Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales

Tyler Wagner [ID],[1]* Erin M. Schliep[2]
[1]U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, University Park, Pennsylvania
[2]Department of Statistics, University of Missouri, Columbia, Missouri

Abstract

Empirical nutrient models that describe lake nutrient, productivity, and water clarity relationships among lakes play a prominent role in limnology. Landscape-based regressions are also used to understand macroscale variability of lake nutrients, clarity, and productivity (hereafter referred to as nutrient-productivity). Predictions from both models are used to inform eutrophication management globally. To date, these two classes of models are generally conducted separately, which ignores the known dependencies among nutrient-productivity variables. We present a statistical model that integrates nutrient-productivity and landscape-based regressions—where lake nutrients, productivity, and clarity variables are modeled jointly. We fitted a joint nutrient-productivity model to over 7000 lakes with three nutrients (total phosphorus, total nitrogen, nitrate concentrations), chlorophyll *a* concentrations, and Secchi disk depth as response variables and landscape features as predictor variables. Because lakes in different regions respond to landscape features differently, we focused our analysis on two subregions with different dominant land uses, the agricultural Midwest and the forested Northeast U.S. Predictive performance was enhanced by modeling nutrient-productivity variables jointly. We also found strong evidence that nutrient-productivity variables were coupled, and that only nitrate may be decoupled from other nutrient-productivity variables in the forested region. We speculate that these regional differences may be related to differences in the strength of biogeochemical cycles and stoichiometric controls between these regions. Jointly modeling nutrient-productivity variables in lakes effectively integrates the two dominant approaches for studying lakes nutrient-productivity relationships and provides novel insight into macroscale patterns of the coupling of nutrients, chlorophyll, and water clarity in lakes.

The development of empirical models to predict nutrient concentrations, measures of primary producer biomass (e.g., chlorophyll *a* concentrations; CHL) and water clarity (e.g., Secchi disk depth), has a rich history in limnology (Dillon and Rigler 1974; Canfield and Bachmann 1981; Peters 1986). The classic example—a log-linear relationship between CHL and total phosphorus (TP)—is commonly used to inform the development of lake water quality criteria (Havens and Walker 2002; U.S. EPA 2010; Huo et al. 2014). These models used to describe lake nutrient, productivity, and clarity relationships (hereafter referred to as empirical nutrient-productivity models) are integral in informing lake management decisions, testing basic limnological principles, and for advancing our understanding of the controls and drivers of

water quality in lakes (Pace 2001). Nutrient-productivity models have been widely applied to individual lakes and, more recently, to populations of lakes in efforts to improve our understanding of the spatial heterogeneity in the response of lakes to environmental stressors (Malve and Qian 2006; Phillips et al. 2008; Wagner et al. 2011).

Although the development and application of empirical nutrient-productivity models are common, models that include information about lake morphometry and the natural and anthropogenic watershed features of lakes are increasingly being used to identify and understand the importance of landscape drivers of lake nutrients and measures of productivity (Wagner et al. 2011; Nielsen et al. 2012; Read et al. 2015). For instance, landscape-based regressions are used to identify the relative importance of different land use types as non-point sources of nutrients to lakes and for explaining macroscale variation in nutrients, primary

*Correspondence: txw19@psu.edu

producer biomass, and other indicators of water quality (Arbuckle and Downing 2001; Jones et al. 2004; Wagner et al. 2011). These landscape-based regressions are also used to highlight potential management activities that could help meet water quality goals and have been shown to be useful for prediction (Meeuwig and Peters 1996). Because landscape-based regression models use predictors that are largely derived from widely available geospatial datasets, these models are well suited for predicting nutrients and productivity across large spatial extents—at regional to continental scales (Cheruvelil et al. 2013; Collins et al. 2017).

The increase in use of landscape-based regression models is driven by the importance of lake morphometric properties and landscape characteristics to the source, delivery, and processing of nutrients in lakes (Soranno et al. 1996; Carpenter et al. 1998; Collins et al. 2017). Likewise, empirical nutrient-productivity models are widely used because many nutrient-productivity variables are correlated with one another (Ostrofsky and Rigler 1987; Phillips et al. 2008). The correlation among nutrient-productivity variables is partly due to coupled biogeochemical cycles (Schlesinger et al. 2011; Gibson and O'Reilly 2012). For example, nutrients and other elements do not cycle independently, as illustrated by co-limitation of growth of primary producers (Sterner 2008; Harpole et al. 2011). For other variables, such as Secchi disk depth, the correlation with nutrients is because water clarity is measuring, in part, an outcome (i.e., algal biomass) driven by the coupling of biogeochemical cycles. This correlation among nutrient-productivity variables has important implications for limnological modeling and prediction and for furthering our understanding of the ecological processes that influence lake water quality.

From a limnological modeling perspective, combining nutrient-productivity and landscape-based regressions is desirable because both approaches are useful for furthering our understanding of nutrient dynamics and water quality in lakes and making informed predictions for unobserved lakes. In fact, there are important relationships between and among nutrients and landscape variables that may be ignored or not realized by not integrating these two approaches. Efforts to integrate nutrient-productivity and landscape-based regressions have been limited due to challenges with missing data and multicolinearity. Instead, researchers study the different nutrient-productivity variables individually by developing univariate nutrient regressions, where, for example, independent TP, total nitrogen (TN), and CHL regressions are developed and modeled as a function of one or more landscape predictors and then compared (Jones et al. 2004; Carle et al. 2005; Chen et al. 2015; Soranno et al. 2015). Including additional nutrient-productivity variables as predictors into these models can be problematic because there is often missing nutrient-productivity data (Soranno et al. 2017), and these nutrient-productivity variables will likely be confounded with each other and with

other landscape predictors. The missing nutrient-productivity data can be problematic because—although accommodating missing data during analysis is a statistical issue—commonly used statistical software programs for fitting regression models require values of all predictor variables for each lake, either observed or imputed. Missing data is less problematic for landscape-based predictors which are often derived from satellite or remotely sensed data sources. The presence of missing predictor variables is most commonly addressed by using a complete-case analysis; whereby, those lakes with missing data are removed from the analysis (Little 1992; Fergus et al. 2016). Discarding lakes with missing data results in a loss of information.

The second major challenge with integrating these two types of empirical models is the high correlation that often exists among predictors (particularly among nutrient-productivity variables and between nutrient-productivity variables and landscape predictors) which leads to uninterpretable regression coefficients associated with multicollinearity (Doubek et al. 2015). This is important because there is interest in interpreting regression coefficients within the context of understanding drivers of lake nutrients and productivity. In addition, by independently modeling nutrient-productivity variables, we are ignoring the inherent dependence among them, which may be due to coupling, similar environmental drivers, or some combination of both. This dependency affects predictive performance and limits our ecological understanding of how these indicators covary over space and time. Ideally, we could utilize both approaches—nutrient-productivity and landscape-based regressions—that is, jointly model nutrient-productivity variables as a function of landscape-based predictors.

