LIMNOLOGY AND OCEANOGRAPHY
Letters

Open Access

Limnology and Oceanography Letters 5, 2020, 228–235
© 2019 The Authors. *Limnology and Oceanography Letters* published by Wiley Periodicals, Inc.
on behalf of Association for the Sciences of Limnology and Oceanography.
doi: 10.1002/lol2.10134

LETTER

Increasing accuracy of lake nutrient predictions in thousands of lakes by leveraging water clarity data

Tyler Wagner,[1]* *Noah R. Lottig*,[2] *Meridith L. Bartley*,[3] *Ephraim M. Hanks*,[3] *Erin M. Schliep*,[4] *Nathan B. Wikle*,[3] *Katelyn B. S. King*,[5] *Ian McCullough*,[5] *Jemma Stachelek*,[5] *Kendra S. Cheruvelil*,[5,6] *Christopher T. Filstrup*,[7] *Jean Francois Lapierre*,[8] *Boyang Liu*,[9] *Patricia A. Soranno*,[5] *Pang-Ning Tan*,[9] *Qi Wang*,[9] *Katherine Webster*,[5] *Jiayu Zhou*[9]

[1]U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, University Park, Pennsylvania; [2]Center for Limnology Trout Lake Station, University of Wisconsin-Madison, Boulder Junction, Wisconsin; [3]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania; [4]Department of Statistics, University of Missouri, Columbia, Missouri; [5]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan; [6]Lyman Briggs College, Michigan State University, East Lansing, Michigan; [7]Natural Resources Research Institute, University of Minnesota Duluth, Duluth, Minnesota; [8]Sciences Biologiques, Universite de Montreal, Montreal, Quebec, Canada; [9]Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan

Scientific Significance Statement

A major challenge facing aquatic scientists is to predict the response of lake eutrophication to changing land use and climate globally. Such predictions require robust, accurate information about nutrient concentrations in thousands of lakes in diverse natural and human-impacted settings. However, generating predictions of nutrient concentrations in all lakes is difficult because of the complex nature of lakes and the challenges both logistically and financially of obtaining in situ monitoring data from a large number of lakes, and these predictions often have high levels of uncertainty. This study demonstrates a new approach for reducing uncertainty in predictions of lake nutrients by conditioning predictions on readily available water clarity data, which improves our ability to predict nutrient concentrations of unmonitored lakes.

Abstract

Aquatic scientists require robust, accurate information about nutrient concentrations and indicators of algal biomass in unsampled lakes in order to understand and predict the effects of global climate and land-use change. Historically, lake and landscape characteristics have been used as predictor variables in regression models to generate nutrient predictions, but often with significant uncertainty. An alternative approach to improve predictions is to leverage the observed relationship between water clarity and nutrients, which is

*Correspondence: txw19@psu.edu

Author Contribution Statement: All authors contributed to the development of the paper. T.W. and N.R.L. led the writing of the manuscript. E.M.S., E.M.H., N.B.W., and M.L.B. performed the statistical analysis. K.B.S.K., I.M., and J.S. performed database queries and summaries. All coauthors edited and contributed to writing. After the coleads, authors are listed in alphabetical order by groups according to level of contribution.

Data Availability Statement: All data are available from the LAGOSNE R package (https://cran.r-project.org/web/packages/LAGOSNE). The code and data required to perform the analysis described in this article are located at https://doi.org/10.5281/zenodo.3484680.

Associate editor: Mark Scheuerell

Additional Supporting Information may be found in the online version of this article.

possible because water clarity is more commonly measured than lake nutrients. We used a joint-nutrient model that conditioned predictions of total phosphorus, nitrogen, and chlorophyll *a* on observed water clarity. Our results demonstrated substantial reductions (8–27%; median = 23%) in prediction error when conditioning on water clarity. These models will provide new opportunities for predicting nutrient concentrations of unsampled lakes across broad spatial scales with reduced uncertainty.

