The Implications of Simpson's Paradox for Cross-Scale Inference Among Lakes

Song S. Qian,[*,†] Craig A. Stow,[‡] Farnarz Nojavan A.,[¶] Joseph Stachelek,[§] Yoonkyung Cha,[‖] Ibrahim Alameddine,[⊥] and Patricia Soranno[§]

†*Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606*

‡*Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI 48108*

¶*Center for Industrial Ecology, Yale University, New Haven, CT*

§*Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI*

‖*School of Environmental Engineering, University of Seoul, Seoul, South Korea*

⊥*Department of Civil and Environmental Engineering, American University of Beirut, Beirut, Lebanon*

E-mail: song.qian@utoledo.edu

Abstract

Vollenweider's revolutionary work in assessing the cause of lake eutrophication not only implicated phosphorous as the main culprit of algal growth, but also validated the approach of data synthesis by grouping data from multiple lakes. Over the decades since Vollenweider's report to OECD, limnologists routinely use sample averages from numerous individual lakes to examine patterns across lakes. The assumption behind the use of cross-lake data is often that responses within and across lakes are identical. Using two large cross-lake datasets, we demonstrate that this assumption is usually unjustified. Through comparisons of an empirical model of the effect of nutrients on algal growth fit to several data sets, we discuss the cognitive importance of distinguishing factors affecting lake eutrophication operating at different spatial and temporal scales, and present an analytic tool for properly structuring the data analysis when data from multiple lakes are employed.

key words: NLA, LAGOS, multilevel/hierarchical model, chlorophyll a

Introduction

Limnologists have a long history of using data from multiple lakes, summarized at various levels of spatial and temporal aggregation, to estimate empirical models. Dillon and Rigler[1] set an early precedent using sample averages from a combination of 46 lakes, lake years, and segments of lakes to estimate a simple linear regression model relating chlorophyll a (*chla*) concentration to total phosphorus (TP) concentration. Numerous papers followed, applying regression approaches to estimate similar models using data from other lakes, sometimes comparing their estimated equations to the equation obtained by Dillon and Rigler[2–5]. The practice of estimating models using data from multiple lakes is common, fostered by increases in computational capacity and corresponding advances in statistical software which now facilitates the estimation of nonlinear models, using large data sets[6].

These approaches are typically based on an implicit assumption that the *chla* and TP means from multiple lakes can be described by a dose-response equation such as:

$$\log(\mu_{Chla}) = \beta_0 + \beta_1 \log(\mu_{TP}) + \varepsilon \tag{1}$$

where μ_{Chla} is the mean of chla concentration for a specified time period (such as summer of a particular year) and lake (or lake segment), μ_{TP} is the mean TP concentration for a corresponding, but not necessarily the same, time period (spring TP may be related to summer *chla*, for example), β_0 and β_1 are the intercept and slope parameters, respectively, and ε is the model error term usually assumed to be normally distributed with a constant variance. Because the underlying "true" mean values are always unknown, sample averages are typically used as surrogates, although occasionally sample medians have been used (Reckhow 1988). This regression-based modeling approach has influenced lake management practices beyond the modeling of the *chla*-nutrient relationship. For example, Yuan and Pollard[7] used data from the National Lake Assessment (NLA), a cross-lake data set including randomly selected lakes in all 48 contiguous states of the United States[8], to develop a dose-response model to describe the relationship between microcystin (MC) concentration and total nitrogen (TN) concentration. The resulting model was used to propose a national nitrogen criterion for controlling harmful algal blooms.

The implicit premise of this approach is that a relationship estimated using sample averages from many lakes can be applied to set criteria for individual lakes, because criteria compliance assessment is typically lake-specific. However, there are two potential problems with this supposition:

1. Using sample averages as surrogates for the "true," unknown means, violates two assumptions of regression analysis: the variance of the response variable is constant and the predictor variables are observed without error. On the one hand, violating the equal variance assumption makes estimated parameter and model error variances ambiguous; it is unclear what uncertainty bands calculated from these values, such as 95% credible or prediction intervals, represent. On the other hands, violating the observation error assumption has been well-studied; it is widely recognized that this "errors-in-variables" problem causes slope coefficient estimators to be biased toward zero[9,10].

