacteria abundance in a reservoir, with particular consideration for the tendency for cyanobacteria abundance to be highly imbalanced with a high proportion of zero values. Two models that can account for the high proportionality of zero measurements, including a ZINB and hurdle NB, were compared. An NB model was also applied to act as a baseline approach that does not account for excess zero counts.

Based on fit determined from LOO-CV, it was found that the ZINB and hurdle NB models performed significantly better than the NB model. The observed improvement of fit when using models that account for excessive zero counts supports the hypothesis that inflated zero counts are important to consider when modelling cyanobacteria abundance. Furthermore, a slight increase of fit was observed when using ZINB compared to the hurdle NB approach. ZINB models can account for

zero measurements being present either from the cell number being below detection limits, or from the true absence of cyanobacteria. As such, the improvement of fit using ZINB illustrates that both mechanisms of zero generation should be considered when modelling cyanobacteria.

The ZINB model was then applied to predict cyanobacteria levels using a separated test set. Although the performance was poor when predicting narrow alert level bands, precision and recall were high (0.62 and 0.99, respectively) for binary prediction of elevated vs. low risk levels of cyanobacteria. The established model utilizes a limited number of easy-to-measure parameters including Chl *a*, total phosphorous, pH, temperature, and solar radiation to generate these predictions. Furthermore, the predictions produced from the Bayesian approach utilized in this paper are probabilistic. The uncertainty from the data and interactions in the system are carried through the modelling process to produce an estimated cell count with an associated level of uncertainty. The high uncertainty levels in parameter estimates demonstrate that cyanobacteria count prediction is difficult, and the impact of influencing factors is complex. As such, we believe that the presented modelling process is well suited to inform the management of complex systems with high uncertainty, such as early warning of cyanobacteria blooms in surface waters.

ACKNOWLEDGEMENT

The research was funded through Canada's NSERC Discovery program (RGPIN-2019-05449).

DATA AVAILABILITY STATEMENT

All relevant data are available from https://github.com/meowliono/ZINB_cyanobacteria.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Anderson, D. M., Glibert, P. M. & Burkholder, J. M. 2002 Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries* **25**, 704–726.

Berkhof, J., Van Mechelen, I. & Hoijtink, H. 2000 Posterior predictive checks: principles and discussion. *Computational Statistics* **15** (3), 337–354.

Bishop, C. M. 2006 *Pattern Recognition and Machine Learning*. Springer, New York.

Bownik, A. 2016 Harmful algae: effects of cyanobacterial cyclic peptides on aquatic invertebrates – a short review. *Toxicon* **124**, 26–35.

Brooks, S. P. & Gelman, A. 1998 General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7** (4), 434–455.

Carlson, R. E. & Simpson, J. 1996 A coordinator's guide to volunteer lake monitoring methods. *North American Lake Management Society* **96**, 305.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. 2017 Stan: a probabilistic programming language. *Journal of Statistical Software* **76** (1), 1–32.

Catalina, A., Bürkner, P. C. & Vehtari, A. 2020 Projection Predictive Inference for Generalized Linear and Additive Multilevel Models. *arXiv preprint arXiv:2010.06994*.

Catherine, Q., Susanna, W., Isidora, E. S., Mark, H., Aurelie, V. & Jean-François, H. 2013 A review of current knowledge on toxic benthic freshwater cyanobacteria–ecology, toxin production and risk management. *Water Research* **47** (15), 5464–5479.

Cha, Y., Park, S. S., Kim, K., Byeon, M. & Stow, C. A. 2014 Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. *Water Resources Research* **50** (3), 2518–2532.

Chaffin, J. D., Kane, D. D., Stanislawczyk, K. & Parker, E. M. 2018 Accuracy of data buoys for measurement of cyanobacteria, chlorophyll, and turbidity in a large lake (Lake Erie, North America): implications for estimation of cyanobacterial bloom parameters from water quality sonde measurements. *Environmental Science and Pollution Research* **25** (25), 25175–25189.

Chapra, S. C., Boehlert, B., Fant, C., Bierman, V. J. Jr., Henderson, J., Mills, D., Mas, D. M. L., Rennels, L., Jantarasami, L., Martinich, J., Strzepek, K. M. & Paerl, H. W. 2017 Climate change impacts on harmful algal blooms in U.S. freshwaters: a screening-level assessment. *Environmental Science & Technology* **51**, 8933–8943.

Chorus, I. & Welker, M. 2021 *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*. Taylor & Francis, p. 858.

Christensen, V. G., Graham, J. L., Milligan, C. R., Pope, L. M. & Ziegler, A. C. 2006 *Water Quality and Relation to Taste-and-Odor Compounds in North Fork Ninnescah River and Cheney Reservoir, South-Central Kansas, 1997-2003 (No. 2006-5095)*. US Geological Survey.