Lake eutrophication is one of the most pressing global issues facing aquatic ecosystems (Smith and Schindler 2009; Schindler 2012). Anthropogenically derived nutrients and the subsequent stimulation of primary production, including potentially harmful algal blooms, have far-reaching ecological and socioeconomic implications (Dodds et al. 2009; Smith and Schindler 2009). Consequently, predicting major nutrients (e.g., total phosphorus [TP] and nitrogen [TN] concentrations) and measures of primary producer biomass (e.g., chlorophyll *a* concentrations [CHL]) for lakes has long been considered a critical component of eutrophication management (Canfield et al. 1984; Ostrofsky and Rigler 1987), with even more urgency within the context of observed and anticipated widespread effects of climate change and land use intensification on the eutrophication of inland waters. For instance, direct and indirect effects of climate change, acting synergistically with increased nutrient loads, may promote the dominance of harmful, bloom-forming cyanobacteria (Elliott 2012; Paerl and Paul 2012). However, because climate and land use change are not uniform across the globe, it is not clear which lakes in which regions and continents will be most eutrophied.

Some of the most important drivers of lake nutrients and productivity (which includes indicators of trophic state, such as CHL and Secchi disk depth), such as land use and climate change, are operating across broad spatial extents, and, as such, it is important to understand the relationships and consequences of those drivers for lakes at large spatial scales. Large-scale assessments of inland lakes have been initiated by the U.S. Environmental Protection Agency National Lake Assessment program (USEPA 2009) and the European Union Water Framework Directive (The EU Water Framework Directive 2014). These programs typically sample a subset of lakes from the entire population of lakes that are of interest because collecting in situ nutrient and productivity measurements (e.g., TP, TN, and CHL concentrations) is logistically and financially difficult to achieve. The resulting sample size of lakes for which in situ measurements are available may be a small proportion of the total population of lakes (e.g., < 1% of the population) from a subset of landscape settings. Many of the subsequent statistical models developed for predicting lake nutrient concentrations and productivity at unsampled locations can be described as univariate landscape-based regressions—where lake morphometric and landscape features that characterize sources and processing of nutrients in lakes are used as predictor variables (Wagner et al. 2011; Collins et al. 2017). In addition, the landscape predictors used in these regressions may not adequately capture the landscape setting of unsampled lakes (i.e., these unsampled lakes may be outside the range of individual predictors or represent novel combinations of predictors). These situations can result in high uncertainty in the prediction of nutrients and productivity in unsampled lakes (i.e., high prediction errors).

Given the challenges of collecting data from a large number of lakes, a potential alternative approach to improve the predictive performance of models for lake nutrients and productivity is to leverage water clarity data (e.g., Secchi disk depth) that are relatively easily obtained and are correlated with nutrients and CHL (Carlson 1977; Wagner and Schliep 2018). Citizen scientists participate in the collection of extremely reliable water clarity information (Canfield et al. 2002; Poisson et al. In press), and "leveraging" these data in joint nutrient-productivity models (JNPMs) is a promising approach to improving predictions of lake nutrient concentrations. Importantly, for the goal of improving predictions, understanding the mechanisms that lead to correlations among water clarity and nutrient concentrations is less important than the existence of a correlation.

JNPMs are models where multiple nutrient-productivity variables are modeled jointly—as opposed to fitting separate univariate regression models, one for each nutrient or productivity variable (Wagner and Schliep 2018). When JNPMs model TP, TN, and CHL jointly with water clarity information, conditional predictions of TP, TN, and CHL—where predictions are obtained by conditioning on the observed water clarity information—can also be obtained. JNPMs also allow for all nutrient-productivity variables to be predicted at unsampled locations.

In this article, our objective was to assess the relative importance of conditioning predictions of nutrients and CHL on water clarity vs. simply increasing the sample size of lakes with in situ measurements of TP, TN, and CHL. We addressed this objective by fitting a Bayesian JNPM model to five data sets that represent different sampling scenarios that varied in the number of lakes sampled with in situ measurements. Sampling scenarios included randomly sampling 1%, 5%, 10%, 25%, or 75% of the population of lakes for model building. For each scenario, we generated out-of-sample (OOS) predictions for TP, TN, and CHL that were conditioned on water clarity and predictions that were generated without conditioning estimates on water clarity values. This process of randomly sampling lakes and making OOS predictions was repeated

10 times for each sampling scenario. Because information is shared between water clarity and TP, TN, and CHL, we would expect conditioned predictions to have less uncertainty compared to predictions of TP, TN, and CHL that were not conditional on water clarity (i.e., marginal predictions). We also expect that models fitted to the larger sample size data sets would perform better than the smaller sample size scenarios because the large sample size presumably captures a broader range of variation in landscape drivers and nutrients.

