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

Song S. Qian[a,*], Craig A. Stow[b], Farnaz Nojavan A.[c], Joseph Stachelek[f], Yoonkyung Cha[d], Ibrahim Alameddine[e], Patricia Soranno[f]

[a]*Department of Environmental Sciences, University of Toledo, 2801 W. Bancroft Street, MS# 604, Toledo, OH, USA.*
[b]*Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI, USA*
[c]*Center for Industrial Ecology, Yale University, New Haven, CT, USA*
[d]*School of Environmental Engineering, University of Seoul, Seoul, South Korea*
[e]*Department of Civil and Environmental Engineering, American University of Beirut, Beirut, Lebanon*
[f]*Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA*

Abstract

Using cross-sectional data for making ecological inference started as a practical means of pooling data to enable meaningful empirical model development. For example, limnologists routinely use sample averages from numerous individual lakes to examine patterns across lakes. The basic assumption behind the use of cross-lake data is often that responses within and across lakes are identical. As data from multiple study units across a wide spatiotemporal scale are increasingly accessible for researchers, an assessment of this assumption is now feasible. In this study, we demonstrate that this assumption is usually unjustified, due largely to a statistical phenomenon known as the Simpson's paradox. Through

*Corresponding author, Phone: 419.530.4230, Fax: 419.530.4421
Email address: song.qian@utoledo.edu (Song S. Qian)

comparisons of a commonly used empirical model of the effect of nutrients on algal growth developed using several data sets, we discuss the cognitive importance of distinguishing factors affecting lake eutrophication operating at different spatial and temporal scales. Our study proposes the use of the Bayesian hierarchical modeling approach to properly structure the data analysis when data from multiple lakes are employed.

Keywords: NLA, LAGOSSE, multilevel/hierarchical model, chlorophyll a

1. Introduction

Ecologists have a long history of using data from multiple lakes, summarized at various levels of spatial and temporal aggregation, to estimate empirical models (Vollenweider , 1968, 1975, Schindler, 1977, Wagner et al., 2011). Dillon and Rigler (1973) set an early precedent using reported sample averages from a combination of 46 North American 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 (Jones and Bachmann, 1976, Canfield and Bachmann, 1981, Canfield, 1983, Prepas and Trew, 1983). 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 (Filstrup et al., 2014).

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 (e.g., McCauley et al. (1989)) 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 (Qian, 2016). 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 (2017) 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 (Pollard et al., 2018), 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, we see 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 an estimated parameter and model error variances ambiguous; it is unclear what uncertainty bands calculated from these values, such as 95% confidence or prediction intervals, represent. On the other hands, the consequence of 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 (Fuller, 1987, Carroll et al., 2006).
2. Lake-specific factors may cause individual lakes to exhibit differing stressor-response relationships (Jones and Bachmann, 1976, Wagner et al., 2011, Malve and Qian, 2006). 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), an example of Simpson's Paradox (Simpson, 1951). Clearly, such a model should not be used to develop lake-specific management strategies (Smith and Shapiro, 1981, Reckhow, 1993, Liang et al., 2018).

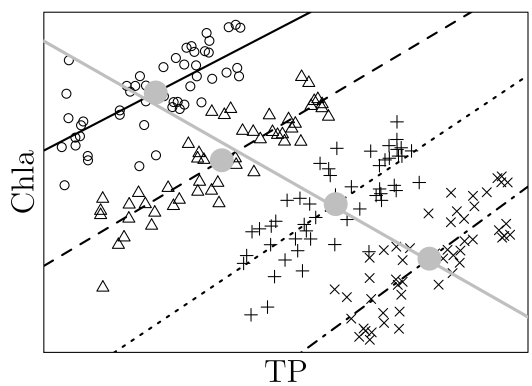

Figure 1: Hypothetical data from four lakes illustrate the worst case scenario for combining lake-means for developing empirical models. Within each lake, *chla* is positively correlated with TP (black lines). The correlation between lakes means of *chla* and TP is, however, negative (shaded dots and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