Jointly modeling multiple response variables allows for the integration of nutrient-productivity and landscape-based regressions and overcomes some of the above challenges (Clark et al. 2014; Warton et al. 2015; Schliep et al. 2017). To date, however, jointly modeling multiple nutrient-productivity variables is rarely done (but see Cha et al. 2016 for an example of jointly modeling N and P). These models, which we will refer to as joint nutrient-productivity models, quantify the effects of landscape drivers on nutrient-productivity variables, while at the same time account for nutrient-productivity dependence through the residuals. The joint models enable observed nutrient-productivity variables to inform on the prediction of unobserved variables after accounting for the effects of landscape predictors. For example, by understanding how relationships between nutrient-productivity variables, such as TP and Secchi disk depth—which is widely measured (Lottig et al. 2014)—vary across a population of lakes, we can leverage this information to make more informed predictions of TP at unobserved lakes. Joint nutrient-productivity models can also provide new insight into fundamental limnological relationships by decomposing the correlations among nutrient-productivity
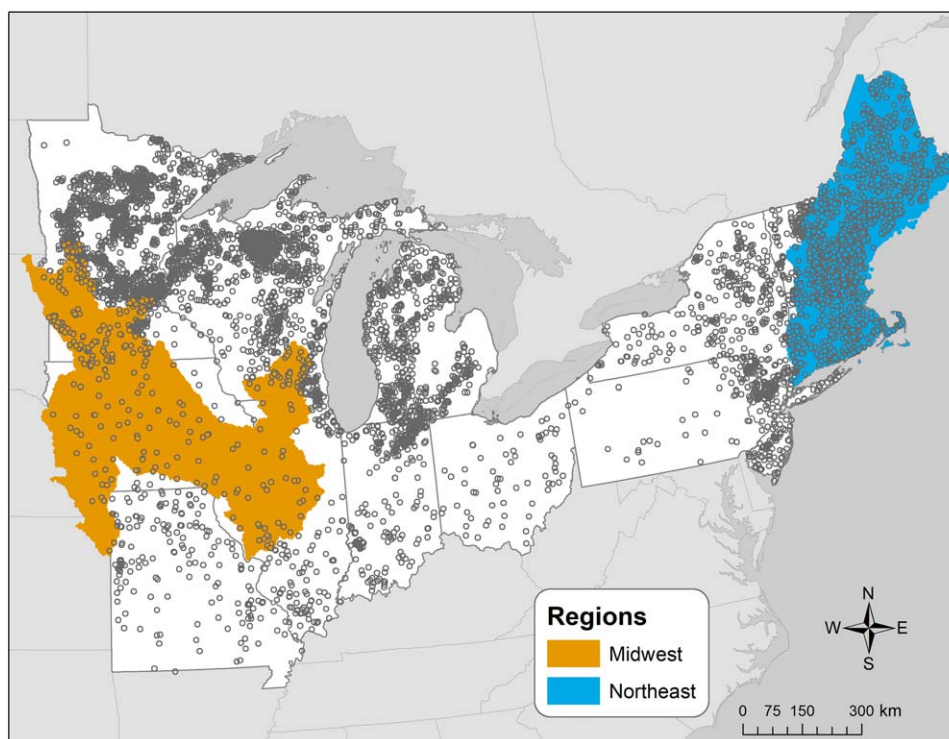
Fig. 1. Map showing study lakes across the LAGOS-NE study extent and the Midwest and Northeast subregions.

variables that are due to shared landscape drivers vs. those that may be due to ecological processes that result in strong ecological coupling.

Our study area contains over 7000 lakes and five nutrient-productivity variables. We compared the predictive performance of the joint nutrient-productivity model to the traditional univariate models where residual dependence is ignored. Previous research has shown that lakes in regions with different dominant land uses and eco-climatic zones respond differently to different landscape drivers. Our study area contains two regions with different dominant land uses—an agriculturally and forested-dominated region. We fitted a nutrient-productivity model for each region to test our expectation that lake nutrient and productivity measures in different regions are likely to be coupled in different ways and are likely to respond to different dominant landscape drivers.

Methods

Nutrient-productivity data

We used lake nutrient-productivity data for 7184 lakes located in the Midwest and Northeastern United States. Nutrient-productivity variables included nutrients—TP (μg L^{-1}), TN (μg L^{-1}), and nitrate concentrations (μg L^{-1}; NO3-N)—and indicators of algal biomass (CHL concentration; μg L^{-1}), and water clarity (Secchi disk depth [m]). All data came from the Lake Multi-Scaled Geospatial and Temporal

Database (LAGOS) of the Northeast U.S. (LAGOS-NE$_{LIMNO}$ v. 1.087.1; Soranno and Cheruvelil 2017a,b; Soranno et al. 2017) using the LAGOS R package (Stachelek and Oliver 2017). LAGOS-NE is a subcontinental scale database that includes approximately 1,800,000 km^2 over a 17-state region in the Midwestern and Northeastern United States. We used a subset of lakes with a surface area $\geq$ 4 ha that had at least one of the five nutrient-productivity variables quantified. Lake nutrient data were restricted to epilimnetic samples taken during the summer months (15 June–15 September) spanning the years 1990–2011. We retained the most recent sampling occasion for every lake. For lakes sample more than once over time (n = 5081 lakes), we used a two-staged approach to ensure that we obtained lakes that were sampled for TN, because TN was not measured as often compared to the other nutrient-productivity variables. First, we selected all lakes that had observed values of TN and used the most recent observation for those lakes. For all other lakes sampled multiple times, we used the most recent observation. This process resulted in a single observation of at least one of the five nutrient-productivity variables per lake (Fig. 1).

Study subregions

We selected two subregions within the LAGOS-NE study area, previously delineated by Collins et al. (2017) in a study that examined drivers of lake nutrient stoichiometry across the LAGOS-NE extent. The regions were created to capture the gradient in forested and agricultural land use present in

the LAGOS study extent and to represent regions dominated by extremes in these two land use/cover types. The focus was on contrasting agricultural and forested landscapes because of the strong relationship between agricultural land use and nutrient inputs into inland lakes (Collins et al. 2017). The two regions (Fig. 1) were created by aggregating regional major river watersheds (HUC 4) based on the proportion of agricultural land use cover in a given HUC 4 where adjacent HUC 4 watersheds with similar land use characteristics were combined. This approach resulted in two contrasting regions—the Northeast region ($n = 1655$ lakes), composed of ten HUC 4 watersheds, that had very low proportions of agricultural land use ($< 10\%$), and the Midwest region ($n = 434$ lakes), which was composed of seven HUC 4 watersheds with relatively high proportions of agricultural land use ($> 50\%$).

Landscape predictor variables

We chose landscape predictor variables that represented important sources of nutrients (e.g., land use) or the transportation of materials to lakes (e.g., stream density), and that are associated with internal processing of nutrients in lakes (e.g., lake depth; Collins et al. 2017). All geospatial lake predictor variable data came from LAGOS$_{GEO}$ v. 1.05 (Soranno and Cheruvelil 2017b). Except for lake maximum depth and lake area, which are lake-scale properties, all geospatial summaries were derived at the lake watershed scale.

Statistical model

We modeled nutrient-productivity variables jointly to account for correlations among variables. Let $\mathbf{Y}(\boldsymbol{s}) = (Y_1(\boldsymbol{s}), \ldots, Y_K(\boldsymbol{s}))\prime$ denote a vector of length K of lake nutrient-productivity variables for lake $\boldsymbol{s}$. The joint nutrient-productivity model can be written as:

$$\mathbf{Y}(\boldsymbol{s}) = \mathrm{BX}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \qquad (1)$$

where B is a matrix of coefficients such that B_{kp} is the coefficient of the pth predictor variable for the kth variable. Additionally, $\epsilon(\boldsymbol{s})$ is an error vector of length K. We model

$$\epsilon(\boldsymbol{s}) \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \Sigma)$$

where Σ is a $K \times K$ covariance matrix capturing the dependence between nutrient-productivity variables that is not accounted for by the regression. These errors are assumed independent and identically distributed across lakes.