2. Lake-specific factors may cause individual lakes to exhibit differing stressor-response

relationships[2]. Using aggregated measures, such as sample averages to estimate among-lake relationships can produce results that poorly represent the individual lakes in the analysis. In extreme cases, the sign of the estimated slope parameter can be reversed (Figure 1), a situation known as Simpson's Paradox[11]. Clearly, such a model should not be used to develop lake-specific management strategies[12–14].

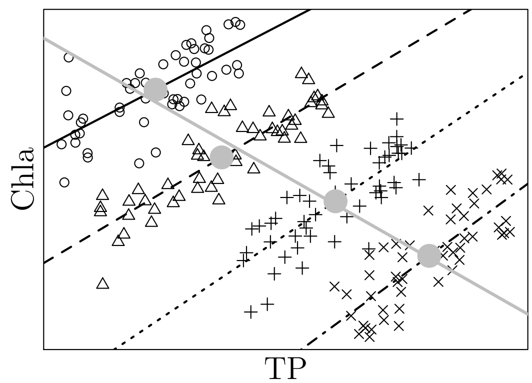

Figure 1: Hypothetical data from four lakes illustrate the worst case scenario for combining lake-means for developing empirical models. Within each lake, *chla* is positively correlated with TP (black lines). The correlation between lakes means of *chla* and TP is, however, negative (shaded dots and line). The best case scenario is realized when the four data sets overlap (four lakes are identical).

Cha and Stow[15] demonstrated a modeling approach that addresses problem 1; in this paper, we use two large data sets to illustrate the potential hazards of using data from multiple lakes without properly addressing the among-lake variation that is often defined as changes in regression model coefficients when the model is fit to data from different lakes. The among-lake variation can also be reflected in the changes in model coefficients when the same model is fit using two data sets collected using the same protocol, even when the number of lakes included in the data is large. We illustrate the effects of among-lake variation on regression-based lake models by comparing models fit using lake sample averages from several cross-sectional data sets. We then present a Bayesian hierarchical modeling (BHM) approach for the hierarchical data structure and an empirical Bayes interpretation of a BHM's hyper-parameter distribution to facilitate the use of cross-lake data for lake-specific inference.

Materials and Methods

Data

We used data from both the National Lakes Assessment (NLA) conducted by the US Environmental Protection Agency (EPA)[16,17] and the LAke multiscaled GeOSpatial and temporal database (LAGOSNE)[18] to illustrate potential statistical issues that may arise when analyzing large data sets encompassing multiple lakes. The NLA consists of 1152 lakes sampled in 2007 (NLA2007) and 1099 lakes sampled in 2012 (NLA2012). Data were collected in each year using an identical sampling protocol. Lakes included in the NLA were selected using a probabilistic sampling design in an attempt to accurately represent the overall population of lakes in the United States. In contrast to the NLA, the LAGOSNE database contains information on lakes with monitoring data from federal, state, or citizen science monitoring programs across 17 states in the north-east of the US. We used 27 lakes from LAGOSNE that were also included in NLA2007 for detailed analysis. These lakes have at least 10 observations in LAGOSNE (Figure 2). The selection of these 27 lakes was for the purpose of methods comparison only. A summary of the data is in Table 1.

Table 1: Summary of data used in the analysis

	NLA2007	NLA2012	LAGOSNE
# obs.	1328	1230	1340
# of lakes	1152	1099	27
# obs per lake	1-2	1-2	17-192
# of years	1	1	9-29

Data from LAGOSNE represent the 27 lakes with more than 10 observations that are also present in NLA2007.

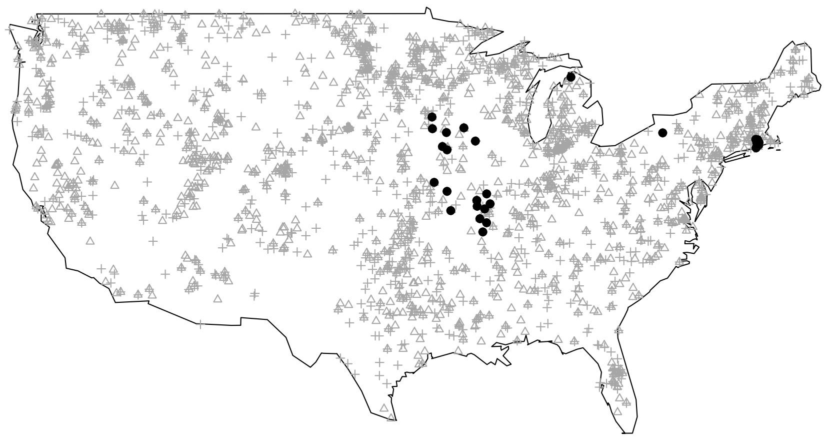

Figure 2: Locations of NLA2007 lakes (pluses), NLA2012 lakes (triangles), and the 27 lakes included in both NLA2007 and LARGOENE (black dots)