Simpson's paradox is a well-discussed topic in social and political sciences. An early case was the Berkeley graduate admission paradox (Bickel et al., 1975), where the campus-wide aggregated graduate admission rate showed a bias against female applicants, whereas disaggregated data showed neutral or favorable rates towards female applicants in most departments. More recently, the apparent switch of allegiance of the two major US political parties (blue states are more affluent than red states) was contradicted by data showing that wealthy people are more likely to vote for Republican candidates (Gelman, 2009). There are numerous statistical studies on the topic, with two that are particularly helpful in developing strategies to avoid the paradox. Lindley and Novick (1981) explained the paradox from a statistical inference perspective, that is, statistical inference is the application of a model developed based on data from the population to a new individual. They suggested that the cause of Simpson's paradox is that the new individual is not "exchangeable" with individuals in the population. In Figure 1, we present two groups of models: models for individual lakes and the model of lake means. From a statistical inference perspective, both groups of models are valid. But the models are intended for two different populations: individual observations in a particular lake and lake means of *chla* and TP. The model developed using lake means may give the false impression that *chla* and TP are inversely correlated. Such inverse correlations can often be explained by factors not included in the model, as suggested by Pearl et al. (2016): Simpson's paradox is a problem of confounding factors and thus can be easily resolved under a causal inference framework, where effects of these confounders are explicitly accounted through the use of a causal diagram. This conclusion is supported by many cross-scale studies. For example, Li et al. (2019) show that parameters of a precipitation-stream flow model vary by region due to region-specific confounding factors.

In lake eutrophication studies, quantifying the effects of nutrients (nitrogen and phosphorous) on algal growth is almost always the primary concern, given that excessive nutrient input is a well-established cause of algal proliferation. If we can identify important confounding factors of this relationship, than adopting the causal inference approach is likely more suitable. When analyzing data from multiple lakes (as in Figure 1), each lake may have different confounding factors, statistical inference using a hierarchical modeling approach, such as the ones used in Cha and Stow (2014) may be more effective.

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 the among-lake variation on regression-based lake models by comparing models fit using lake sample averages from several cross-sectional datasets. 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. As the BHM approach is consistent with the shrinkage estimator of Stein's paradox (Qian et al., 2015), our paper provides a Stein's paradox solution to a Simpson's paradox problem.

2. Materials and Methods

2.1. Data

We used data from both the National Lakes Assessment (NLA) conducted by the US Environmental Protection Agency (EPA) (U.S. EPA, 2009, 2016) and the LAke multiscaled GeOSpatial and temporal database (LAGOSNE) (Soranno et al., 2017) to illustrate potential statistical issues that may arise when analyzing large data sets encompassing multiple lakes. The NLA consists of 1,152 lakes sampled in 2007 (NLA2007) and 1,099 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 northeast 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
No. of obs.	1328	1230	1340
No. of lakes	1152	1099	27
No. of obs per lake	1-2	1-2	17-192
No. of years	1	1	9-29

NLA2007: data from 2007 NLA; NLA2012: data from 2012 NLA; LAGOSNE: data from 27 lakes in LAGOSNE with more than 10 observations that were also present in NLA2007.

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 to 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

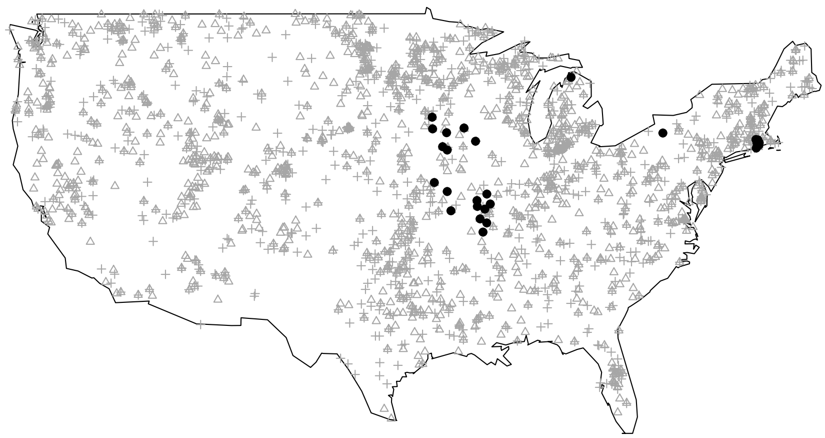

Figure 2: Locations of NLA2007 lakes (shaded pluses), NLA2012 lakes (shaded triangles), and the 27 lakes included in both NLA2007 and LAGOSNE (black dots)

models developed 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 (models developed based on NLA2007, NLA2012, and NLA2007+NLA2012) to a model fit to a subset of LAGOSNE that includes 27 lakes that are represented in NLA2007 with at least 10 observations in each lake. 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).

