Description of Updates Made to the Lake Criterion Models

September 8, 2022

The 2017 National Lakes Assessment data were added to the data used to estimate national nutrient criterion models. The larger data set provided an opportunity to explore the effects of variables that might modify stressor-response relationships. Based on these analyses, minor changes were implemented in each of the criterion models.

Microcystin model

Incorporating 2017 data increased the data set from 2352 samples collected from 1116 lakes to 3535 samples from 1633 lakes. Using this larger data set we examined the effects of other variables on the two statistical relationships that define the microcystin model: (1) the relationship between Chl *a* and cyanobacterial relative abundance, and (2) the relationship between cyanobacterial biovolume and microcystin.

Exploratory analysis

Generalized additive models (GAMs) were used to assess the effects of different covariates on the two relationships of interest. GAMs were ideally suited for data exploration because they could be estimated quickly for many different model configurations. First, base models were estimated for the two relationships. For cyanobacterial relative abundance, the following base model was estimated:

$$P_{cyano} = s(\log(Chl)) \quad (1)$$

Where P_{cyano} is cyanobacterial relative abundance, *Chl* is Chl *a* concentration, and *s(Chl)* is a non-parametric smooth function estimated by the model. Cyanobacterial relative abundance is bounded by a maximum value of 1 and a minimum value of 0, so a quasibinomial sampling distribution was used in the model.

For microcystin the base model was expressed as:

$$MC = s(\log(B_{cyano})) \quad (2)$$

Where *MC* is the observed microcystin concentration, and B_{cyano} is the cyanobacterial biovolume. The sampling distribution of microcystin concentrations was assumed to be a negative binomial distribution to account for its lower bound of zero and infrequent occurrences of large values.

To assess the effect of each candidate covariate, the base GAM was refit with an additional term that modeled the effect of the covariate. For example, the effect of lake depth on the Chl *a* – cyanobacterial relative abundance model was assessed by fitting the following model:

$$P_{cyano} = s(\log(Chl)) + s(Depth) \quad (3)$$

Where *s(Depth)* is a non-parametric smooth function of lake maximum depth. The influence of each candidate covariate was quantified by computing the deviance explained by the modified model and comparing it to the deviance explained by the base model. Eleven candidate covariates were considered

based on data availability (Table 1). Ecoregion is a discrete variable and was included in the models as an additive factor. All other candidate covariates were modeled with non-parametric smooth functions as described above.

Of the modifying variables considered, ecoregion accounted for the greatest percentage increase in the deviance explained by the model (Table 1). Of the continuous variables, lake depth accounted for the greatest increase in deviance explained in the Chl *a* – cyanobacterial relative abundance model, while DOC accounted for the greatest increase in the deviance explained in the cyanobacterial biovolume – MC model.

Table 1. Percentage increase in deviance explained by including the indicated modifying variable in the GAM.

Modifying variable	Cyanobacterial biovolume – MC	Chl *a* – Cyanobacterial relative abundance
Elevation	3	5
DOC	19	3
Sampling day	4	2
Lake temperature	-1	8
Depth	3	16
Area	1	15
Conductivity	8	15
Color	1	7
Longitude	5	8
Latitude	3	3
Ecoregion	43	37

Incorporating classification variables in the criterion model

Including ecoregion as a predictor yielded a large increase in the deviance explained but required at least an additional 85 degrees of freedom in each model because different values for each model coefficient are estimated for each ecoregion. More specifically, with a total of 3535 samples, each ecoregion-specific coefficient that was included in model reduced the ratio between the number of independent samples and number of parameters by a factor of approximately 20. Testing with independent validation data (see below) indicated that including ecoregion for both relationships overfit the available data, and so, ecoregion was included as a classifying variable only in the cyanobacterial biovolume – MC model. For the Chl *a* – cyanobacterial relative abundance model, maximum lake depth was classified into 4 groups and incorporated in the model.

In the MC criterion model, depth-specific values of the coefficients that quantified the Chl *a* – cyanobacterial relative abundance model and ecoregion-specific values for the coefficients that quantified the cyanobacterial biovolume – MC model were estimated as follows,

$$logit\left(P_{cyano}\right) = f_{1,j} + f_{2,j}\log(Chl) + f_{3,j}\log\left(Chl\right)^2 \quad (4)$$

$$MC = b_{1,i} + b_{2,i}\log\left(B_{cyano}\right) \quad (5)$$

Where different values of the coefficients $f_{1,j}$, $f_{2,j}$, and $f_{3,j}$ are estimated for each of four depth classes. These depth classes were defined to ensure that approximately the same number of samples were included in each class. Class boundaries were as follows: D ≤ 2.6 m, 2.6 < D ≤ 4.75 m, 4.75 < D ≤ 9.1 m, and D > 9.1 m.