These data sets were used to illustrate (1) the effects of among-lake variation on regression-based lake modeling and (2) the Bayesian hierarchical modeling approach for properly account for the among-lake variation.

The two NLA data sets include a large number of lakes and were collected to be representative of lakes in the US. Using these two data sets, we illustrate how the among-lake variation may be reflected in regression models fit using the data sets separately, and fit to the combined data. To contrast the NLA which includes only a small number of observations for each lake (such that lakes means are highly variable), we compare the three models fit using NLA data sets to a model fit to a subset of LAGOSNE that includes 27 lakes that are represented in NLA2007. For this comparison, we use lake mean concentrations of *chla*, TP, and TN as the observations for developing the regression model discussed in the next section.

Using data of the 27 lakes in LAGOSNE we show how Bayesian hierarchical modeling approach can be used to partially pool data from different lakes to avoid the potential problems of Simpson's paradox(Figure 1).

Statistical Modeling

Illustrating Among-Lake Variation in Model Coefficients

We first developed a regression model (equation (2)) to demonstrate the variability of model coefficients between data sets. The model used both TP, TN, and their interaction as predictor variables:

$$\log(chla_j) = \beta_0 + \beta_1 \log(TP_j) + \beta_2 \log(TN_j) + \beta_3 \log(TP_j)\log(TN_j) + \varepsilon_j \qquad (2)$$

where $chla_j$, TP_j, and TN_j are sample average concentrations for chla, TP, and TN for the jth lake. Frequently, TP is used as the only predictor because phosphorus is usually assumed as the limiting nutrient; we did not make that *a priori* assumption for all the lakes in the data[19]. Furthermore, TP and TN are often correlated, which can imply an interaction effect[20]. For example, an oligotrophic lake may be limited by both phosphorus and nitrogen; thus increasing phosphorus may lead to an increased nitrogen demand, constituting a positive interaction. In an analysis of Finnish lakes, Malve and Qian[19] and Qian[20] showed that including both TP and TN, and their interaction term can lead to a more informative model. Specifically, the magnitude of the coefficient β_3 may be indicative of a lake's trophic level[20]. A lake is likely to be oligotrophic when $\beta_3 > 0$ (both P and N are limiting), mesotrophic when $\beta_3 \approx 0$ (P is likely the limiting nutrient), and eutrophic when $\beta_3 < 0$ (perhaps neither P nor N is limiting). Because of the inclusion of the interaction term, the effects of TP and TN on *chla* are no longer constants. The effect of TP depends on the value of TN and vice versa. The meanings of software reported values of β_1 and β_2 are the TP and TN effects for specific values of TN and TP, respectively[20]. Specifically, the reported β_1 (β_2) is the TP (TN) effect when $\log(TN) = 0$ ($\log(TP) = 0$). In this paper, we centered both predictors by subtracting the respective log means of TP and TN; such that, the reported slopes (i.e., $\hat{\beta}_1$ and $\hat{\beta}_2$) are the TP and TN effects when the other predictor value is at the geometric mean of 27 LAGOSNE lakes. Because the geometric means of 27 LAGOSNE lakes do not have the same reference value for all lakes (e.g., the geometric mean of TP represents a high phosphorus level for some lakes and a low level for other lakes), software reported β_1 and β_2 values are not comparable among lakes. Consequently, we focus on the comparisons of β_0 and β_3.