2.2. Statistical Modeling

2.2.1. 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:

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

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 (Malve and Qian, 2006). Furthermore, TP and TN are often correlated, which can imply an interaction effect (Qian, 2016). 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. The most commonly used statistical modeling approach to account for the interaction effect is to include the product of the two predictors (known as the interaction term in statistics (Qian, 2016)) in the regression model. For example, in an analysis of Finnish lakes, Malve and Qian (2006) and Qian (2016) 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 (Qian, 2016). 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 (Qian, 2016). 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), the 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. See Qian (2016) for more detailed explanations.

2.2.2. 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):

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

where the subscript ij represents the ith observation from the jth lake. Above the individual lake level, the BHM captures 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 model coefficients $\beta_{0:3j}$ may differ by lake. Statistically, these lakes are regarded as exchangeable with respect to model coefficients 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) combined 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. (2015) 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 (Stow et al., 2009).

2.3. Modeling Road Map

Our analyses consist of two parts:

1. The model represented by 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 as a function of the data set used.
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 to

properly account for the among-lake variation.

All models were fit with log TP and log TN centered at the respective means of log TP and TN concentrations 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 (R Core Team, 2018), using function `lm()` for linear regression models and the function `lmer` from package `lme4` (Bates and Maechler, 2010) for BHM in equations (3) and (4) (Gelman and Hill, 2007). Annotated R code can be found at GitHub (https://github.com/songsqian/simpsons).

3. Results

3.1. 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 coefficient estimated for 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 1,000 lakes. The estimated model coefficients based on NLA2007 and NLA2012 also differ, and the model coefficients based on the combined NLA data are 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 discussed in the captions of Figure 1 is highly unlikely.

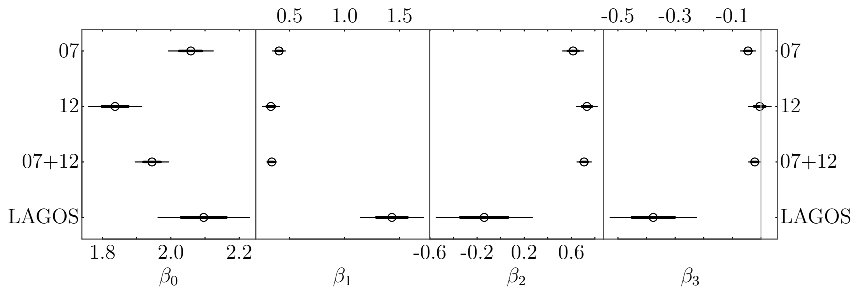

Figure 3: Model coefficients ($\beta_{0:3}$) estimated using lake mean concentrations from NLA2007 (07), NLA2012 (12), NLA2007 and NLA2012 combined (07+12), and the 27 LAGOSNE lakes (LAGOS). Dots are the estimated means and thin and thick horizontal lines are the mean plus one and two standard errors,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

3.2. 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.

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.

4. Discussion

Lakes in both NLA2007 and NLA2012 were selected based on a probabilistic sampling protocol such that analytical results can be "(extrapolated) to national scales" (Pollard et al., 2018). It is tempting to interpret the difference in model coefficients between NLA2007 and NLA2012 (e.g., a decrease in β_0) as a result of improved overall lake condition from 2007 to 2012. Because these coefficients were estimated using lake sample average concentrations of *chla*, TP, and TN, we cannot directly interpret the differences in the models as a 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 (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 model parameters were estimated using statistical methods (Reckhow and Chapra, 1983). These models relied on data from multiple lakes, with each lake or lake segment contributing one observation (Stow and Reckhow, 1996). 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. Our study demonstrates 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." When a dominant confounding factor can be identified, we can incorporate the confounding factor into the BHM (also known as the multilevel model) framework (Tang et al., 2019).