The coefficients, $b_{1,i}$, and $b_{2,i}$, were estimated for each ecoregion, *i*. For these coefficients, a hierarchical structure was imposed to shrink the value of each coefficient toward the overall mean and to allow ecoregions with smaller amounts of data to "borrow strength" from other ecoregions. To that end, ecoregion-specific values of each coefficient were assumed to be drawn from a common normal distribution. For example, the values of b_1 were assumed to be drawn from one normal distribution as follows,

$$b_1 \sim Normal(\mu_{b1},\ s_{b1}) \quad (6)$$

Where the normal distribution is characterized by a mean of μ_{b1} and a standard deviation of s_{b1}. Imposition of this hierarchical structure ensured that relationships estimated in ecoregions with smaller amounts of data were still consistent with the trends that characterized the full data set. An identical normal distribution was specified for the coefficient, b_2.

The formulation of the cyanobacterial biovolume – MC model was simplified to reduce the overall number of parameters that were estimated in this model. As shown in Equation (5), a simple linear model is used to represent this relationship, whereas in the original model, a piecewise linear model requiring two additional coefficients was used. Because the range of possible cyanobacterial biovolumes within an ecoregion is more limited than the range observed across the full data set, this linear approximation could still accurately represent the underlying relationship.

Performance of revised model

Model performance was assessed by testing the model using NLA data collected in 2007. These data were not used to fit the model and therefore provided independent validation data. A predicted mean microcystin concentration was calculated for each Chl *a* concentration observed in 2007 NLA data. Predictions were then binned into groups of ~40 samples with similar predicted values of mean MC. Within each group, the mean observed MC concentration was calculated and compared with the predicted mean (Figure 1).

Predictions of 2007 observations from both the original and revised models exhibited a similar mean bias in the predictions. Log-transformed MC concentrations were greater than predictions by 0.63 and 0.67 units for the original and revised models, respectively. After correcting for this mean bias, the root mean square (RMS) errors were 0.79 and 0.67 for the original and revised models, a difference that corresponded with a 15% improvement in predictive accuracy.

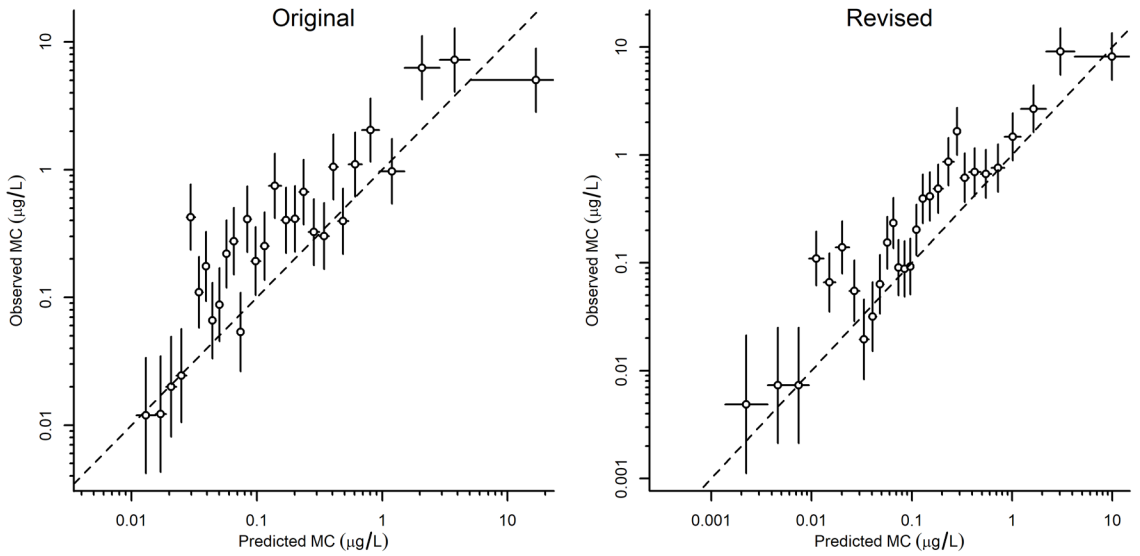

Figure 1. Predictive accuracy of microcystin models. Left panel: original model, right panel: revised model. Open circles: mean observed MC concentration at the indicated mean prediction for the group. Vertical line segments: 95% confidence limit on estimated mean value. Horizontal line segments: range of predicted mean MC values for bin. Dashed diagonal line: 1:1 relationship.

Hypoxia model

The 2017 NLA data increased the number of samples to analyze for the hypoxia model from 488 samples collected at 386 sites to 791 samples collected at 553 sites. Accurately estimating the first day of stratification is a critical aspect of the hypoxia model because hypolimnetic oxygen depletion commences on this day. Therefore, while incorporating 2017 NLA data into the hypoxia criterion model, we re-examined the model predicting the first day of stratification.

