qPCR-Based Prediction of Low-Level Microcystin-LR Using *mcyE* and Passive Sampling Across Multiple Lakes Over Three Years

David J Redden1\*, Clarke Brown1, Morgan Harasymchuk1, Saksham Bafna1, Justin Laforest2, Nicole Taylor2, Lindsay Anderson1,3, and Graham A. Gagnon1

1 Centre for Water Resources Studies, Department of Civil & Resource Engineering, Dalhousie University, 1360 Barrington St. Halifax, Nova Scotia, Canada

2 City of Moncton, 655 Main St., Moncton, New Brunswick, Canada

3 Halifax Water, 450 Cowie Hill Road, Halifax, Nova Scotia, Canada

\* Corresponding author

Manuscript submitted for possible publication in *Water Research X*

# 1. Supplementary Methods

## 1.1 Sample collection

All samples were collected in prepared Nalgene bottles. The bottles for DNA extraction were bathed in 10% bleach for at least 12 h and triple rinsed with sample water before collection. The bottles for all other analyses were washed and triple rinsed with sample water before collection. Following collection, all samples were placed on ice and kept in the dark in a cooler until delivery to the laboratory. During each sampling event, a DNA field blank was placed in the sample cooler and taken to each site. The field blank was prepared by triple rinsing a bleach-bathed Nalgene bottle with deionized (DI) water, filling it with DI water, and then autoclaving the filled bottle.

## 1.2 Physicochemical analyses

Turbidity was measured on a laboratory turbidimeter (HACH Company, Loveland, CO). Unfiltered samples for TOC and TN were placed, headspace-free, in 40 mL glass vials and preserved (pH < 2) with concentrated phosphoric acid for analysis on a TOC-V CPH analyzer (Shimazdu Corporation, Kyoto Japan). DOC was analyzed in the same manner using filtered samples. Elemental composition of unfiltered samples was quantified by inductively coupled mass-spectrometry (ICP-MS) after preserving (pH < 2) with nitric acid. Dissolved anions were quantified using ion chromatography (IC; Dionex Aquion, Thermo Scientific, Waltham, MA) and optical properties (true color and UV254) were characterized on a benchtop UV-Vis spectrophotometer (HACH Company, Loveland, CO).

## 1.3 DNA extraction quality control

In 2022, samples were concentrated on 25 mm membranes using sterile syringe filter holders. Filters were placed into BioGX Bead Lysis Tubes (BioGX Inc, Birmingham, AL) containing 500 µL of lysis buffer (Phytoxigene™, Inc., Akron, OH) and DNA was extracted immediately, following the manufacturer’s protocol. In 2023 and 2024, samples were concentrated onto 47 mm membrane filters using sterilized, reusable polysulfone filter holders (Advantec MFS, Dublin, CA). Filters were stored at -70 °C in bead beating tubes provided in the DNeasy PowerWater Kit (Qiagen, Venlo, The Netherlands), until DNA extraction following the manufacturer’s protocol.

Throughout the monitoring program, the same contamination control and quality assurance protocols were applied, regardless of extraction method. All extractions were conducted using aseptic technique inside a Thermo Scientific 1300 Series Class II Type A2 Biological Safety Cabinet (Thermo Fisher Scientific, Waltham, MA). Field blanks and extraction blanks (empty bead tubes processed alongside samples) were included in every extraction batch. Both types of filter holders were sterilized by washing, soaking in 10% bleach for at least 12 hours, triple rinsing with DI water, and autoclaving.

## 1.4 qPCR methodology

### 1.4.1 Assay preparation and thermocycling conditions

Both CyanoDTec assays were prepared according to the manufacturer’s protocols in 96 well plates (Bio-Rad Laboratories Inc., Cepheid, CA) and amplified using a Bio-Rad CFX Opus 96-well thermocycler (Bio-Rad Laboratories Inc.). Thermocycling conditions were: initial denaturation at 95 °C for 2 min, followed by denaturation at 95 °C for 15 seconds, and then 40 cycles of annealing and extension at 60 °C for 45 seconds.

### 1.4.2 Quantification of gene concentrations

Five-point standard curves, used to convert the cycle of quantification () to a concentration of gene copies (GC), were run in duplicate by each analyst for each lot of reagents. The standards, Phytoxigene™ CyanoNAS Nucleic Acid Standards (Phytoxigene™, Akron, OH), comprised five concentrations (20, 200, 2,000, 20,000, and 200,000 GC µL-1) of each gene target. All standard curves had a R2 values > 0.99 with efficiencies between 90 and 110%. After converting the to a qPCR concentration, the volume of eluent used in the extraction, and the total volume of water concentrated onto the filter membrane were used to compute the equivalent concentration per sample (Eq. 1). According to the manufacturer, the limit of detection (LOD) — the concentration at which there is a 95% chance of detection — for each Phytoxigene™ CyanoDTec assay is 9 GC μL−1. The LOD for each sample was scaled according to Eq. 1 based on the volume of sample that was concentrated on the filter membrane and thus varies by sample.