Without properly modeling the hierarchical structure, we risk misinterpreting the data (e.g., Figure 1), a situation that has long been recognized in statistics as the Simpson's paradox (Simpson, 1951). 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.

Developing "national" nutrient criteria using models based on lake average concentrations 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 (2017), the resulting criterion is most likely too stringent, and thereby unachievable, for most lakes. This result is not surprising as 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 (Pollard et al., 2018).

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) (Overton and Stehman, 1996). These programs are purposefully designed to best support a limited number of objectives (Messer et al., 1991). 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.

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 (Efron and Morris, 1977, Efron, 1978). The difficulty in using a Bayesian method is in obtaining proper informative priors. The most important contribution of our paper is the recognition 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 (Qian et al., 2015).

5. Conclusions

- Empirical models developed using lake average concentrations of *chla*, TP, and TN are unlikely coincide with models developed using data from individual lakes – a statistical phenomenon known as the Simpson's paradox in statistics literature and "ecological fallacy" in social science literature.

- Regional differences in relevant natural (e.g., climate, weather, watershed soil) and cultural (e.g., land use) variables are attributed as the cause of the phenomenon. These relevant variables are known as confounding factors in causal analysis literature.

- When using cross-sectional data without detailed information about the confounding factors, a Bayesian hierarchical modeling approach is an appropriate analytic tool.

Acknowledgement

We thank Zutao Ouyang and colleagues at the Center for Global Change and Earth Observations at Michigan State University for feedback when the idea of the project was discussed. Constructive comments and recommendations from the associate editor and two anonymous reviewers are greatly appreciated. This work is partially supported by the Great Lakes Environmental Research Laboratory of the US National Oceanic and Atmospheric Administration (NOAA-GLERL contribution # 1918).

Douglas Bates and Martin Maechler. *lme4: Linear mixed-effects models using S4 classes*, 2010. URL `http://CRAN.R-project.org/package=lme4`. R package version 0.999375-33.

P.J. Bickel, E.A. Hammel, and J.W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404, 1975.

D.E. Canfield. Prediction of chlorophyll a concentrations in Florida lakes: The importance of phosphorus and nitrogen. *Journal of the American Water Resources Association*, 19(2):255–262, 1983.

D.E. Canfield and R.W. Bachmann. Prediction of total phosphorus concentrations, chlorophyll a, and secchi depths in natural and artificial lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, 38(4): 414–423, 1981.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2006.

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

P. J. Dillon and F. H. Rigler. Phosphorus-chlorophyll relationship in lakes. *Limnology and Oceanography*, 19:767–773, 1973.

B. Efron. Controversies in the foundations of statistics. *The American Mathematical Monthly*, 85(4):231–246, 1978.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.

C.T. Filstrup, T. Wagner, P.A. Soranno, E.H. Stanley, C.A. Stow, K.E. Webster, and J.A. Downing. Regional variability among nonlinear chlorophyll—phosphorus relationships in lakes. *Limnology and Oceanography*, 59(5):1691–1703, 2014.

W.A. Fuller. *Measurement Error Models*. Wiley Series in Probability and Statistics. Wiley, New York, 1987.

A. Gelman. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do - Expanded Edition*. Princeton University Press, 2009. ISBN 9781400832118.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

J.R. Jones and R.W. Bachmann. Prediction of phosphorus and chlorophyll levels in lakes. *Journal of Water Pollution Control Federation*, 48(9): 2176–2182, 1976.

Yanzhong Li, Changming Liu, Wenjun Yu, Di Tian, and Peng Bai. Response of streamflow to environmental changes: A Budyko-type analysis based on 144 river basins over China. *Science of the Total Environment*, 664(10):824–833, 2019.

Z. Liang, H. Chen, S. Wu, X. Zhang, Y.H. Yu, and Y. Liu. Exploring dynamics of the chlorophyll a-total phosphorus relationship at the lake-specific scale: a bayesian hierarchical model. *Water, Air, & Soil Pollution*, 229(1):21, 2018.