Using BHM to Account for Among-Lake Variation

Next, we developed a Bayesian hierarchical or multilevel model to incorporate the hierarchical structure inherent in multi-lake data. We constructed a two-tier multilevel model; at the lake level, we use a form of equation (2):

$$\log(chla_{ij}) = \beta_{0j} + \beta_{1j}\log(TP_{ij}) + \beta_{2j}\log(TN_{ij}) + \beta_{3j}\log(TP_{ij})\log(TN_{ij}) + \varepsilon_{ij} \quad (3)$$

where the subscript ij represents the ith observation from the jth lake. Above the individual lake level, we capture the variation of among lake-specific model coefficients. As the regression model represents a basic well-studied limnological relationship, we expect that the log-log linear relationship to hold for all lakes, but that model coefficients $\beta_{0:3j}$ may differ by lake. Statistically, these lakes are regarded as exchangeable because without additional information we would not know how these coefficients might differ. Thus, the lake-specific model coefficients are modeled as random variables from a common distribution:

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix} \sim MVN \left[\begin{pmatrix} \mu_{\beta_0} \\ \mu_{\beta_1} \\ \mu_{\beta_2} \\ \mu_{\beta_3} \end{pmatrix}, \Sigma \right] \tag{4}$$

where MVN represents a multivariate normal distribution. Equations (3) and (4) form a two-tier hierarchical model. The multivariate normal distribution on the right-hand-side of equation (4) is often known as the hyper-parameter distribution. The rationale of using the BHM is discussed by Qian et al.[21] in the context of estimating mean concentrations of water quality variables for multiple water bodies. Compared to coefficients estimated using lake-specific data (one lake at a time), BHM estimated model coefficients are more accurate overall. More importantly, the hierarchical model specified in equations (3) and (4) separates within-lake models (specified by $\beta_{0:3j}$) from the among-lake model ($\mu_{\beta_{0:3j}}$). As a result, a lake-specific inference can be made more accurately[22].

Modeling Road Map

Our analyses consist two parts:

1. The model of equation (2) was fit to lake sample average *chla*, TP, and TN concentrations from (1) NLA2007 data alone, (2) NLA2012 alone, (3) combined NLA2007 and NLA2012 data, and (4) LAGOSNE to illustrate the variability of the estimated model coefficients.

2. The hierarchical model of equations (3) and (4) was fit using data from the 27 lakes in LAGOSNE to demonstrate the use of a BHM for properly account for the among-lake variation.

All models were fit with log TP and log TN centered at the respective log means of TP and TN of the 27 lakes in LAGOSNE. As a result, the intercept (β_0) of these models represents the log mean *chla* concentrations when TP and TN are at the (log) mean levels of the 27 lakes (log TP mean of 3.112, or geometric mean of 22.5 μg/L, and log TN mean of 6.296, or geometric mean of 542.7 μg/L).

All statistical models were implemented in R[23], using function `lm()` for linear regression models and the function `lmer` from package `lme4`[24] for BHM in equations (3) and (4). Annotated R code can be found at GitHub (https://github.com/songsqian/simpsons).

Results

Variability in Model Coefficients

The linear model fit to the 27 LAGOSNE lakes has a much smaller $\hat{\beta}_3$, as compared to the same of the three linear models fit to NLA2007, NLA2012, and NLA2007+NLA2012 (Figure 3, Table 2). In addition, the LAGOSNE model coefficients have much larger standard errors because the LAGOSNE model is based on 27 sets of lake sample average concentrations ($n = 27$) whereas the three NLA models are based on sample averages from over 1000 lakes. The estimated model coefficients based on NLA2007 and NLA2012 also differ, and the model based on the combined NLA data is closer to coefficients of the model fit to NLA2012. The interpretations of these model coefficients, especially the slopes, are ambiguous. β_0 is the expected log *chla* for lakes with TP and TN concentrations near the respective geometric means of the 27 LAGOSNE lakes. However, the meanings of the three slopes of these models are no longer clear. Mathematically, β_1 is the expected change in $\log(chla)$ for every unit change in $\log(TP)$, while TN is held unchanged. By using a regression model, we assume that changes in $\log(chla)$ due to factors not included in the model will not affect the estimated slope and can be lumped into the error term. This assumption, however, requires that the within-lake and among-lake relationship between $\log(chla)$ and $\log(TP)$ be the same. As shown in the four hypothetical lakes in Figure 1, this assumption is likely unrealistic. The ambiguity of model coefficients manifested in the differences among the estimated coefficients of the four models, suggests that the practice of using lake means for developing an empirical model is potentially misleading. The difference in the estimated model coefficients from the two data sets collected for the same purposes (NLA2007 and NLA2012) suggests that the best case scenario (Figure 1) is highly unlikely.