### 1.4.3 qPCR quality control

The Environmental Microbiology Minimum Information (EMMI) guidelines for qPCR experiments (Borchardt et al., 2021) were consulted and followed, where applicable. Each qPCR run included a positive control and a nuclease-free water no-template control (NTC). The positive control was one of the five CyanoNAS standards used to create the standard curves. Auto-thresholding was used to determine the fluorescent baseline of the thermocycler. The values of the positive controls in each run were compared to their corresponding value in the standard curve to ensure that positive control values were always within one cycle of their  value in the standard curve. Negative controls — NTCs, field blanks, and extraction blanks — were used to assess contamination during qPCR, sampling, and extraction. A detection in these controls was defined as any non-NA value. None of the target genes were ever detected in any of the negative control samples. The CyanoDTec Total Cyanobacteria assay contains an internal amplification control (IAC) that is used to verify that PCR reactions are not inhibited. According to the manufacturer, a sample for which the value of the IAC was more than 1.5 cycles greater than the IAC of the NTC should be considered inhibited and rerun after dilution. No samples were found to be inhibited.

All qPCR master mix was preparaed in a MicroZone Cleanceil laminar flow hood (Design Filtration Microzone Inc., Ottawa, ON) immediately before use. Master mix and qPCR template were combined in 96 well plates in a Mystaire® MY-PCR prep station (Mystaire Inc, Creedmore, NC) laminar flow hood immediately before PCR amplification. Sterile, filtered pipette tips were used throughout.

## 1.5 MC-LR analytical methods

### 1.5.1 Grab sample processing

Grab samples were processed via three cycles of freezing at -20°C for at least one hour or until frozen, followed by thawing. Samples then had an internal standard mix added before being centrifuged for 10 minutes at 10 000 rpm (~10 845 relative centrifugal force; RCF) to remove debris. An aliquot of this sample was then transferred to a LC vial for analysis.

### 1.5.2 Passive sample processing

The passive samples comprised 1000 mg of Oasis HLB adsorbent (Waters Limited, Stamford Ave., UK) in a heat-sealed nylon mesh bag with 25 µm pore size. The fill mesh bags were placed in a plastic sampling cage that was modified from the work of Hayes et al. (2021) and affixed at the sampling location — either affixed via fishing line to the shoreline or to buoys used to indicate the sampling location. After sampler collection, the entire sampling bag was removed from the plastic cage, placed in a falcon tube, rinsed with 10 mL of DI water by shaking vigorously for 1 minute, transferred to a new falcon tube containing 10 mL of LC-MS grade methanol and shaken again for 1 minute to extract MC-LR. The falcon tube was then centrifuged for 10 minutes at 8000 rpm (~9874 RCF) to settle the sorbent and any remaining debris. A 500 µL aliquot of the supernatant was transferred to an LC vial and diluted by a factor of 2 with DI water and the internal standard mix.

### 1.5.3 LS-MS/MS instrumentation

The LC-MS/MS comprised a binary pump LC system (Agilent 1260) coupled with a triple quadrupole mass spectrometer (Agilent 6460) configured in positive electrospray ionization (ESI) mode. The LC separation was accomplished using an Agilent InifinityLab Poroshell EC C-18 (4.6 x 150 mm x 2.7 µm) column, preceded by an Agilent InfinityLab Poroshell EC C-18 (4.6 x 5 mm x 2.7 µm) guard column.

The instrument was operated using multiple reaction monitoring (MRM) mode with two transitions providing a quantifier and qualifier ion for identification and quantitation. The LC column was held at a constant temperature of 40°C. The mobile phase consisted of 0.4% (v/v) acetic acid in DI water (Solvent A) and 70%:30% (v/v) acetonitrile: methanol (Solvent B). The separation solvent gradient proceeded as follows:

| Time (minutes) | % Solvent A | % Solvent B |
| --- | --- | --- |
| 0 | 70 | 30 |
| 0.1 | 70 | 30 |
| 5 | 2 | 98 |
| 7 | 2 | 98 |
| Post time: 3 minutes |  |  |

The flow rate was constant at 750 µL min-1 and 40 µL of sample was injected per run. Samples were held at 4°C in the autosampler tray. The ESI source parameters were set to the following values: gas temperature, 350°C; gas flow rate, 12 L min-1; nebulizer pressure, 50 psi; sheath gas temperature, 400°C; sheath gas flow rate, 12L min-1; capillary voltage, 3500 V; nozzle voltage,0 V; delta EMV cell acceleration voltage, 200V. The MS parameters were as follows:

Insert table here

The MC-LR internal standard was purchased from Caymen Chemicals (Ann Arbor, MI, USA). LC-MS grade methanol, acetonitrile, and acetic acid were purchased from Fisher Scientific (Ottawa, ON, CA). The LC-MS/MS data analysis was completed using the Agilent MassHunter Software (Version 12.0, build 12.0.893.1).