Cirés, S., Wörmer, L., Timón, J., Wiedner, C. & Quesada, A. 2011 Cylindrospermopsin production and release by the potentially invasive cyanobacterium *Aphanizomenon ovalisporum* under temperature and light gradients. *Harmful Algae* **10** (6), 668–675.

Davidson, K., Gowen, R. J., Tett, P., Bresnan, E., Harrison, P. J., McKinney, A., Milligan, S., Mills, D. K., Silke, J. & Crooks, A.-M. 2012 Harmful algal blooms: how strong is the evidence that nutrient ratios and forms influence their occurrence? *Estuarine, Coastal and Shelf Science* **115**, 399–413.

Depaoli, S., Winter, S. D. & Visser, M. 2020 The importance of prior sensitivity analysis in Bayesian statistics: demonstrations using an interactive Shiny App. *Frontiers in Psychology* **11**, 608045.

Dolman, A. M., Rücker, 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.

Dzialowski, A. R., Smith, V. H., Huggins, D. G., Denoyelles, F., Lim, N. C., Baker, D. S. & Beury, J. H. 2009 Development of predictive models for geosmin-related taste and odor in Kansas, USA, drinking water reservoirs. *Water Research* **43** (11), 2829–2840.

Dzialowski, A. R., Smith, V. H., Wang, S. H., Martin, M. C. & deNoyelles Jr., F. 2011 Effects of non-algal turbidity on cyanobacterial biomass in seven turbid Kansas reservoirs. *Lake and Reservoir Management* **27** (1), 6–14.

Gelman, A. 2005 Comment: fuzzy and Bayesian p-values and u-values. *Statistical Science* **20** (2005), 380–381.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. 2013 *Bayesian Data Analysis*. CRC Press, New York.

Hamilton, D. P., Salmaso, N. & Paerl, H. W. 2016 Mitigating harmful cyanobacterial blooms: strategies for control of nitrogen and phosphorus loads. *Aquatic Ecology* **50** (3), 351–366.

Harris, T. D. & Graham, J. L. 2017 Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. *Lake and Reservoir Management* **33** (1), 32–48.

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications *Biometrika*. Volume 57, Issue 1, April 1970, Pages 97–109. https://doi.org/10.1093/biomet/57.1.97.

He, H. & Garcia, E. A. 2009 Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1263–1284.

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.

Ibelings, B. W., Backer, L. C., Kardinaal, W. E. A. & Chorus, I. 2014 Current approaches to cyanotoxin risk assessment and risk management around the globe. *Harmful Algae* **40**, 63–74.

Japkowicz, N. & Stephen, S. 2002 The class imbalance problem: A systematic study. *Intelligent data analysis* **6** (5), 429–449.

Jöhnk, K. D., Huisman, J. E. F., Sharples, J., Sommeijer, B. E. N., Visser, P. M. & Stroom, J. M. 2008 Summer heatwaves promote blooms of harmful cyanobacteria. *Global Change Biology* **14** (3), 495–512.

Klemer, A. R. & Konopka, A. E. 1989 Causes and consequences of blue-green algal (cyanobacterial) blooms. *Lake and Reservoir Management* **5** (1), 9–19.

Korner-Nievergelt, F., Roth, T., von Felten, S., Guélat, J., Almasi, B. & Korner, P. 2015 Posterior predictive model checking and proportion of explained variance. In: *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN*. Elsevier, pp. 161–174. doi:10.1016/B978-0-12-801370-0.00010-1.

Kuk, A. Y., Li, J. & John Rush, A. 2014 Variable and threshold selection to control predictive accuracy in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63** (4), 657–672.

Lambert, D. 1992 Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** (1), 1–14.

Lambert, B. 2018 *A Student's Guide to Bayesian Statistics*. Sage.

LeBlanc Renaud, S., Pick, F. R. & Fortin, N. 2011 Effect of light intensity on the relative dominance of toxigenic and nontoxigenic strains of *Microcystis aeruginosa*. *Applied and Environmental Microbiology* **77** (19), 7016–7022.

Lee, T. A., Rollwagen-Bollens, G., Bollens, S. M. & Faber-Hammond, J. J. 2015 Environmental influence on cyanobacteria abundance and microcystin toxin production in a shallow temperate lake. *Ecotoxicology and Environmental Safety* **114**, 318–325.

Lee, J., Lee, S. & Jiang, X. 2017 Cyanobacterial toxins in freshwater and food: important sources of exposure to humans. *Annual Review of Food Science and Technology* **8**, 281–304.

Litke, D. W. 1999 *Review of Phosphorus Control Measures in the United States and Their Effects on Water Quality*, Vol. 99, No. 4007. US Department of the Interior, US Geological Survey.