Methods

Water quality data

We used nutrient, CHL, and water clarity data (measures of water quality) for 8187 lakes located in the Midwest and Northeastern United States (Fig. 1). Data were 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 2017*a*,*b*, Soranno et al. 2017) accessed using the LAGOSNE R package (Stachelek and Oliver 2017). LAGOS-NE is a subcontinental scale database that includes lakes with surface area ≥ 4 ha within an approximately 1,800,000 km^2 extent over a 17-state region in the Midwestern and Northeastern United States (Fig. 1). Water quality variables included nutrients—TP (μg L^{-1}) and TN (μg L^{-1})—and indicators of algal biomass (CHL; μg L^{-1}), and water clarity (Secchi disk depth [m]). Water quality data were restricted to epilimnetic samples taken during the summer stratification months (15 June–15 September) spanning the years 1990–2011. For lakes that were sampled more than once across the entire time period, we retained the sampling observation that had the greatest number of water quality response variables measured, which resulted in a single water quality sample per lake. Lakes across the study extent had a wide range of nutrient concentrations, algal biomass, and water clarity. Median TP, TN, CHL, and water clarity were 16.0 μg L^{-1}, 600 μg L^{-1}, 4.8 μg L^{-1}, and 2.5 m, respectively (Supporting Information Table S1). Not all lakes had simultaneous in situ measurements or all water quality variables, but most lakes had at least water clarity observations. The proportion of the 8187 lakes with missing water quality data was 0.33 (*n* = 2708), 0.53 (*n* = 4308), 0.31 (*n* = 2506), and 0.09 (*n* = 698) for TP, TN, CHL, and water clarity, respectively.

Lake and landscape predictor variables

Eighteen predictor variables were selected 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., maximum lake depth; Read et al. 2015; Collins et al. 2017). All landscape predictor variable data came from LAGOS$_{GEO}$ v. 1.05 (Soranno and Cheruvelil 2017*b*) and have been shown to be important predictors for lake nutrients by past work (e.g., Read et al. 2015). While a few of these predictor variables were highly correlated (Supporting Information Fig. S1), we did not address any issues of collinearity during model fitting as our primary interest was prediction rather than inference on the coefficients (Graham 2003).