## 1.6 Bayesian mixed model descriptions

We fit five hierarchical logistic regression models to estimate the probability of detecting MC-LR, incorporating smooth seasonal effects, fixed effects for key predictors, and random effects to account for location-level heterogeneity. Missing and censored values in the predictor variables were jointly modeled using the *mi()* function in *brms* (Bürkner, 2017), allowing uncertainty from imputation to propagate through to posterior estimates of the regression coefficients. Each imputation submodel followed a Student’s *t*-distribution, with a conditional mean defined using observed covariates and a location-specific random intercept (see Section 1.6 for full submodel specifications).

The five models were:

1. **Null model**

* This model included a fixed intercept, with location- and year-specific random intercepts:
* where is the population-level intercept, and the deviations and account for random variation across location and year, respectively.

1. ***mcyE*** **with varying intercepts** This model included a population-level intercept and fixed effects for *mcyE* and year. A smooth function of week-of-year () was estimated separately for each location () and year () combination:

* Missing values in *mcyE* were imputed using a Student’s *t*-distribution with a mean modeled as a function of year and a location-specific random intercept.

1. **mcyE with varying intercepts and slopes**  
   This model extended the previous one by allowing the effect of *mcyE* to vary by location. Both the intercept and the slope for *mcyE* were modeled as location-specific effects drawn from a multivariate normal distribution:

* where is a covariance matrix capturing the correlation between location-specific intercepts and slopes for *mcyE*.

1. **Shared predictor model**  
   This model included the top eight predictors shared between the top passive and grab sample random forest models: *mcyE*, lag *mcyE*, TDS, conductivity, chloride, , true colour, and the 7-d total precipitation. A smooth function of week-of-year was fit for each location, and both intercepts and slopes were allowed to vary by location. Each predictor had a fixed effect and a location-level deviation modeled jointly via a multivariate normal distribution:

* where is a covariance matrix allowing for correlation among location-specific intercepts and slopes.

1. **Passive predictor model**  
   This model followed the same structure as the shared predictor model but used the top five predictors from the passive sample random forest model: *mcyE*, lag *mcyE*, TDS, conductivity, and chloride.

* The random effects for the intercept and five predictors were again modeled using a multivariate normal distribution:

Each model was fit using brms with a Bernoulli likelihood, logit link, and the factor-smooth basis for the location-specific spline on week-of-year.

## 1.7 Imputation model specifications

Each partially observed predictor was modeled using a Student’s *t*-distribution, with a predictor-specific mean structure and a random intercept by location. Imputed values were incorporated into the main regression model via the *mi()* syntax in *brms*. Below, we describe each submodel.

For all imputed predictors :

### 1.7.1 Predictors modeled as a function of year and location:

For :

### 1.7.2 Predictors modeled as a function of conductivity, year, and location:

For :

### 1.7.3 Predictors modeled as a function of true colour, year, and location:

For :

### 1.7.4 Random intercept for location:

For all :

All submodels were estimated jointly with the main regression model using the brms multivariate modeling framework, ensuring that posterior uncertainty in the imputed values was fully incorporated into predictions of MC-LR probability.

## 1.8 Prior specifications

For all models, weakly informative priors were used. The only priors explicitly specified were:

* A Student’s prior on all population-level regression coefficients in the MC-LR detection model:
* A normal prior on the intercept :

All other priors were left at their brms defaults:

* Flat (improper uniform) priors on unconstrained population-level effects for other responses (e.g., imputation models),
* Student- priors on standard deviations for group-level effects,
* LKJ(1) priors on correlation matrices among random effects,
* Gamma(2, 0.1) priors on the degrees of freedom in Student’s distributions used for imputation,
* Student- priors on residual standard deviations in the imputation submodels.