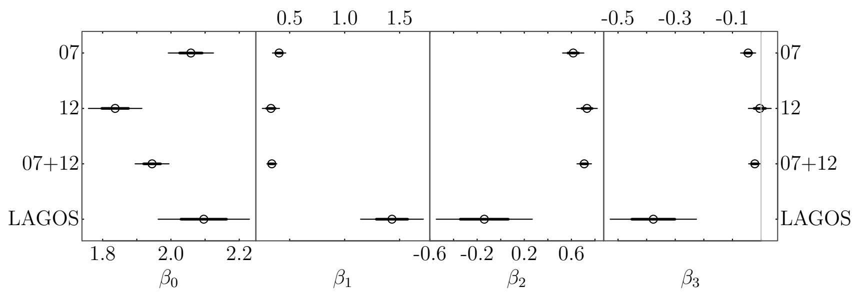

Figure 3: Model coefficients ($\beta_{0:3}$) estimated using lake mean concentrations from NLA2007 (07), NLA2012 (12), NLA2007 and NLA2012 combined (07+12), and the 27 LAGOSNE lakes (LAGOS). Dots are the estimated means and thin and thick horizontal lines are the mean plus one and two standard errors, respectively. The shaded vertical line references $\beta_3 = 0$.

BHM for Among-Lake Variation

The hierarchical model fit to data from the 27 LAGOSNE lakes shows a large among-lake variation in model coefficients (Figure 4). The estimated intercepts ($\hat{\beta}_0$) are the expected log *chla* concentration for these 27 lakes when they all have the same TP and TN concentrations (the respective geometric means). As such, values of β_0 in Figure 4 show the relative productivity of the 27 lakes (sorted based on their intercept values). The visible opposite trends between β_0 and β_3 are indicative of the value of β_3 in understanding a lake's trophic level. Because the value of β_0 is dependent on the baseline values of TP and TN, while the value of β_3 is invariant, the interaction slope

β_3 is a more direct indicator of a lake's trophic status. The wide range of β_3 shows that these lakes have different trophic levels, indicating that nutrient effects on lake primary productivity vary by lake.

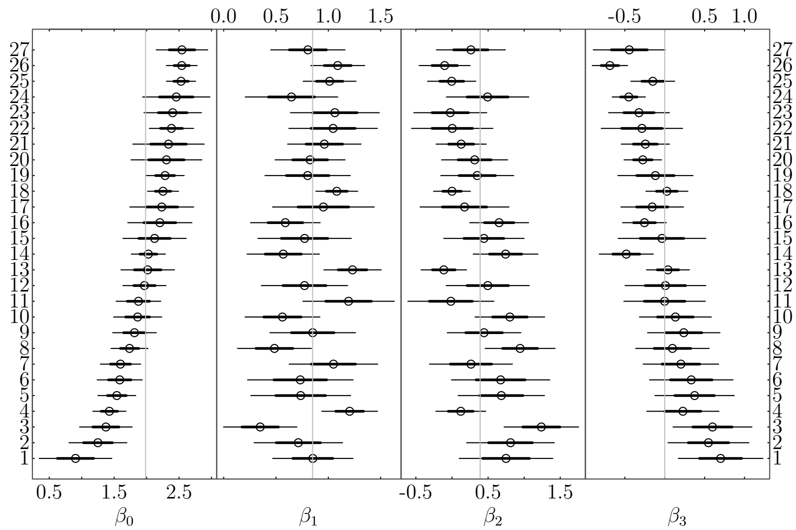

Figure 4: BHM estimated lake-specific model coefficients (β_{0j} – β_{3j}) shown a strong negative correlation between β_{0j} and β_{3j}. Dots are the estimated means and thin and thick horizontal lines are the mean plus one and two standard errors, respectively. The shaded vertical lines for β_0, β_1, and β_2 show the estimated respective hyper-parameters ($\mu_{\beta_0}, \mu_{\beta_1}$, and μ_{β_2}), the vertical line in the β_3 panel references $\beta_3 = 0$.