First day of stratification and spring turnover models

The first day of stratification and day of spring turnover are closely related dates for a dimictic lake. During spring turnover, the water column is isothermal, and a small wind stress can completely mix the lake. In most lakes, the first day of stratification follows soon after spring turnover as warming at the surface of the lake establishes a layer of water at the surface that is less dense than deeper waters. As warming continues, the density gradient becomes more pronounced and larger wind stresses are required to disturb the layers.

For the hypoxia criterion model, the day of spring turnover is most relevant to the predictions of deep-water oxygen concentration because spring turnover fixes the point from which the depletion of dissolved oxygen begins. The first day of stratification is closely related to spring turnover, but as described above, vertical transport through the lake water column is limited as soon as warming begins at the surface (i.e., immediately after spring turnover). Different methods have been used to quantitatively define a stably stratified lake (e.g., maximum temperature/density gradient, buoyancy frequency, density difference between the top and bottom of the lake), but for the hypoxia model, we are most interested in the day that the lake water column is isothermal (and dissolved oxygen concentrations are uniform throughout the water column).

Measurements of spring turnover are rare, but recently published studies indicated that spring turnover in lakes can be estimated as the day that surface temperatures in a lake reach 4°C (Woolway *et al.*, 2021), and surface temperatures in large lakes can be measured remotely. So, to better understand the factors that influence the day of spring turnover, remotely sensed lake surface temperature measurements were downloaded from the Copernicus Global Land Service (https://land.copernicus.eu/global/products/lswt). In this data set, satellite measurements of lake surface temperature were reported every ten days for 1018 lakes worldwide. Data were downloaded starting from 2007 (the first year of data of the NLA) up to 2020. Data were not available from 2012 – 2016 due to changes in the availability of different satellites. At the center of each lake, the day of the year that the lake temperature reached 4°C (WATER4) was estimated by linearly interpolating among the 10-day temperature measurements. Lakes where temperatures did not decrease below 4°C were excluded from the data set, as were the Great Lakes, endorheic lakes, and cold monomictic lakes. A total of 265 estimates of WATER4 at 31 different lakes in the conterminous U.S. were available for analysis (Figure 2).

Figure 2. Lakes with satellite temperature data.

Three predictors of the day of spring turnover were evaluated: air temperature, lake depth, and lake area. Mean annual air temperature has been used to predict the date of stratification onset (Demers & Kalff, 1993), but for this application a new air temperature metric was developed that better represents air temperature conditions immediately prior to lake turnover. This new metric, AIR4, is defined as the day of the year that mean daily air temperature reaches 4°C. To estimate AIR4 at all locations in the conterminous U.S., 30-year average monthly air temperatures across the U.S. at a 4 km scale were obtained from the PRISM web site (https://prism.oregonstate.edu/normals/). AIR4 was then estimated at all locations by linearly interpolating among monthly mean temperatures. Lake depth and lake area quantify morphological characteristics of a lake that can influence the rate of warming. More specifically, the depth of the mixed layer is related to the average magnitude of wind stress on the lake surface, which in turn, is related to the fetch and lake area. Similarly, lake depth is associated with the depth of the mixed layer, as deeper lakes have a greater potential for deeper mixed layers.

All three predictors were significantly associated with WATER4. Overall, AIR4, log-transformed surface area, and log-transformed depth accounted for 79% of the observed variance in WATER4. AIR4 was strongly and positively associated with WATER4 (Figure 3). Similarly, increased lake area and increased lake depth were both associated with later WATER4 values, as would be expected. These parameters were incorporated into the hypoxia model to improve predictions of the date of spring turnover when estimating the effects of deep-water hypoxia.

Performance of the revised model

The relevance of the day of spring turnover is evident when t_0, estimated from the hypoxia model, is compared with directly measured WATER4 from satellite data. Recall from the criterion document that t_0 is estimated as the day at which dissolved oxygen (DO) in a particular lake is equivalent to DO at a temperature of 4°C and that corresponds with a linear decrease in dissolved oxygen concentrations. That is, t_0 is estimated by projecting backward in time from a temporal history of DO measurements from lakes with similar characteristics. These estimated values of t_0 (open circles, Figure 3) were comparable to directly measured spring turnover days (filled circles, Figure 3). More specifically, the linear relationship between t_0 and AIR4 was very similar to the relationship between directly measured WATER4 and AIR4, and the overall mean value of t_0 was somewhat earlier than the mean value for WATER4, reflecting the fact that NLA lakes had smaller surface areas than those that could be resolved with satellite data.