D.V. Lindley and M.R. Novick. The role of exchangeability in inference. *The Annals of Statistics*, 9(1):45–58, 1981.

O. Malve and S.S. Qian. Estimating nutrients and chlorophyll a relationships in Finnish lakes. *Environmental Science and Technology*, 40 (24):7848–7853, 2006.

E. McCauley, J.A. Downing, and S. Watson. Sigmoid relationships between nutrients and chlorophyll among lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, 46:1171–1175, 1989.

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

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

J Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics*. Wiley, Chichester, UK, 2016.

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

E.E. Prepas and D.O. Trew. Evaluation of the phosphorus–chlorophyll relationship for lakes off the precambrian shield in western canada. *Canadian Journal of Fisheries and Aquatic Sciences*, 40(1):27–35, 1983.

S.S. Qian. *Environmental and Ecological Statistics with R*. Chapman and Hall/CRC Press, 2nd edition, 2016.

S.S. Qian, C.A. Stow, and Y.K. Cha. Implications of Stein's Paradox for environmental standard compliance assessment. *Environmental Science and Technology*, 49(10):5913–5920, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

K.H. Reckhow. A random coefficient model for chlorophyll-nutrient relationships in lakes. *Ecological Modelling*, 70(1):35 – 50, 1993. ISSN 0304-3800.

K.H. Reckhow and S.C Chapra. *Engineering Approaches for Lake Management: Data analysis and empirical modeling*, volume 1. Ann Arbor Science, Butterworth Publishers, 1983.

D.E. Schindler. Evolution of phosphorus limitation in lakes. *Science*, 195: 260–262, 1977.

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

Val H Smith and Joseph Shapiro. Chlorophyll-phosphorus relations in individual lakes. their importance to lake restoration strategies. *Environmental Science and Technology*, 15(4):444–451, 1981.

P.A. Soranno, L.C. Bacon, M. Beauchene, K.E. Bednar, E.G. Bissell, C.K. Boudreau, M.G. Boyer, M.T. Bremigan, S.R. Carpenter, J.W. Carr, K.S. Cheruvelil, S.T Christel, M. Claucherty, S.M. Collins, J.D. Conroy, J.A. Downing, J. Dukett, C.E. Fergus, C.T. Filstrup, C. Funk, M.J. Gonzalez, L.T. Green, C. Gries, J.D. Halfman, S.K. Hamilton, P.C. Hanson, E.N. Henry, E.M. Herron, C. Hockings, J.R. Jackson, K. Jacobson-Hedin, L.L. Janus, W.W. Jones, J.R. Jones, C.M. Keson, K.B.S. King, S.A. Kishbaugh, J.F. Lapierre, B. Lathrop, J.A. Latimore, Y. Lee, N.R. Lottig, J.A. Lynch, L.J. Matthews, W.H. McDowell, K.E.B. Moore, B.P Neff, S.J. Nelson, S.K. Oliver, M.L. Pace, D.C. Pierson, A.C. Poisson, A.I. Pollard, D.M. Post, P.O. Reyes, D.O. Rosenberry, K.M. Roy, L.G. Rudstam, O. Sarnelle, N.J. Schuldt, C.E. Scott, N.K. Skaff, N.J. Smith, N.R. Spinelli, J.J. Stachelek, E.H. Stanley, J.L. Stoddard, S.B. Stopyak, C.A. Stow, J.M. Tallant, P.N. Tan, A.P. Thorpe, M.J. Vanni, T. Wagner, G. Watkins, K.C. Weathers, K.E. Webster, J.D. White, M.K. Wilmes, and S. Yuan. LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of us lakes. *GigaScience*, 6(12), 10 2017. ISSN 2047-217X.

Craig A. Stow and Kenneth H. Reckhow. Estimator bias in a lake phosphorus model with observation error. *Water Resources Research*, 32 (1):165–170, 1996.

Craig A. Stow, E. Conrad Lamon, Song S. Qian, Patricia A. Soranno, and Kenneth H. Reckhow. *Bayesian Hierarchical/Multilevel Models for Inference and Prediction Using Cross-System Lake Data*, pages 111–136. Springer New York, New York, NY, 2009.