Table 2: Model Coefficients Estimated Using Different Methods

Models	07	12	07+12	LAGOS	BHM
β_0	2.058 (0.033)	1.837 (0.039)	1.9448 (0.025)	2.096 (0.067)	1.984 (0.098)
β_1	0.404 (0.030)	0.330 (0.039)	0.3376 (0.022)	1.430 (0.143)	0.850 (0.073)
β_2	0.616 (0.045)	0.732 (0.044)	0.7088 (0.031)	-0.139 (0.204)	0.390 (0.104)
β_3	-0.045 (0.013)	-0.004 (0.020)	-0.0218 (0.011)	-0.377 (0.075)	-0.014 (0.091)

Estimation standard errors are in parentheses. Models: "07" is the model fit to NLA2007 data, "12" is fit to NLA2012, "07+12" is fit to the combined NLA data, "LAGOS" is fit using the mean concentrations of the 27 lakes from LAGOSNE, BHM is the Bayesian hierarchical model (hyper-parameters, μ_β's).

The difficulty in interpreting linear regression model slopes disappears when the coefficients are allowed to differ by lake. The hierarchical model estimated $\beta_{0:3j}$ are lake-specific, while the hyper-parameters $\mu_{\beta_{0:3}}$ are the means of the respective lake-specific coefficients. Consequently, the meaning of these estimated coefficients is unambiguous.

Conclusions and Discussion

We showed that empirical models based on mean concentrations of *chla*, TP, and TN from NLA2007 and NLA2012 have visibly different coefficients. Lakes in both NLA2007 and NLA2012 were selected based on a probabilistic sampling protocol such that analytical results can be "(extrapolated) to national scales"[8]. It is tempting to interpret the difference in model coefficients between NLA2007 and NLA2012 as a result of improved overall lake condition from 2007 to 2012. Yet, Because these coefficients were estimated using lake sample average concentrations of *chla*, TP, and TN, we cannot directly interpret the differences in the models of NLA2007 and the model of NLA2012 as a direct result of changes in lake conditions over time. A more reasonable explanation of these difference is the random sampling variability. Furthermore, the large variability in lake-specific model coefficients as shown in Figure 4 suggests that an overall "average" model is unlikely to be informative, especially for developing management strategies that will be implemented to individual lakes.

Many early lake water quality models were based on simple mechanistic principles and were parameterized using statistical methods[25]. These models relied on data from multiple lakes, with each lake or lake segment contributing one observation[26]. As we accumulated a larger amount of data from multiple lakes, these simple modeling methods are increasingly being used as the basis for analyzing cross-sectional data. In the age of fast computers, the successful tools of the past can be easily applied to big data. In this paper, we used a common regression model in the limnological literature to demonstrate the potential problems of treating "big" (multiple lakes) data using conventional methods. The hierarchical structure in the data (i.e., from individual observations to lake-specific features to regional characteristics shared by many lakes) should be properly reflected in our empirical models. The Bayesian hierarchical modeling approach provides a flexible tool for modeling the hierarchical structure inherent to most of our "big data."

Without properly modeling the hierarchical structure, we risk misinterpreting the data (e.g., Figure 1), a situation has long been recognized in statistics as the Simpson's paradox[11]. Although the mathematics behind the Simpson's paradox is straightforward, the implications of the paradox are still not widely recognized in our field. Frequently, we do not analyze data at different levels of aggregation, thereby we fail to notice the paradoxical phenomenon, which can lead to misinterpretation of the results. Lakes are naturally different (Figure 4); forcing a single model on all lakes is undesirable.

When used to develop management strategies for eutrophication control, models based on lake mean concentrations are likely to fail when used in compliance assessment. Developing "national" nutrient criteria is likely counterproductive as nutrient concentrations are only one of many factors affecting a lake's trophic status. A national standard would be inevitably too stringent for some lakes and too loose for others. When the among-lake variance is considered as in Yuan and Pollard[7], the resulting criterion is most likely too stringent, and thereby unachievable, for most lakes. Consequently, a lake-specific approach is necessary.

When developing models for individual lakes, mathematical theories show that a Bayesian estimator with a proper (informative) prior is always better (compared to a non-Bayesian estimator) in terms of a model's predictive accuracy[27,28]. In fact, Bayes himself showed that the Bayes estimator minimizes the squared error associated with both observed means and the underlying true mean[29]. In a regression problem, errors associated with the observed means are the residuals. A regression model would minimize the residual sum of squares; whereas, a Bayesian regression model would also minimize the error associated with the estimated model coefficients. The difficulty in using a Bayesian method is in obtaining informative priors. The BHM approach suggests that such informative prior can be obtained by analyzing data from multiple lakes. The hyper-parameter distribution (right-hand-side of equation (4)) is naturally such a proper prior. In other words, an important and valuable result of analyzing data from multiple lakes is the hyper-parameter distribution, which can be used as a proper informative prior for analyzing data from individual lakes that are not included in the data used to develop the hierarchical model. This conclusion is not limited to limnological modeling[21].