Malve, O., Laine, M., Haario, H., Kirkkala, T. & Sarvala, J. 2007 Bayesian modelling of algal mass occurrences – using adaptive MCMC methods with a lake water quality model. *Environmental Modelling & Software* **22** (7), 966–977.

Mangan, N. M., Flamholz, A., Hood, R. D., Milo, R. & Savage, D. F. 2016 Ph determines the energetic efficiency of the cyanobacterial CO_2 concentrating mechanism. *Proceedings of the National Academy of Sciences* **113** (36), E5354–E5362.

Meng, X. L. 1994 Posterior predictive p-values. *The Annals of Statistics* **22** (3), 1142–1160.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21** (6), 1087–1092.

Min, Y. & Agresti, A. 2005 Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5** (1), 1–19.

Newcombe, G., House, J., Ho, L., Baker, P. & Burch, M. 2010 *Management Strategies for Cyanobacteria (Blue-Green Algae): A Guide for Water Utilities*. Water Quality Research Australia (WQRA). Reserach Report 74, pp. 60–76.

Oberemm, A., Becker, J., Codd, G. A. & Steinberg, C. 1999 Effects of cyanobacterial toxins and aqueous crude extracts of cyanobacteria on the development of fish and amphibians. *Environmental Toxicology: An International Journal* **14** (1), 77–88.

O'neil, J. M., Davis, T. W., Burford, M. A. & Gobler, C. J. 2012 The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. *Harmful Algae* **14**, 313–334.

Paerl, H. W. & Otten, T. G. 2013 Harmful cyanobacterial blooms: causes, consequences, and controls. *Microbial Ecology* **65**, 995–1010.

Piironen, J., Paasiniemi, M. & Vehtari, A. 2020 Projective inference in high-dimensional problems: prediction and feature selection. *Electronic Journal of Statistics* **14** (1), 2155–2197.

Pyo, J., Park, L. J., Pachepsky, Y., Baek, S. S., Kim, K. & Cho, K. H. 2020 Using convolutional neural network for predicting cyanobacteria concentrations in river water. *Water Research* **186**, 116349.

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.

Sinha, R. P., Kumar, A., Tyagi, M. B. & Hader, D. 2005 Ultraviolet-B-induced destruction of phycobiliproteins in cyanobacteria. *Physiology and Molecular Biology of Plants* **11** (2), 313.

Taranu, Z. E., Gregory-Eaves, I., Leavitt, P. R., Bunting, L., Buchaca, T., Catalan, J., Domaizon, I., Guilizzoni, P., Lami, A., McGowan, S., Moorhouse, H., Morabito, G., Pick, F. R., Stevenson, M. A., Thompson, P. L. & Vinebrooke, R. D. 2015 Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. *Ecology Letters* **18** (4), 375–384.

Thomas, M. K. & Litchman, E. 2016 Effects of temperature and nitrogen availability on the growth of invasive and native cyanobacteria. *Hydrobiologia* **763** (1), 357–369.

US Geological Survey 2015 *USGS National Water Information System*. http://dx.doi.org/10.5066/F7P55KJN.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. & Yau, C. 2021 Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1** (1), 1–26.

Verspagen, J. M., Van de Waal, D. B., Finke, J. F., Visser, P. M., Van Donk, E. & Huisman, J. 2014 Rising CO_2 levels will intensify phytoplankton blooms in eutrophic and hypertrophic lakes. *PLoS ONE* **9** (8), e104325.

Walls, J. T., Wyatt, K. H., Doll, J. C., Rubenstein, E. M. & Rober, A. R. 2018 Hot and toxic: temperature regulates microcystin release from cyanobacteria. *Science of the Total Environment* **610**, 786–795.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. 1996 Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* **88** (1–3), 297–308.

Wenger, S. J. & Freeman, M. C. 2008 Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology* **89** (10), 2953–2959.

Xue, L., Zhang, Y., Zhang, T., An, L. & Wang, X. 2005 Effects of enhanced ultraviolet-B radiation on algae and cyanobacteria. *Critical Reviews in Microbiology* **31** (2), 79–89.

Zhang, F., Lee, J., Liang, S. & Shum, C. K. 2015 Cyanobacteria blooms and non-alcoholic liver disease: evidence from a county level ecological study in the United States. *Environmental Health* **14** (1), 1–11.

Zhao, C. S., Shao, N. F., Yang, S. T., Ren, H., Ge, Y. R., Feng, P., Dong, B. E. & Zhao, Y. 2019 Predicting cyanobacteria bloom occurrence in lakes and reservoirs before blooms occur. *Science of the Total Environment* **670**, 837–848.

First received 30 December 2022; accepted in revised form 31 July 2023. Available online 14 August 2023