Tang, Q. and L. Peng and Y. Yang and S.S. Qian and B.P. Han. Total phosphorus-precipitation and Chlorophyll a-phosphorus relationships of lakes and reservoirs mediated by soil iron at regional scale. *Water Research*, 154: 136–143, 2019.

U.S. EPA. National Lakes Assessment: A Collaborative Survey of the Nation's Lakes. Technical Report EPA 841-R-09-001, U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Washington D.C., 2009.

U.S. EPA. National Lakes Assessment 2012: A Collaborative Survey of Lakes in the United States. Technical Report EPA 841-R-16-113, U.S.

Environmental Protection Agency, Office of Water and Office of Research and Development, Washington D.C., December 2016.

R.A. Vollenweider. Scientific fundations of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorus as factors in eutrophication. Organization for Economic Co-operation and Development, Technical Report DAS/CSI/68.27. 250p, 1968.

R.A. Vollenweider. Input-output models with special reference to phosphorus loading concept in limnology. *Schweizerische Zeitschrift für Hydrologie-Swiss*, 37:53–84, 1975.

T. Wagner, P.A. Soranno, K.E. Webster, and K.S. Cheruvelil. Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes. *Freshwater Biology*, 56:1811–1824, 2011.

Lester L. Yuan and Amina I. Pollard. Using national-scale data to develop nutrient–microcystin relationships that guide management decisions. *Environmental Science and Technology*, 51(12):6972–6980, 2017.

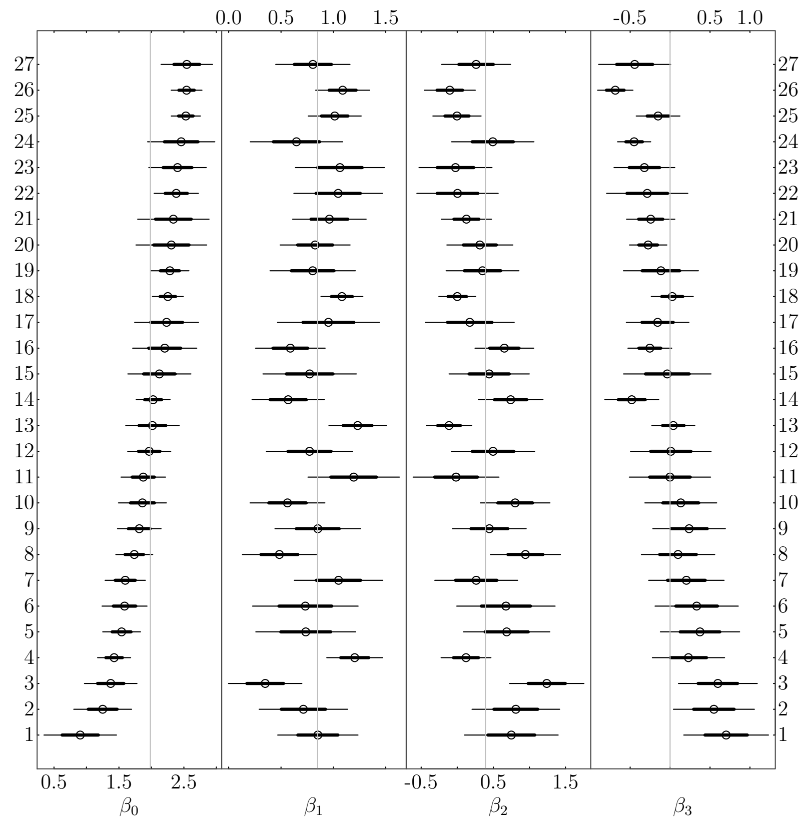

Figure 4: BHM estimated lake-specific model coefficients (β_{0j} – β_{3j}) shown a strong negative correlation between β_{0j} and β_{3j}. Dots are the estimated means and thin and thick horizontal lines are the mean plus one and two standard errors, respectively. The shaded vertical lines for β_0, β_1, 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 $\beta_3 = 0$.