Our analyses suggest that data such as NLA may be ill-suited for developing lake-specific *chla*-nutrient models because of the limited lake-specific sample size. In fact, with only 10% of the lakes were sampled twice[8], fitting BHM is impossible. This outcome is not surprising because the NLA program was designed to answer two questions (what is the current condition of lakes? and how is this condition changing over time?) that are not directly related to the quantification of the *chla* nutrient relationship[8]. The goals of the NLA monitoring program are similar to those of EPA's Environmental Monitoring and Assessment Program (EMAP), which is optimized for estimating the mean and variance of individual environmental/ecological indicators over a national/regional scale, or of a stratified subpopulation (e.g., small lakes)[30]. These programs are purposefully designed to best support a limited number of objectives[31]. As a result, when data from programs such as EMAP and NLA are used beyond their original design goals, we need to incorporate these data collection design parameters and plan our analysis accordingly.

In this paper, our objectives were to (1) illustrate the potential problems of developing empirical models using cross-lake data and (2) demonstrate the use of BHM for properly modeling the hierarchical structure of the data. Although the data we used are ideal for both objectives, our BHM model from LAGOSNE may not be of any practical interest because the 27 lakes were selected to illustrate the potential issues and for demonstrating methods. These lakes do not represent any particular subpopulation of lakes. That is, the resulting models are of no particular practical purposes. For the estimated hyper-parameter distribution to be practically meaningful, lakes used for developing the hierarchical model should be selected to represent the subpopulation of interest. As such, the values of large cross-lake data such as NLA lie in their wide coverage that can be used to guide stratifying lakes into subpopulations, within which lakes are "exchangeable," to facilitate the proper data selection for lake-specific inference. This process of careful data selection is necessitated by the recognition that "correlation does not imply causation" (commonly attributed to the Irish philosopher George Berkeley); statistical analysis of observational data must be done only after properly balancing "confounding factors"[32–34] and in the context of intended goals.

References

(1) Dillon, P. J.; Rigler, F. H. Phosphorus-Chlorophyll Relationship in Lakes. *Limnology and Oceanography* **1973**, *19*, 767–773.

(2) Jones, J.; Bachmann, R. Prediction of Phosphorus and Chlorophyll Levels in Lakes. *Journal of Water Pollution Control Federation* **1976**, *48*, 2176–2182.

(3) Canfield, D.; Bachmann, R. Prediction of Total Phosphorus Concentrations, Chlorophyll a, and Secchi Depths in Natural and Artificial Lakes. *Canadian Journal of Fisheries and Aquatic Sciences* **1981**, *38*, 414–423.

(4) Canfield, D. Prediction of chlorophyll a concentrations in Florida lakes: The importance of phosphorus and nitrogen. *Journal of the American Water Resources Association* **1983**, *19*, 255–262.

(5) Prepas, E.; Trew, D. Evaluation of the Phosphorus–Chlorophyll Relationship for Lakes Off the Precambrian Shield in Western Canada. *Canadian Journal of Fisheries and Aquatic Sciences* **1983**, *40*, 27–35.

(6) Filstrup, C.; Wagner, T.; Soranno, P.; Stanley, E.; Stow, C.; Webster, K.; Downing, J. Regional variability among nonlinear chlorophyll—phosphorus relationships in lakes. *Limnology and Oceanography* **2014**, *59*, 1691–1703.

(7) Yuan, L. L.; Pollard, A. I. Using National-Scale Data To Develop Nutrient–Microcystin Relationships That Guide Management Decisions. *Environmental Science and Technology* **2017**, *51*, 6972–6980.

(8) Pollard, A.; Hampton, S.; Leech, D. The promise and potential of continental-scale limnology using the U.S. Environmental Protection Agency's National Lake Assessment. *Limnology and Oceanography Bulletin* **2018**, *May*, 36–41.

(9) Fuller, W. *Measurement Error Models*; Wiley Series in Probability and Statistics; Wiley: New York, p 440.