Nutrient-productivity variables were modeled on the log$_e$-transformed scale. Because of highly skewed distributions, all proportional predictor variables (e.g., land use) were logit transformed and standardized, while non-proportion predictors were log$_e$-transformed and standardized prior to analysis. The model was fitted to three datasets. The first model used the entire LAGOS-NE study extent. The second and third were fitted to the Midwest and Northeast subregions. All

models were fitted using the program WinBUGS (Lunn et al. 2000) called from within the program R (R Core Team 2017) using the R2WinBUGS package (Sturtz et al. 2005). Independent, diffuse normal priors were used for all coefficient parameters in B, and the variance-covariance matrix, Σ, was modeled using the scaled inverse-Wishart distribution (Gelman and Hill 2007). We ran three parallel Markov chains beginning each chain with random starting values. Each chain was run for 15,000 iterations, from which the first 5000 samples were discarded. This resulted in 30,000 samples used to summarize posterior distributions. Convergence was assessed both visually through the use of trace plots and quantitatively using the Brooks-Gelman-Rubin statistic. Residual plots were examined to assess the assumption of normality. We determined predictor variable significance by evaluating whether or not the 95% credible interval of the coefficient overlapped with zero.

Model performance measures and decomposing correlations

We calculated root mean squared prediction error (RMSPE) using 10-fold cross validation, where the model was fitted 10 times to 90% of the data with 10% retained for out-of-sample prediction. To evaluate the potential predictive power gained by modeling nutrient-productivity variables jointly, we compare marginal predictions (RMSPE$_M$) to conditional predictions (RMSPE$_C$) of the nutrient-productivity variables obtained at out-of-sample locations. Under the multivariate normal distribution assumed in Eq. 1, the marginal predictive distributions are equivalent to the predictive distributions that would result from modeling each nutrient-productivity variable independently. Therefore, marginal predictions are obtained for each nutrient-productivity variable without reference to the values of the other variables. The conditional predictions are obtained for each nutrient-productivity variable by conditioning on the values of all other variables, e.g., we predict TN conditionally at location $\boldsymbol{s}^*$ given its observations of TP, NO3-N, CHL, and Secchi disk depth. Assuming that there is information shared across variables, we would expect to see more accurate and more precise predictions from the conditional predictive distributions than the marginal predictive distributions. If there is very little dependence between the nutrient-productivity variables after accounting for the predictors in the model, the conditional and marginal predictions will be approximately equivalent. As an additional measure of model performance, we calculated both the marginal and conditional predictive R^2, R_M^2, and R_C^2, respectively.

To quantify the potential spatial variability in both landscape drivers (i.e., the estimated regression coefficients) and the residual covariance structure among the response variables on predictive performance, we calculated RMSPE$_M$, RMSPE$_C$, R_M^2, and R_C^2 for the Midwest and Northeast regions using parameter estimates from the model fitted to all lakes

Table 1. Medians, followed by the first and third quartiles, of the landscape and lake-scale predictors and nutrient-productivity response variables for the LAGOS-NE study extent and Midwest and Northeast subregions.

	LAGOS-NE	Midwest	Northeast
Maximum depth (m)	8.6 (4.9, 14.2)	5.8 (3.0, 10.3)	7.6 (4.2, 13.4)
Lake area (ha)	54.5 (21.0, 145.5)	82.1 (30.0, 234)	35.1 (13.9, 108)
Watershed : lake area	8.4 (3.9, 21.6)	8.5 (3.6, 28.7)	10.2 (5.5, 22.9)
Proportion urban land use	0.06 (0.03, 0.12)	0.08 (0.06, 0.16)	0.05 (0.01, 0.09)
Proportion agricultural land use	0.05 (0.0, 0.35)	0.66 (0.40, 0.80)	0.01 (0.0, 0.05)
Proportion wetland land cover	0.07 (0.02, 0.16)	0.03 (0.0, 0.07)	0.04 (0.02, 0.08)
Road density (m ha^{-1})	25.8 (16.3, 40.4)	28.2 (20.1, 44.2)	21.7 (12.0, 35.2)
Stream density (m ha^{-1})	3.7 (0.02, 8.0)	4.0 (0.7, 8.0)	6.5 (2.1, 10.3)
TP (μg L^{-1})	16.0 (10.0, 34.0)	68.2 (33.5, 138)	10.0 (7.0, 16.0)
TN (μg L^{-1})	600 (366, 1000)	1450 (994, 2300)	284 (195, 420)
Chlorophyll *a* (μg L^{-1})	5.1 (2.7, 14.0)	25.9 (9.1, 62.3)	3.9 (2.5, 6.6)
Secchi disk depth (m)	2.4 (1.3, 3.9)	0.9 (0.5, 2.0)	3.9 (2.4, 5.6)
Nitrate (μg L^{-1})	20 (5.0, 50.0)	60 (18.0, 200)	50.0 (20.0, 50.0)

in the LAOGS-NE study extent. If there were differences in the marginal predictions between the region-specific and LAGOS-NE models, then this would suggest the need for spatially varying coefficients. If there were differences in conditional predictions (and marginal predictions were the same), then this may suggest the need for spatially varying covariance structure. If both marginal and conditional predictions were different, this would suggest the need for both spatially varying coefficients and covariances.

Nutrient-productivity variable correlations were decomposed into residual correlations and correlations due to shared environmental drivers. Residual correlations were obtained from the off-diagonal elements of Σ. The correlations due to shared environmental responses were calculated following the methods of Pollock et al. (2014), where $\mathbb{P}_{kk'}$ denotes the correlation between nutrient-productivity variables k and $k\prime$, and is a function of the regression vectors β_{kp} and $\beta_{k\prime p}$ and the covariances of the environmental variables. Strong residual correlation may suggest strong coupling of nutrient-productivity variables or the need for the inclusion of more predictors in the model. A strong correlation due to the environment may suggest similar landscape and lake-scale drivers.

Results

Study lakes

Lakes across the study region varied substantially in their geophysical, chemical, and biological properties, and anthropogenic settings (Table 1). Median values of TP, TN, and NO3-N across all 7184 study lakes were 16.0 μg L^{-1}, 600 μg L^{-1}, and 20 μg L^{-1}, respectively. Median CHL was 5.1 μg L^{-1} and median Secchi disk depth was 2.4 m. Not all lakes had all five nutrient-productivity response variables observed. The proportion of lakes with missing observations was 0.29

($n = 2093$), 0.44 ($n = 3135$), 0.55 ($n = 3983$), 0.32 ($n = 2309$), and 0.18 ($n = 1265$) for TP, TN, NO3-N, CHL, and Secchi disk depth, respectively.

Lakes within the study extent also varied widely in the amount of urban and agricultural land use present in their watersheds, with lake watersheds ranging from 0% to 95% urban and from 0% to 100% agricultural land use. The median percentage of agricultural land use in a lake's watershed in the Midwest subregion was 66% and only 1% for lakes within the Northeast subregion. As expected, the two subregions also differed substantially in lake chemistry and landscape settings (Table 1). For example, median TP for lakes in the Midwest subregion was 68.2 μg L^{-1}; whereas, median TP in the Northeast subregion was 10.0 μg L^{-1}.

