Short-term Probabilistic Microcystin Prediction using Bayesian Model Averaging Supporting Information

Song S. Qian[a], Craig A. Stow[b], Sabrina Jaffe[a]

[a]*Department of Environmental Sciences, The University of Toledo 2801 West Bancroft Street Toledo, OH, 43606, USA*
[b]*Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, 4840 South State Road Ann Arbor, MI, 48108, USA*

1. Additional Figures

The evaluation results are presented graphically by superimposing the sequentially updated weekly/biweekly model predictions onto the scatter plot of MC versus Chla of the entire sampling year, with data used for model updating and data for model evaluation highlighted (Figures S1-S3).

At the beginning and, to a lesser extend, near the end of each sampling season, most measured MC values were below the method detection limit of 0.1 μg/L and these censored values were shown in Figures S1-S3 as being 0.1. As a result, the models appeared to have consistently under predicted MC values in the first few sampling periods, an artifact of how censored data were recorded and presented in these figures. This apparent under prediction also occurred in 2021, a year with unusually low level of algal blooms.

Email address: song.qian@utoledo.edu (Song S. Qian)

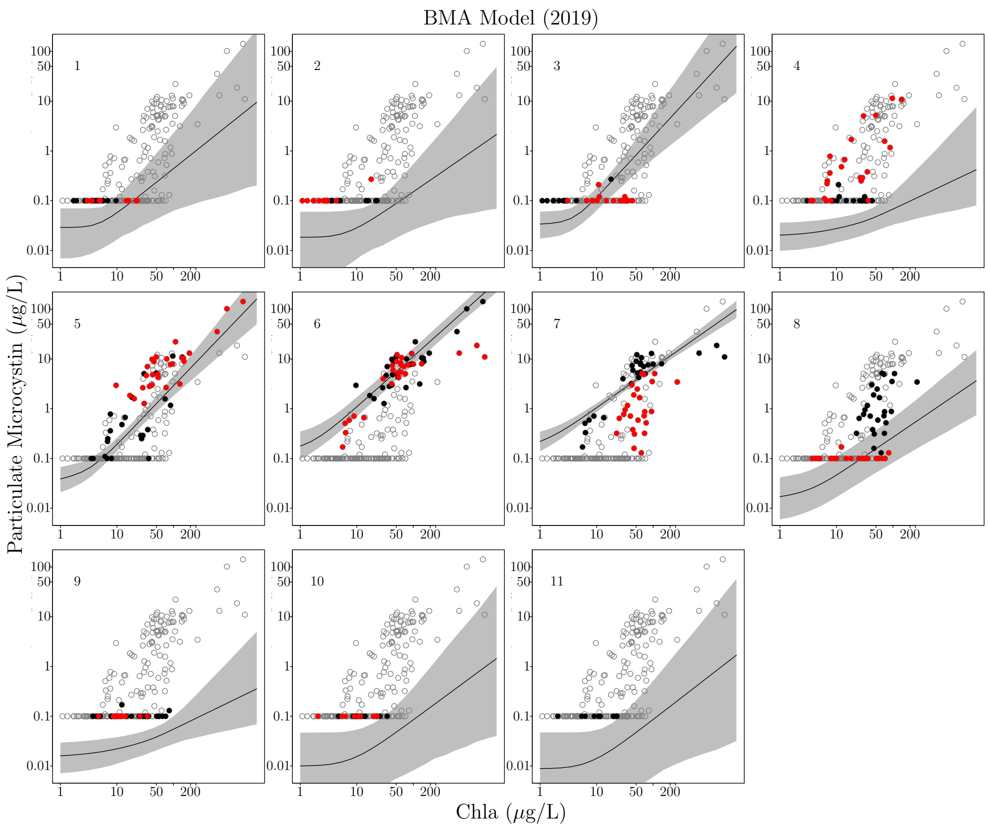

Figure S1: Sequentially updated weekly/biweekly models were fit to data from 2019. From left to right and top to bottom, each panel show a sampling event in chronological order (numbers shown in top left corners). Inside each panel, black dots are data from the sampling event, red dots are data from the next sampling event, and the gray circles are data from the rest of the year. Inside each panel, the four seasonal model predictions are averaged using BMA (the black line), which are intended to predict the red dots in the same panel. The gray polygons are the 95% credible intervals of the fitted mean. Data point with MC concentration of 0.1 μg/L represent values below 0.1.