(10) Carroll, R.; Ruppert, D.; Stefanski, L.; Crainiceanu, C. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*; Chapman & Hall/CRC Monographs on Statistics & Applied Probability; CRC Press, 2006; p 488.

(11) Simpson, E. The interpretation in contingency table. *Journal of Royal Statistical Society (B)* **1951**, *13*, 238–241.

(12) Smith, V. H.; Shapiro, J. Chlorophyll-phosphorus relations in individual lakes. Their importance to lake restoration strategies. *Environmental Science and Technology* **1981**, *15*, 444–451.

(13) Reckhow, K. A random coefficient model for chlorophyll-nutrient relationships in lakes. *Ecological Modelling* **1993**, *70*, 35 – 50.

(14) Liang, Z.; Chen, H.; Wu, S.; Zhang, X.; Yu, Y.; Liu, Y. Exploring Dynamics of the Chlorophyll a-Total Phosphorus Relationship at the Lake-Specific Scale: a Bayesian Hierarchical Model. *Water, Air, & Soil Pollution* **2018**, *229*, 21.

(15) Cha, Y.; Stow, C. A Bayesian network incorporating observation error to predict phosphorus and chlorophyll a in Saginaw Bay. *Environmental Modelling and Software* **2014**, *57*, 90–100.

(16) U.S. EPA, *National Lakes Assessment: A Collaborative Survey of the Nation's Lakes*; 2009.

(17) U.S. EPA, *National Lakes Assessment 2012: A Collaborative Survey of Lakes in the United States*; 2016.

(18) Soranno, P. A. et al. Building a multi-scaled geospatial temporal ecology database

from disparate data sources: fostering open science and data reuse. *GigaScience* **2015**, *4*, 28.

(19) Malve, O.; Qian, S. Estimating Nutrients and Chlorophyll a Relationships in Finnish Lakes. *Environmental Science and Technology* **2006**, *40*, 7848–7853.

(20) Qian, S. *Environmental and Ecological Statistics with R*, 2nd ed.; Chapman and Hall/CRC Press, 2016.

(21) Qian, S.; Stow, C.; Cha, Y. Implications of Stein's Paradox for Environmental Standard Compliance Assessment. *Environmental Science and Technology* **2015**, *49*, 5913–5920.

(22) Stow, C. A.; Lamon, E. C.; Qian, S. S.; Soranno, P. A.; Reckhow, K. H. In *Real World Ecology: Large-Scale and Long-Term Case Studies and Methods*; Miao, S., Carstenn, S., Nungesser, M., Eds.; Springer New York: New York, NY, 2009; pp 111–136.

(23) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2018.

(24) Bates, D.; Maechler, M. lme4: Linear mixed-effects models using S4 classes. 2010; R package version 0.999375-33.

(25) Reckhow, K.; Chapra, S. *Engineering Approaches for Lake Management: Data analysis and empirical modeling*; Ann Arbor Science, Butterworth Publishers, 1983; Vol. 1.

(26) Stow, C. A.; Reckhow, K. H. Estimator Bias in a Lake Phosphorus Model with Observation Error. *Water Resources Research* **1996**, *32*, 165–170.

(27) Efron, B.; Morris, C. Stein's Paradox in Statistics. *Scientific American* **1977**, *236*, 119–127.

(28) Efron, B. Controversies in the foundations of statistics. *The American Mathematical Monthly* **1978**, *85*, 231–246.

(29) Berger, J. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer-Verlag: New York, 1985.

(30) Overton, W.; Stehman, S. Desirable design characteristics for long-term monitoring of ecological variables. *Environmental and Ecological Statatistics* **1996**, *3(4)*, 349–361.

(31) Messer, J.; Linthurst, R.; Overton, W. An EPA program for monitoring ecological status and trends. *Environmental Monitoring and Assessment* **1991**, *17(1)*, 67–78.

(32) Rubin, D. *Matched Sampling for Causal Effects*; Cambridge University Press, Cambridge, UK, 2006.

(33) Qian, S.; Harmel, R. Applying statistical causal analyses to agricultural conservation: a case study examining P loss impact. *Journal of the American Water Resources Association* **2016**, *52*, 198–208.

(34) Nummer, S.; Qian, S.; Harmel, R. A meta-analysis on the effect of agricultural conservation practices on nutrient loss. *Journal of Environmental Quality* **2018**,