Landscape predictors and predictive performance

For all three models (LAGOS-NE extent, Midwest, and Northeast) and across all five nutrient-productivity variables, the predictive R^2 due to landscape predictors alone ranged from 6% (NO3-N in the Northeast subregion)—61% (Secchi disk depth in the Northeast subregion; R_M^2 values; Table 2). On average, predictive R^2 was greatest for Secchi disk depth (average $R_M^2 = 47\%$), followed by TP (average $R_M^2 = 41\%$), TN (average $R_M^2 = 38\%$), CHL (average $R_M^2 = 22\%$), and NO3-N (average $R_M^2 = 11\%$). Some important similarities and differences in predictors of nutrient-productivity variables were detected across indicator and region (Fig. 2). For instance, lake depth was consistently negatively correlated with TP, TN, and CHL and positively correlated with Secchi disk depth. In addition, the proportion of agricultural land use was positively correlated to TP, TN, CHL, and NO3-N and negatively correlated with Secchi disk depth across the LAGOS-NE extent; however, the effect of agricultural land use on nutrient-productivity variables varied by region. The largest differences among the three models were observed in

Table 2. Margnal (*M*) and conditional (*C*) root mean squared predictive error (RMSPE) and predictive R^2 from joint nutrient-productivity models for the Midwest and Northeast subregions. RMSPE and R^2 values with a LAGOS subscript are values for the Midwest and Northeast subregions calculated using the model fitted to the entire LAGOS-NE study extent.

	TP	TN	CHL	Secchi	NO3-N
Midwest					
$RMSPE_M$	0.762	0.552	1.125	0.743	1.689
$RMSPE_C$	0.637	0.428	0.775	0.527	1.404
$RMSPE_{M\,LAGOS}$	0.927	0.658	1.148	0.770	1.804
$RMSPE_{C\,LAGOS}$	0.662	0.437	0.765	0.531	1.498
R^2_M	0.48	0.26	0.13	0.34	0.12
R^2_C	0.64	0.55	0.59	0.67	0.39
$R^2_{M\,LAGOS}$	0.23	−0.06	0.10	0.29	0.00
$R^2_{C\,LAGOS}$	0.61	0.53	0.60	0.66	0.32
Northeast					
$RMSPE_M$	0.640	0.431	0.805	0.391	0.962
$RMSPE_C$	0.588	0.349	0.714	0.332	0.878
$RMSPE_{M\,LAGOS}$	0.722	0.549	0.847	0.504	1.120
$RMSPE_{C\,LAGOS}$	0.614	0.400	0.773	0.397	1.103
R^2_M	0.34	0.44	0.21	0.61	0.06
R^2_C	0.44	0.63	0.38	0.72	0.21
$R^2_{M\,LAGOS}$	0.16	0.09	0.12	0.34	−0.28
$R^2_{C\,LAGOS}$	0.39	0.52	0.27	0.59	−0.24

the agriculturally dominated Midwest subregion. Fewer landscape predictors were important in predicting all five nutrient-productivity variables in the Midwest subregion compared to all the lakes in the study area and the Northeast subregion. The larger uncertainty in parameter estimates in the Midwest subregion primarily reflects the smaller sample size in this region compared to the other two models.

Jointly modeling nutrient-productivity variables and leveraging information about the dependence among nutrient-productivity variables led to substantial gains in predictive performance and, in particular, the precision of estimates. The increased precision of the conditional predictions can be seen when comparing the marginal and conditional posterior predictive distributions in Fig. 3 for two lakes from the Midwest subregion. These lakes were randomly chosen for illustrative purposes. For a given lake, the conditional distributions are obtained by conditioning on whichever other nutrient-productivity variable is/are observed. For lake #1, all five nutrient-productivity variables were observed. Therefore, the conditional distribution for each variable was obtained given the other four variables. For this lake, the highest density of each conditional posterior predictive distribution is closer to the observed value (improved accuracy), and the distributions are narrower (i.e., increased precision) when compared to the marginal posterior predictive distributions (Fig. 3). Similar patterns

are observed for lake #2; however, for this lake, TP and TN were not observed. The conditional posterior predictive distributions for these unobserved quantities have less uncertainty and predict greater concentrations compared to the marginal predictions (Fig. 3). In addition, note that the observed value of Secchi disk depth for this lake is slightly lower than the mean of the predictive distribution given the predictors. Therefore, since there is negative dependence between Secchi disk depth and TP, TN, CHL, and NO3-N, the conditional distributions of TP, TN, CHL, and NO3-N for this lake are shifted right. The shift is least pronounced for NO3-N, which had the lowest residual correlation with Secchi disk depth. Comparisons of the marginal and conditional RMSPE and marginal and conditional predictive R^2 values in Table 2 summarize the gain in predictive performance—where smaller RMSPE and larger predictive R^2 values indicate better predictive performance. For example, in the Midwest study region, RMSPE for TP decreased from 0.762 to 0.637 when predictions were conditional on the observed values of each of the other nutrient-productivity variables. In addition to the decrease in RMSPE, the predictive R^2 increased from 0.48 (R^2_M) to 0.64 (R^2_C; Table 2). Similar gains in predictive performance were observed for all nutrient-productivity variables and across all three regions. Importantly, however, there were differences between the subregions in predictive performance for some nutrient-productivity variables. For instance, the predictive R^2 increased substantially for NO3-N in the Midwest subregion ($R^2_M = 0.12$ vs. $R^2_C = 0.39$); however, a similar gain in predictive R^2 was not observed for NO3-N in the Northeast subregion ($R^2_M = 0.06$ vs. $R^2_C = 0.21$; Table 2).