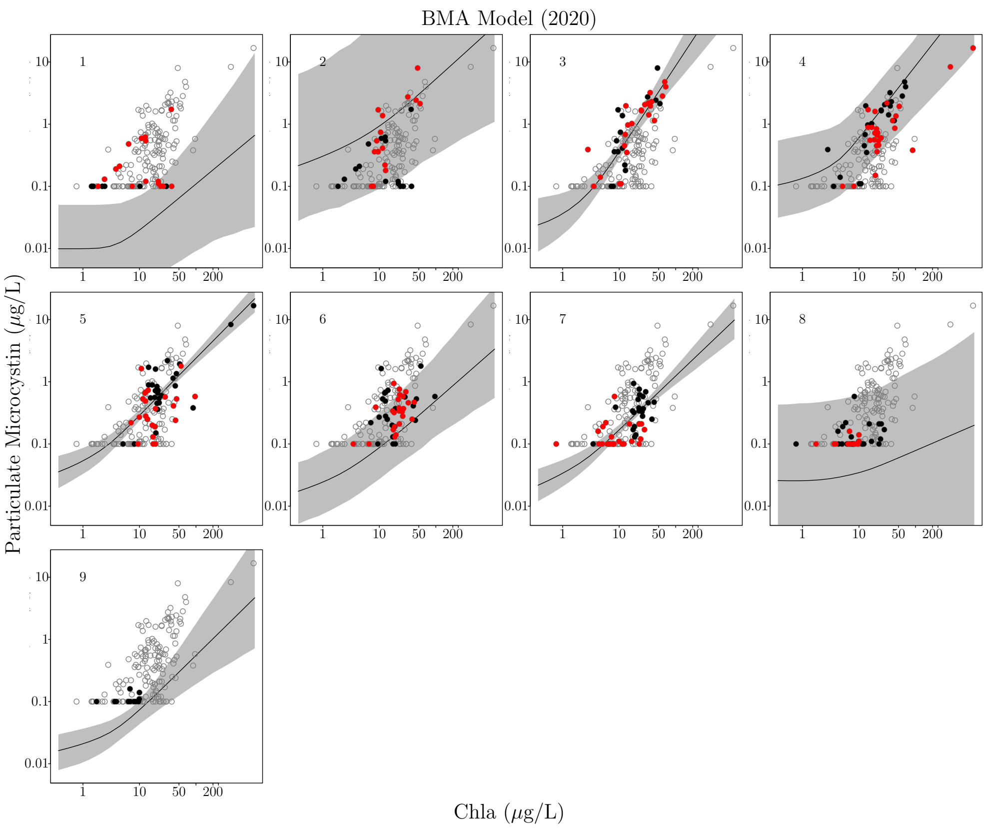

Figure S2: Same as in Figure S1, for 2020.

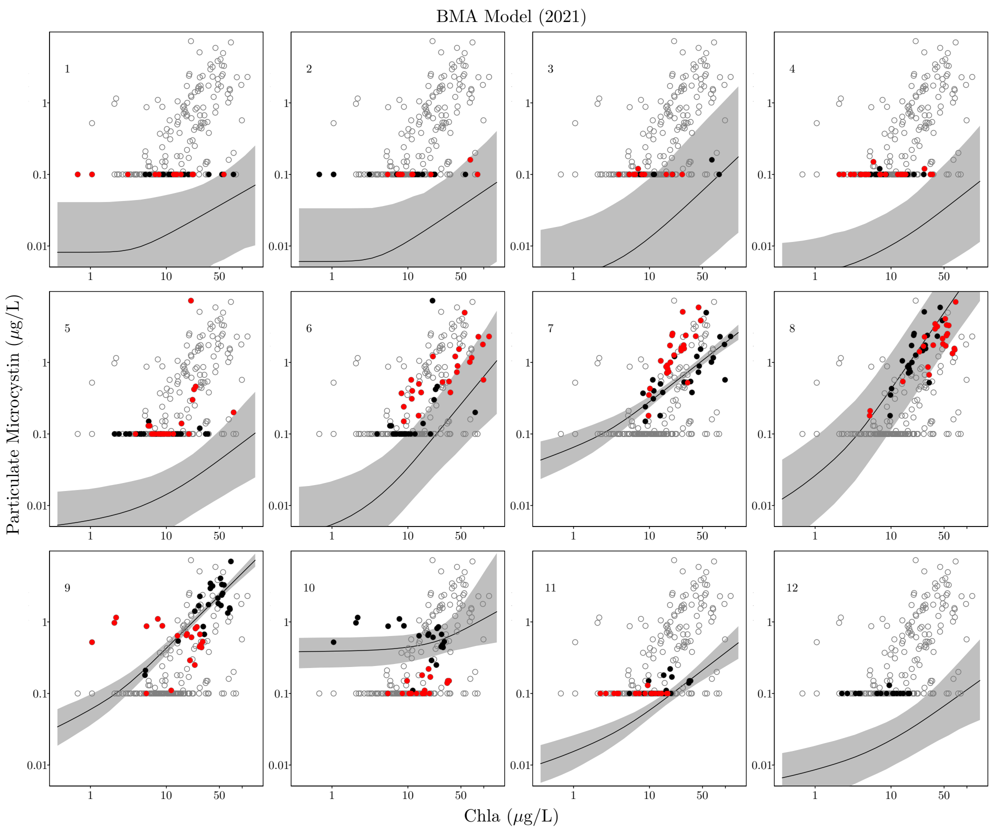

Figure S3: Same as in Figure S1, for 2021.

2. Why Not Bi-weekly Additive Model

There are two reasons for not considering an additive model for biweekly seasonal specification.

- First is a computational reason. We need to define a week as from Monday to Sunday. Consequently, there are upto 54 weeks in a calendar year because January 1 can fall on any weekday. When naming the week as from 1 through 54, date (e.g., May 24) can fall in different weeks in different years, which seems to us as unreasonable. Alternatively, we can specify weeks as consequtive 7 days starting from January 1, resulting in weeks in a year starting on different weekdays.

- Season is a concept representing a temporal aggregation. As such, the temporal scale is a consideration when speaking of seasonal effects. The word season is mostly used to describe seasons as in winter, spring, summer, and fall, an aggregation of 3 months, because of the distinct differences among seasons. As a result, an additive seasonal model is reasonable. (But not very useful for our purposes.) When using calendar month to represent season, such differences are less pronounced from month to month, thereby we found that an additive model is not conssitent with the data. When the time scale is reduced to a week or two weeks, we have no reason to believe an additive seasonal effect, as in the effect of week 24 is always higher than the effect of week 23. As such, an additive model is of no practical use in our modeling situation.

3. Computational Details

The Bayesian sequential updating process is implemented in Stan through the R function `rstan`. Computational details (including R and Stan code and data) can be found in the GitHub repository (`https://github.com/songsqian/ErieMC2023/R`).