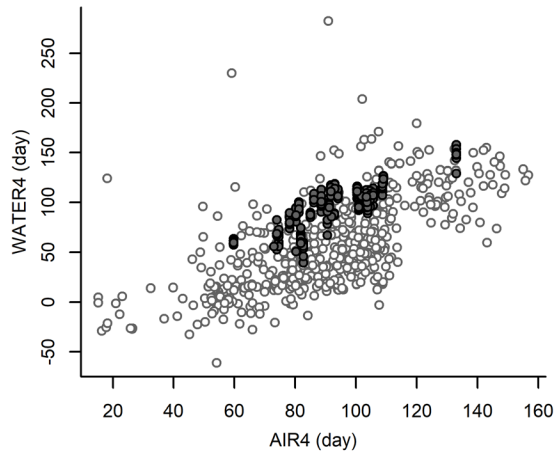

Figure 3. AIR4 vs. WATER4. AIR4: day of the year that mean air temperature is 4°C. WATER4: day of the year that surface lake water temperature is 4°C. Open circles: estimate of spring turnover from NLA hypoxia model. Filled circles: measured spring turnover day from satellite data.

RMS error for predictions of DO with the revised model for spring turnover day and the larger data set was 1.49°C. Running the original model with the larger data set yielded an RMS error of 1.75°C, so the revisions to the model improved prediction accuracy by 15%.

Zooplankton model

Incorporating data from the 2017 NLA increased the data available for analysis, but an additional restriction was imposed, limiting the analysis to lakes deeper than 3m (see below). The final combined data set consisted of 1591 measurements of zooplankton biomass collected at 1106 lakes.

Exploratory analysis

Adding the 2017 NLA data increased the number of measurements of zooplankton biomass to 1625, collected at 1138 sites. The correction to the NLA data for the cowl diameter of the towed net was also incorporated in this reanalysis. While incorporating this new data, possible classification data was re-evaluated. Based on earlier analyses, four candidate classification variables were considered: conductivity, color, lake maximum depth, and seasonal maximum temperature. Lake locational information (latitude, longitude, and elevation) were also evaluated. To assess each classification

variable, a generalized additive model was fit to predict log-transformed zooplankton biomass as a function of log-transformed chlorophyll concentration and the classification variable. For example, the model to evaluate the effect of depth can be written as follows:

$$\log(Z) = s(\log(Chl)) + s(Depth) + Year \quad (7)$$

Where *Z* is zooplankton biomass, *Chl* is Chl *a* concentration, *Depth* is maximum lake depth, and *s(.)* indicates the use of a non-parametric smooth function. The year that the sample was collected was included as an additive factor to account for small systematic differences in zooplankton biomass across the two surveys.

The proportion of variance that was explained by models including each candidate classification variable was retained. Of the four candidate classification variables, temperature improved the model by the greatest amount, yielding an R^2 of 14%. The R^2 values associated with including conductivity, depth, and color in the model were 8.4%, 5.2%, and 3.9%. Including lake locational information yielded R^2 values of 9.9% and 10.7% for longitude and latitude, respectively, but because of the strong association between maximum lake temperature and location, lake seasonal maximum temperature was selected.

Incorporating classification variables into the criterion model

Four temperature classes were defined with similar number of sites within each group. Temperature thresholds delineating the four classes were T ≤ 22.9°C, 22.9 < T ≤ 24.9°C, 24.9 < T ≤ 27.9°C, and > 27.9°C. Shallow lakes (< 3m) were excluded to limit the effects of benthic invertebrates on the zooplankton biomass measurements.

Within each temperature class, the same model as described in the lake criterion document was fit to the data. Because data collected in two different years were available, year was included as a categorical factor in the model.

A moderately informative prior for the breakpoint, c_p, in the relationship between zooplankton biomass and phytoplankton biovolume, was specified as a normal distribution with a mean value that was the same as the overall mean phytoplankton biovolume in the data set. This prior distribution expresses the idea that the change in the slope should be estimated as occurring somewhere within the range of sampled conditions. Moderately informative priors for the model coefficients expressed the theoretical expectations that at low levels of phytoplankton, the slope of the relationships should be near one, and at high levels of phytoplankton, the slope of the relationship should be near zero.

The revised model reduced the RMS prediction error for zooplankton biomass by 7.6%.

TP-TN-Chl models

Including the 2017 NLA data increased the number of samples for the TP and TN models from 2356 to 3434 distinct samples. The median number of samples per ecoregion increased from 19 to 27 samples.

One minor change to the prior distributions for different parameters was introduced, in which a strong prior value of 0.1 was specified for the sampling error of TP and TN measurements. This prior distribution directly reflects the performance standard for lab measurements for these two parameters, and imposing this distribution yields a more accurate estimate of the limiting relationship for Chl-TP and Chl-TN. No other significant changes in the model structure were implemented.

Model performance was very similar to the original model, and the resulting criterion values differed only slightly from those associated with the original model.