Residual and shared environmental correlations

Pairwise shared environmental correlations were plotted against residual correlations for all nutrient-productivity variable pairs and models (Fig. 4). The partitioning of the effects of shared environmental drivers from residual interactions revealed which nutrient-productivity variables responded similarly to environmental conditions and which ones may be correlated due to ecological processes not accounted for by the predictor variables. For instance, for lakes in the Midwest subregion, TP, TN, and CHL tended to respond similarly to landscape drivers ($\mathbb{P}_{jj}'$: TP,TN = 0.47, TP,CHL = 0.67, and TN,CHL = 0.32) and were also indicative of variables potentially driven by similar ecological processes (residual correlations: TP,TN = 0.35, TP,CHL = 0.54, TN,CHL = 0.41). Whereas, Secchi disk depth tended to be negatively correlated with shared environmental drivers when compared with TP, TN, and CHL ($\mathbb{P}_{jj}'$: Secchi disk depth,TP = −0.68, Secchi disk depth,CHL = −0.91, and Secchi disk depth,TN = −0.35), and to respond in the opposite direction to shared environmental processes (residual correlations: Secchi disk depth,TP = −0.53, Secchi disk
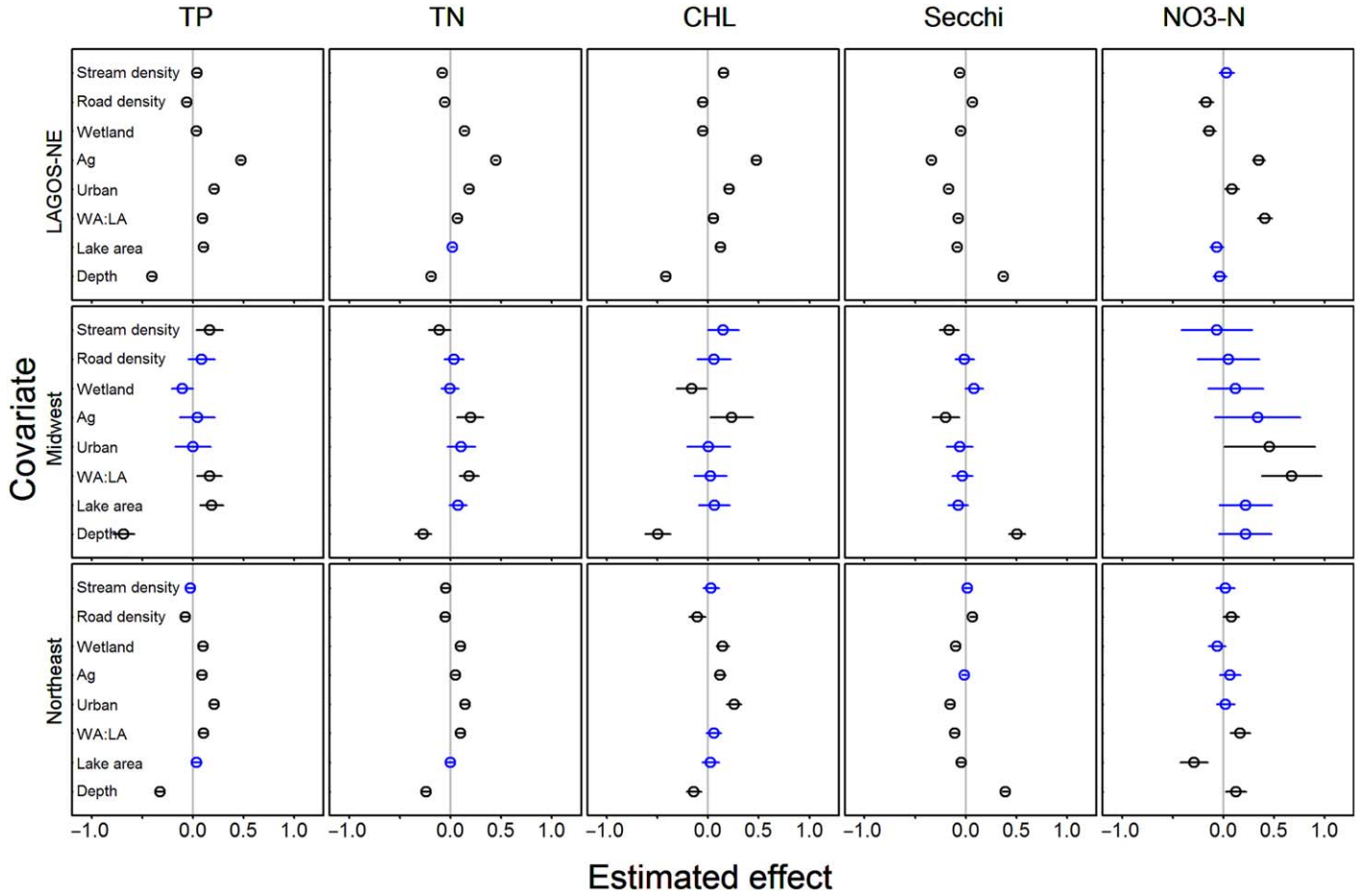
Fig. 2. Estimated effects of landscape and lake-scale predictors on nutrient-productivity variables for the entire LAGOS-NE study extent (row 1), the Midwest subregion (row 2), and the Northeast subregion (row 3). Circles are posterior means and horizontal bars are 95% credible intervals. Effects with 95% credible intervals that overlap with zero are shown in blue. TP, total phosphorus; TN, total nitrogen; CHL, chlorophyll *a*; Secchi, Secchi disk depth; NO3-N, nitrate.

depth,CHL = −0.72, and Secchi disk depth,TN = −0.35). Similar patterns were observed in the other regions for these nutrient-productivity variables, although the magnitude of the correlations varied. One noticeable difference in the correlation partitioning across the two subregions was between NO3-N and the other nutrient-productivity variables. In the Midwest subregion, the correlations due to shared environmental drivers between NO3-N and TP and TN were much larger compared to the Northeast subregion (Midwest: NO3-N,TP = 0.37, NO3-N,TN = 0.76; Northeast: NO3-N,TP = 0.01, NO3-N,TN = 0.06). Residual correlations between NO3-N and TP and TN were also larger in the Midwest subregion compared to those observed in the Northeast subregion, especially for NO3-N and TN (Midwest: NO3-N,TN = 0.49, Northeast: NO3-N,TN = 0.29). In addition, NO3-N shared environmental and residual correlations tended to cluster more closely around zero compared to the Midwest subregion (Fig. 4).

Discussion

Our results demonstrate that jointly modeling nutrient-productivity variables effectively integrates nutrient-productivity and landscape-based regression approaches to understand macroscale drivers of water quality, while simultaneously accounting for dependence among indicators. This approach also easily accommodates missing nutrient-productivity data which allows for the inclusion of lakes into an analysis that otherwise may have been excluded. To date, jointly modeling nutrient-productivity variables is rarely performed. One exception is Cha et al. (2016) who jointly modeled TN and TP to examine N and P limitation in aquatic systems. Their work focused on the spatial and temporal dynamics of N and P limitation, but did not include landscape predictors nor decompose correlation structure. They emphasized the utility of jointly modeling nutrients to further understanding of potential (de)coupling of nutrients across space and time. Our results also highlight the gain in predictive performance that is achieved by
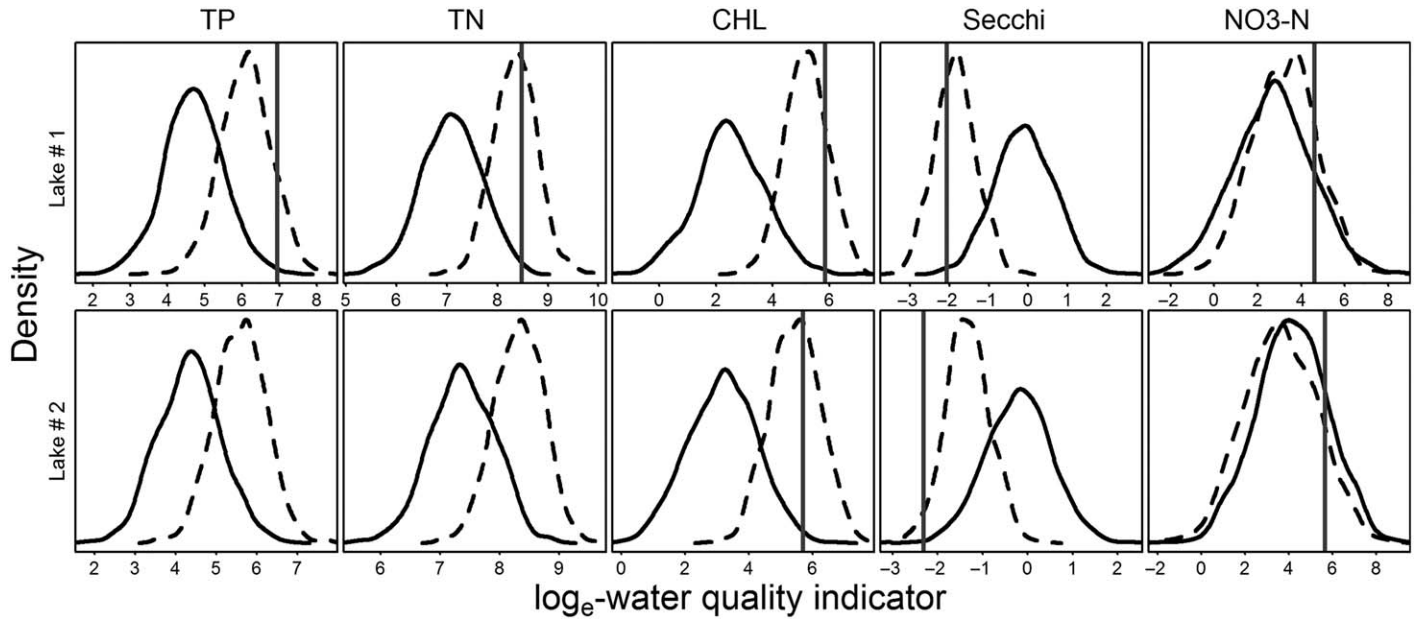
Fig. 3. Marginal (solid black lines) and conditional (dotted lines) posterior predictive distributions from a joint nutrient-productivity model for two lakes in the Midwest subregion. TP, total phosphorus; TN, total nitrogen; CHL, chlorophyll *a*; Secchi, Secchi disk depth; NO3-N, nitrate. Vertical line is observed value. Lake #1 (upper row) had all five nutrient-productivity variables observed; whereas, TP and TN were not observed for lake #2 (bottom row).