# 2. Figures

|  |
| --- |
| Figure 1: Comparison of qPCR results obtained from the two DNA extraction methods used in this work. |

|  |
| --- |
| Figure 2: Distribution of true colour before and after missing data imputation. |

|  |
| --- |
| Figure 3: Distribution of total organic carbon before and after missing data imputation. |

|  |
| --- |
| Figure 4: Distribution of before and after missing data imputation. |

|  |
| --- |
| Figure 5: Distribution of *mcyE* gene concentrations before and after missing data imputation. |

|  |
| --- |
| Figure 6: Distribution of total cyanobacteria 16S rRNA gene concentrations before and after missing data imputation. |

|  |
| --- |
| Figure 7: Correlations between predictor variables. |

|  |
| --- |
| Figure 8: Pairwise comparisons of physicochemical parameters between lakes. The size of the points indicates the magnitude of p-values obtained using Dunn’s test with Benjamini-Hochberg correction. |

|  |
| --- |
| Figure 9: Concentrations of MC-LR in grab samples collected over the duration of the montoring program. The shape of the points denotes the sampling location and the colour denotes the year. |

|  |
| --- |
| Figure 10: Concentrations of *mcyE* and total cyanobacterial 16S rRNA genes at each location. Samples in which MC-LR was detected in passive samples are indicated by red points; non-detects are shown in gray. Locations are ordered top to bottom by the proportion of samples with MC-LR detections, indicated in parentheses (detections / total samples). Spearman correlations coefficients, , were calculated between MC-LR detection rates and the median concentration of gene target across all samples at a location. These values are shown in the top-right of each panel. qPCR concentrations that were below the limit of detection were replaced with a value of for plotting. |

|  |
| --- |
| Figure 11: Receiver operator characteristic curves and variable importance metrics for random forest models trained on data from 2022 and 2023 and tested on data from 2024 for grab (**a**, **b**) and passive (**c**, **d**) MC-LR detections. . |

|  |
| --- |
| Figure 12: Additional posterior predictive checks comparing (**a**) the predicted and observed counts of MC-LR detections and (**b**) the observed and predicted standard deviations of the binomial distributed MC-LR detections. |

|  |
| --- |
| Figure 13: *mcyE* gene concentrations associated with a 75% predicted probability of detectable MC-LR from the hierarchical Bayesian logistic regression model. The lines and ribbons represent the median and interquartile range of the predicted probability across 4000 posterior draws. |

|  |
| --- |
| Figure 14: *mcyE* gene concentrations associated with a 95% predicted probability of detectable MC-LR from the hierarchical Bayesian logistic regression model. The lines and ribbons represent the median and interquartile range of the predicted probability across 4000 posterior draws. |

# 3. Tables

Areas and maximum depths of the lakes included in this study.

| Lake | Area (ha) | Maximum Depth (m) |
| --- | --- | --- |
| Lake Banook | 33.30 | 12 |
| Penhorn Lake | 4.10 | 9 |
| Shubenacadie Canal | 0.55 | unknown (shallow) |
| Lake Charles | 146.00 | 30 |
| Lake Fletcher | 105.20 | 11 |
| Kearney Lake | 63.00 | 25 |
| Cunard Pond | 2.50 | unknown (shallow) |
| Irishtown Nature Park | 98.00 | 9 |
| Tower Road Reservoir | 160.00 | 16 |
| Turtle Creek Reservoir | 73.00 | 18 |

Model comparison using Pareto smoothed important sampling as an approximation of leave-one-out cross validation.

| Model | ELPD | Δ ELPD | LOOIC | Δ LOOIC | No. *k* > 0.7 | No. *k* > 1 |
| --- | --- | --- | --- | --- | --- | --- |
| *mcyE* with varying intercepts | -251.94 | 0.00 | 503.88 | 0.00 | 5 | 0 |
| *mcyE* with varying slopes and intercepts | -252.87 | -0.93 | 505.74 | 1.85 | 6 | 0 |
| Passive predictor model | -263.76 | -11.81 | 527.51 | 23.63 | 27 | 0 |
| Shared predictor model | -264.92 | -12.98 | 529.84 | 25.96 | 45 | 2 |
| Null model | -299.48 | -47.54 | 598.97 | 95.08 | 0 | 0 |

# 4. References

Borchardt, M.A., Boehm, A.B., Salit, M., Spencer, S.K., Wigginton, K.R., Noble, R.T., 2021. The environmental microbiology minimum information (EMMI) guidelines: qPCR and dPCR quality and reporting for environmental microbiology. Environmental Science & Technology 55, 10210–10223. <https://doi.org/10.1021/acs.est.1c01767>

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28. <https://doi.org/10.18637/jss.v080.i01>

Hayes, E.K., Sweeney, C.L., Anderson, L.E., Li, B., Erjavec, G.B., Gouthro, M.T., Krkosek, W.H., Stoddart, A.K., Gagnon, G.A., 2021. A novel passive sampling approach for SARS-CoV-2 in wastewater in a Canadian province with low prevalence of COVID-19. Environmental Science: Water Research & Technology 7, 1576–1586. <https://doi.org/10.1039/D1EW00207D>