jointly modeling nutrient-productivity variables, which is important given the wide-spread use of predictions from limnological models to help inform management decisions (Jones and Bachmann 1976). Specifically, there is a substantial improvement in the precision and accuracy of predictions that is achieved by conditioning on other observed nutrient-productivity variables.

The effects of landscape predictors on nutrient-productivity variables were as expected, and similar to those reported by other studies that examined landscape drivers of lake nutrients across large spatial extents (Wagner et al. 2011; Read et al. 2015; Collins et al. 2017). In addition, we found the effects of landscape predictors to vary spatially (i.e., between subregions), which suggests regional differences in the dominant drivers of, and their effects on, lake nutrients and productivity. Spatially varying effects of landscape predictors on lake nutrients and productivity have been previously identified (Soranno et al. 2014). For example, using spatially varying coefficient models, Fergus et al. (2016) explicitly accounted for spatial heterogeneity in the effects of TP and water color when predicting CHL. Accounting for spatial differences in the effects of predictors on CHL improved model fit and predictive performance compared to models that did not allow predictor effects to vary over space (Fergus et al. 2016). Our results support this notion of the importance of incorporating spatially varying or regionally specific coefficients. For example, predictive performance decreased when using the model fitted to LAGOS-NE to

predict lakes in the Midwest or Northeast subregions compared to using the region-specific models to make predictions.

We observed relatively large, positive residual correlations among TP, TN, and CHL across all analyses. In the case of TP and TN, these positive residual correlations may be the result of coupled biogeochemical cycles and the response of algal communities to increased nutrient loading (Schindler 1978; Cha et al. 2016). Cha et al. (2016) also observed large correlations between N and P in Finnish lakes, which they concluded were indicative of similar rates of N and P biogeochemical cycles. In addition, a relatively large negative residual correlation was observed between Secchi disk depth, TP, TN, and CHL. This was expected, as Secchi disk depth is generally negatively correlated with nutrients and CHL (Canfield and Bachmann 1981). Interestingly, however, there was spatial variability in the residual correlation among NO3-N and other nutrient-productivity variables. For example, the residual correlations of NO3-N with other variables were near zero in the Northeast subregion and positive in the Midwest subregion. These regional differences in NO3-N residual correlations could be the result of differences in the strength of the coupling between the N and carbon biogeochemical cycles—and this coupling may play a larger role in northeastern lakes through stoichiometric controls and microbial processes (Taylor and Townsend 2010). Differences in atmospheric chemistry, nutrient processing dynamics of dominant land cover types, and lake internal processing may also
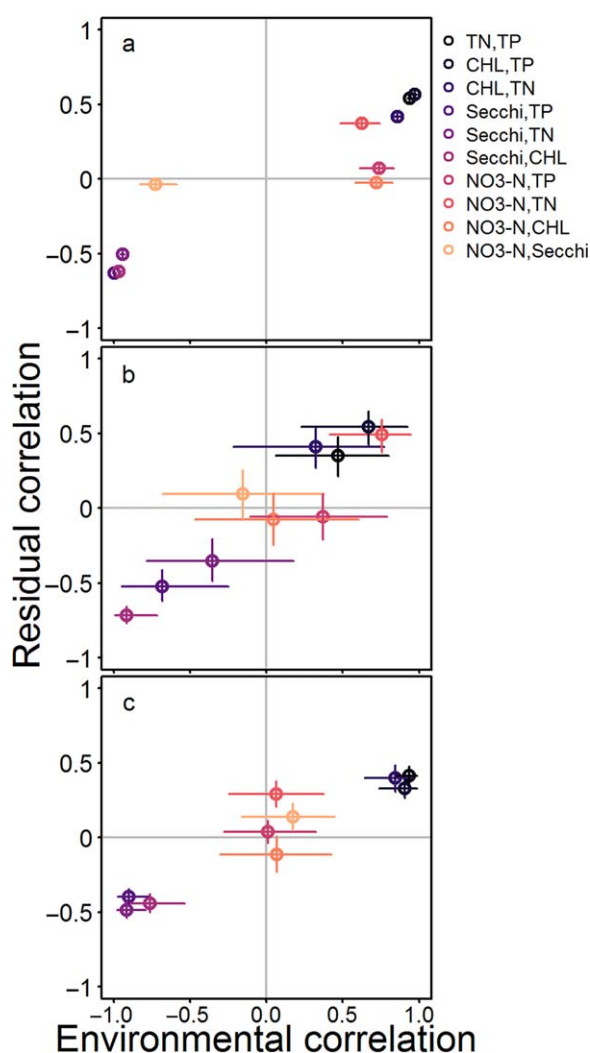
Fig. 4. Estimated residual and environmental correlations between pairs of nutrient-productivity variables for the LAGOS-NE study extent (**a**) and the Midwest (**b**) and Northeast (**c**) subregions. Circles are posterior means and error bars are 95% credible intervals.

contribute to the apparent decoupling of NO3-N with other nutrients in Northeast, and in particular with TN, compared to Midwest lakes (Bernhardt et al. 2002; Goodale et al. 2003). Understanding this potential (de)coupling of nutrients across environmental gradients is important within the context of global change. For example, extreme temperature events may result in the decoupling of some biogeochemical cycles through altering microbial processes (Mooshammer et al. 2017). These results suggest that further investigations into spatially varying covariances among nutrient-productivity variables may be warranted in an effort to understand the implications of spatially varying covariances on model predictive performance and for understanding how (de)coupling may vary spatially and in response to global change.

In addition to relatively large positive residual correlations among TP, TN, and CHL, we also observed strong positive shared environmental correlations among these variables, and strong negative shared environmental correlations among these variables and Secchi disk depth. The positive environmental correlations reflect the positive relationship between nutrient loading and primary production (Schindler 1978). These patterns were consistent across all analyses and highlight the similarity in dominant landscape (e.g., agricultural land use) and lake-level (e.g., lake depth) drivers between our subregions that influence the observed spatial variability of key nutrients and, thus, algal biomass and water clarity. Similar to the spatial variability observed for residual correlations between NO3-N and other nutrient-productivity variables, we also observed spatial variability in the correlations due to shared environmental drivers between NO3-N and other indicators. The regional differences in NO3-N shared environmental correlations may be related to the dominant source of NO3-N to lakes in these two regions. In the Midwest subregion, the dominant source of NO3-N is from agricultural land use practices (Van Metre et al. 2016), which is also a significant source of P. This may result in similar shared environmental correlations among nutrients and biological responses (e.g., CHL concentrations). In contrast, the dominant source of NO3-N in the forested Northeast subregion is from atmospheric deposition (Aber et al. 2003), which may reduce the shared environmental correlations of NO3-N with landscape-derived nutrients (e.g., P). For NO3-N, shared environmental correlations were closer to zero in the Northeast subregion and positive in the Midwest subregion. These patterns also partly reflect the fewer number of landscape-based predictors that were important for predicting nutrients in the Midwest compared to the Northeast subregion—with lake depth, stream density, and agricultural land use playing important roles as a drivers of nutrients and lake productivity in the Midwest. Conversely, a more diverse set of predictors were important in the Northeast subregion, suggesting more than just agricultural inputs and internal lake processing are potentially driving nutrients and productivity in those lakes.

Summary

Understanding the dominant drivers of lake nutrients and productivity, the coupling of biogeochemical cycles, and the use of empirical models to predict water quality is necessary to help guide management and conservation of lake ecosystems. This is particularly the case when examining lakes across macroscales, since some of the primary stressors of freshwater ecosystems operate across large spatial extents. For example, understanding how land use and climate change and the increased human demands on freshwater systems will affect water quality at regional, continental, and global scales is of increasing importance (Woodward et al.

2010). In fact, changes in environmental factors—such as increasing temperatures—may interact with increased nutrient loading to exacerbate the symptoms of eutrophication (Moss et al. 2011). Jointly modeling nutrient-productivity variables provides a useful analytical framework for increasing knowledge of environmental drivers, the coupling of nutrients across space and time, and improving limnological predictions at macroscales.

References

Aber, J. D., C. L. Goodale, S. V. Ollinger, M.-L. Smith, A. H. Magill, M. E. Martin, R. A. Hallett, and J. L. Stoddard. 2003. Is nitrogen deposition altering the nitrogen status of northeastern forests? AIBS Bull. **53**: 375–389. doi: 10.1641/0006–3568(2003)053[0375:INDATN]2.0.CO;2

Arbuckle, K. E., and J. A. Downing. 2001. The influence of watershed land use on lake N: P in a predominantly agricultural landscape. Limnol. Oceanogr. **46**: 970–975. doi: 10.4319/lo.2001.46.4.0970

Bernhardt, E. S., R. O. Hall, Jr., and G. E. Likens. 2002. Whole-system estimates of nitrification and nitrate uptake in streams of the Hubbard Brook Experimental Forest. Ecosystems **5**: 419–430. doi:10.1007/s10021-002-0179-4

Canfield, D. E., and R. W. Bachmann. 1981. Prediction of total phosphorus concentrations, chlorophyll a, and Secchi depths in natural and artificial lakes. Can. J. Fish. Aquat. Sci. **38**: 414–423. doi:10.1139/f81-058

Carle, M. V., P. N. Halpin, and C. A. Stow. 2005. Patterns of watershed urbanization and impacts on water quality. J. Am. Water Resour. Assoc. **41**: 693–708. doi:10.1111/j.1752-1688.2005.tb03764.x

Carpenter, S. R., N. F. Caraco, D. L. Correll, R. W. Howarth, A. N. Sharpley, and V. H. Smith. 1998. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. **8**: 559–568. doi:10.1890/1051-0761(1998)008[0559:NPOSWW]2.0.CO;2

Cha, Y., I. Alameddine, S. S. Qian, and C. A. Stow. 2016. A cross-scale view of N and P limitation using a Bayesian hierarchical model. Limnol. Oceanogr. **61**: 2276–2285. doi:10.1002/lno.10375

Chen, M., G. Zeng, J. Zhang, P. Xu, A. Chen, and L. Lu. 2015. Global landscape of total organic carbon, nitrogen and phosphorus in lake water. Sci. Rep. **5**: 15043. doi: 10.1038/srep15043

Cheruvelil, K., P. Soranno, K. Webster, and M. Bremigan. 2013. Multi-scaled drivers of ecosystem state: Quantifying the importance of the regional spatial scale. Ecol. Appl. **23**: 1603–1618. doi:10.1890/12-1872.1

Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: Forest climate response from joint species distribution models. Ecol. Appl. **24**: 990–999. doi:10.1890/13-1015.1

Collins, S. M., S. K. Oliver, J.-F. Lapierre, E. H. Stanley, J. R. Jones, T. Wagner, and P. A. Soranno. 2017. Lake nutrient stoichiometry is less predictable than nutrient concentrations at regional and sub-continental scales. Ecol. Appl. **27**: 1529–1540. doi:10.1002/eap.1545

Dillon, P., and F. Rigler. 1974. The phosphorus-chlorophyll relationship in lakes. Limnol. Oceanogr. **19**: 767–773. doi: 10.4319/lo.1974.19.5.0767

Doubek, J. P., C. C. Carey, and B. J. Cardinale. 2015. Anthropogenic land use is associated with N-fixing cyanobacterial dominance in lakes across the continental United States. Aquat. Sci. **77**: 681–694. doi:10.1007/s00027-015-0411-x

Fergus, C. E., A. O. Finley, P. A. Soranno, and T. Wagner. 2016. Spatial variation in nutrient and water color effects on lake chlorophyll at macroscales. PLoS One **11**: e0164592. doi:10.1371/journal.pone.0164592

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge Univ. Press.

Gibson, C. A., and C. M. O'Reilly. 2012. Organic matter stoichiometry influences nitrogen and phosphorus uptake in a headwater stream. Freshw. Sci. **31**: 395–407. doi: 10.1899/11-033.1

Goodale, C. L., J. D. Aber, and P. M. Vitousek. 2003. An unexpected nitrate decline in New Hampshire streams. Ecosystems **6**: 0075–0086. doi:10.1007/s10021-002-0219-0

Harpole, W. S., and others. 2011. Nutrient co-limitation of primary producer communities. Ecol. Lett. **14**: 852–862. doi:10.1111/j.1461-0248.2011.01651.x

Havens, K. E., and W. W. Walker. 2002. Development of a total phosphorus concentration goal in the TMDL process for Lake Okeechobee, Florida (USA). Lake Reserv. Manag. **18**: 227–238. doi:10.1080/07438140209354151

Huo, S., C. Ma, B. Xi, Z. Tong, Z. He, J. Su, and F. Wu. 2014. Determining ecoregional numeric nutrient criteria by stressor-response models in Yungui ecoregion lakes, China. Environ. Sci. Pollut. Res. **21**: 8831–8846. doi: 10.1007/s11356-014-2819-6

Jones, J. R., and R. W. Bachmann. 1976. Prediction of phosphorus and chlorophyll levels in lakes. J. Water Pollut. Control Fed. **48**: 2176–2182. http://www.jstor.org/stable/25040000

Jones, J. R., M. F. Knowlton, D. V. Obrecht, and E. A. Cook. 2004. Importance of landscape variables and morphology on nutrients in Missouri reservoirs. Can. J. Fish. Aquat. Sci. **61**: 1503–1512. doi:10.1139/f04-088

Little, R. J. 1992. Regression with missing X's: A review. J. Am. Stat. Assoc. **87**: 1227–1237. doi:10.2307/2290664

Lottig, N. R., T. Wagner, E. N. Henry, K. S. Cheruvelil, K. E. Webster, J. A. Downing, and C. A. Stow. 2014. Long-term citizen-collected data reveal geographical patterns and temporal trends in lake water clarity. PLoS One **9**: e95769. doi:10.1371/journal.pone.0095769

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. Stat. Comput. **10**: 325–337. doi:10.1023/A:1008929526011

Malve, O., and S. S. Qian. 2006. Estimating nutrients and chlorophyll a relationships in Finnish lakes. Environ. Sci. Technol. **40**: 7848–7853. doi:10.1021/es061359b

Meeuwig, J., and R. Peters. 1996. Circumventing phosphorus in lake management: A comparison of chlorophyll a predictions from land-use and phosphorus-loading models. Can. J. Fish. Aquat. Sci. **53**: 1795–1806. doi:10.1139/f96-107

Mooshammer, M., and others. 2017. Decoupling of microbial carbon, nitrogen, and phosphorus cycling in response to extreme temperature events. Sci. Adv. **3**: e1602781. doi:10.1126/sciadv.1602781

Moss, B., and others. 2011. Allied attack: Climate change and eutrophication. Inland Waters **1**: 101–105. doi:10.5268/IW-1.2.359

Nielsen, A., D. Trolle, M. Søndergaard, T. L. Lauridsen, R. Bjerring, J. E. Olesen, and E. Jeppesen. 2012. Watershed land use effects on lake water quality in Denmark. Ecol. Appl. **22**: 1187–1200. doi:10.1890/11-1831.1

Ostrofsky, M., and F. Rigler. 1987. Chlorophyll–phosphorus relationships for subarctic lakes in western Canada. Can. J. Fish. Aquat. Sci. **44**: 775–781. doi:10.1139/f87-094

Pace, M. L. 2001. Prediction and the aquatic sciences. Can. J. Fish. Aquat. Sci. **58**: 63–72. doi:10.1139/f00-151

Peters, R. H. 1986. The role of prediction in limnology. Limnol. Oceanogr. **31**: 1143–1159. doi:10.4319/lo.1986.31.5.1143

Phillips, G., O.-P. Pietiläinen, L. Carvalho, A. Solimini, A. L. Solheim, and A. Cardoso. 2008. Chlorophyll–nutrient relationships of different lake types using a large European dataset. Aquat. Ecol. **42**: 213–226. doi:10.1007/s10452-008-9180-0

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods Ecol. Evol. **5**: 397–406. doi:10.1111/2041-210X.12180

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing; [accessed 2018 May 10]. Available from https://www.R-project.org/

Read, E. K., and others. 2015. The importance of lake-specific characteristics for water quality across the continental United States. Ecol. Appl. **25**: 943–955. doi:10.1890/14-0935.1

Schindler, D. 1978. Factors regulating phytoplankton production and standing crop in the world's freshwaters. Limnol. Oceanogr. **23**: 478–486. doi:10.4319/lo.1978.23.3.0478

Schlesinger, W. H., J. J. Cole, A. C. Finzi, and E. A. Holland. 2011. Introduction to coupled biogeochemical cycles. Front. Ecol. Environ. **9**: 5–8. doi:10.1890/090235

Schliep, E. M., A. E. Gelfand, R. M. Mitchell, M. E. Aiello-Lammens, and J. A. Silander. 2017. Assessing the joint

behaviour of species traits as filtered by environment. Methods Ecol. Evol. **9**: 716–727. doi:10.1111/2041-210X.12901

Soranno, P. A., S. L. Hubler, S. R. Carpenter, and R. C. Lathrop. 1996. Phosphorus loads to surface waters: A simple model to account for spatial pattern of land use. Ecol. Appl. **6**: 865–878. doi:10.2307/2269490

Soranno, P. A., and others. 2014. Cross-scale interactions: Quantifying multi-scaled cause–effect relationships in macrosystems. Front. Ecol. Environ. **12**: 65–73. doi:10.1890/120366

Soranno, P. A., K. S. Cheruvelil, T. Wagner, K. E. Webster, and M. T. Bremigan. 2015. Effects of land use on lake nutrients: The importance of scale, hydrologic connectivity, and region. PLoS One **10**: e0135454. doi:10.1371/journal.pone.0135454

Soranno, P. A., and K. S. Cheruvelil. 2017a. LAGOS-NE-LIMNO v1.087.1: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925–2013. Environmental Data Initiative. Available from http://dx.doi.org/10.6073/pasta/b1b93ccf3354a7471b93eccca484d506

Soranno, P. A., and K. S. Cheruvelil. 2017b. LAGOS-NE-GEO v1.05: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925–2013. Environmental Data Initiative; [accessed 2017 September 26]. Available from https://doi.org/10.6073/pasta/b88943d10c6c5c480d5230c8890b74a8

Soranno, P. A., and others. 2017. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. Giga-Science **6**: 1. doi:10.1093/gigascience/gix101

Stachelek, J., and S. K. Oliver. 2017. LAGOS: R interface to the LAke multi-scaled GeOSpatial & temporal database. Available from https://github.com/cont-limno/LAGOS

Sterner, R. W. 2008. On the phosphorus limitation paradigm for lakes. Int. Rev. Hydrobiol. **93**: 433–445. doi:10.1002/iroh.200811068

Sturtz, S., U. Ligges, and A. Gelman. 2005. R2WinBUGS: A package for running WinBUGS from R. J. Stat. Softw. **12**: 1–16. doi:10.18637/jss.v012.i03

Taylor, P. G., and A. R. Townsend. 2010. Stoichiometric control of organic carbon–nitrate relationships from soils to the sea. Nature **464**: 1178–1181. doi:10.1038/nature08985

U.S. EPA. 2010. Using stressor-response relationships to derive numeric nutrient criteria. Technical report, U.S. EPA.

Van Metre, P. C., J. W. Frey, M. Musgrove, N. Nakagaki, S. Qi, B. J. Mahler, M. E. Wieczorek, and D. T. Button. 2016. High nitrate concentrations in some Midwest United States streams in 2013 after the 2012 drought. J. Environ. Qual. **45**: 1696–1704. doi:10.2134/jeq2015.12.0591

Wagner, T., P. A. Soranno, K. E. Webster, and K. S. Cheruvelil. 2011. Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes. Freshw. Biol. **56**: 1811–1824. doi:10.1111/j.1365-2427.2011.02621.x

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. Hui. 2015. So many variables: Joint modeling in community ecology. Trends Ecol. Evol. **30**: 766–779. doi:10.1016/j.tree.2015.09.007

Woodward, G., D. M. Perkins, and L. E. Brown. 2010. Climate change and freshwater ecosystems: Impacts across multiple levels of organization. Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**: 2093–2106. doi:10.1098/rstb.2010.0055

Conflict of Interest

None declared.