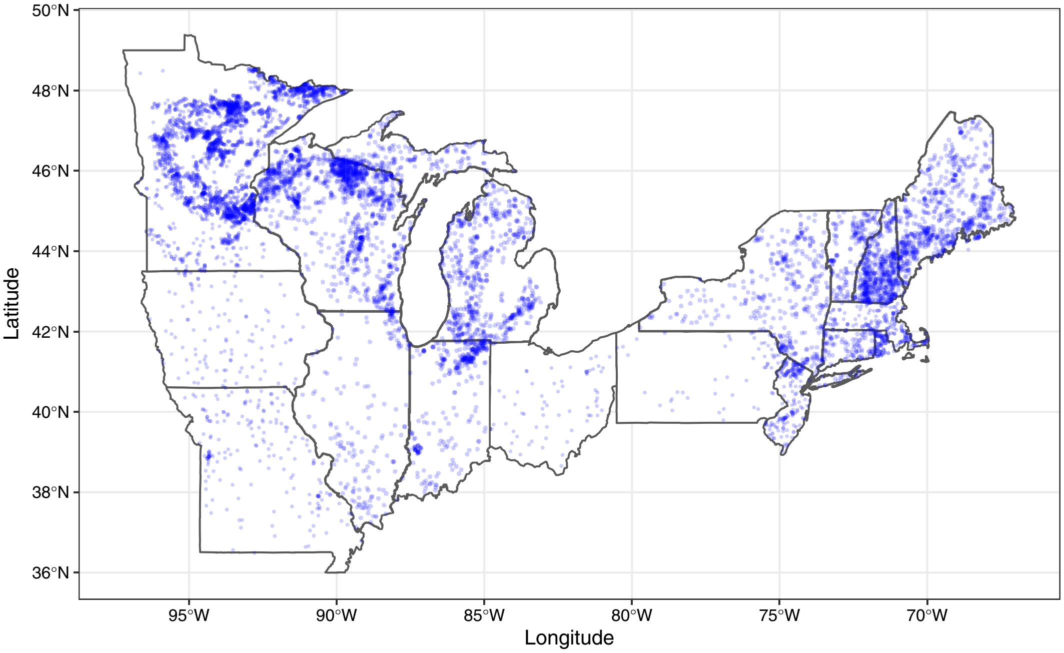

Fig. 1. The population of lakes that were randomly sampled to compare conditional and marginal predictions and assess the importance of sample size using the JNPM.

Statistical model

We fitted a joint distribution model (Wagner and Schliep 2018) to account for correlations among response variables. We begin with the general model notation and then describe the exact specifications used in this analysis. Letting i denote lake where $i = 1, \ldots, n$, the length K vector of lake-nutrient productivity variables is defined $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})'$. Additionally, let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})'$ denote a vector of P predictor variables for lake i. Then, the JNPM is defined

$$\mathbf{Y}_i = B\mathbf{X}_i + \boldsymbol{\varepsilon}_i \qquad (1)$$

where B is a $K \times P$ matrix of coefficients such that B_{kp} is the coefficient of the pth predictor variable for the kth response variable. Additionally, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK})'$ is a vector of length K that defines the residual error for lake i across the K lake-nutrient productivity variables. The residual error vector is modeled

$$\varepsilon_i \sim \mathrm{MVN}(\mathbf{0}, \Sigma) \qquad (2)$$

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

In our analysis, $\mathbf{Y}_i$ is a length 4 vector containing the quantities TP, TN, CHL, and water clarity for lake i. We include 18 predictor variables, including an intercept term, to capture the variation in lake-nutrient productivity variables across the spatial domain. These variables are included in Supporting Information Table S1. Therefore, B is a 4×18 matrix of coefficients allowing for lake-nutrient productivity variable-specific relationships with the predictor variables. Last, Σ is a 4×4 covariance matrix capturing the residual variation in the four lake-nutrient productivity variables after accounting for the predictor variables. The off-diagonal elements of Σ capture the dependence in the residuals across lake nutrient productivity variable at the lake scale.

Water quality variables were modeled on the $\log_e$-transformed scale and all predictor variables were standardized prior to analysis. A constant of 0.1 was added to the response variables prior to log-transformations to accommodate zero values (number of lakes with zero values: TP, $n = 2$; TN, $n = 1$ lake; CHL, $n = 1$ lake; water clarity, $n = 5$ lakes). Adding a constant prior to log-transformation assumes that these values were not true zeros, but that there was some measurement error. We assigned independent Gaussian priors to each element of B and an inverse-Wishart prior to Σ. For each training data set described below, we obtained samples from the posterior distribution of model parameters (B and Σ) using a Markov chain Monte Carlo (MCMC) algorithm coded in R. In addition, missing lake nutrient-productivity variables were treated as random variables and sampled directly from their posterior distributions within the MCMC algorithm. We ran our MCMC sampling algorithm for 2000 iterations, from which the first 200 samples were discarded. This process resulted in 1800 samples used to summarize posterior distributions. We report the posterior mean parameter estimates and corresponding 95% credible intervals and the posterior probability that the estimated coefficient was positive.

Training data sets, conditional predictions, and model performance measures

We created five training data sets to compare the conditional and marginal predictions and assess the importance of sample size using the JNPM. The training data sets represented different sampling scenarios and assumed that we had in situ measurements for either 1%, 5%, 10%, 25%, or 75% of the lakes in our lake population of interest. We used this subset to fit (i.e., train) the JNPM, while the remaining 99%, 95%, 90%, 75%, or 25% of the data set, respectively, was randomly withheld for OOS prediction. The largest sample scenario represented an optimistic scenario where in situ data were available for a large proportion of lakes in the LAGOS-NE study extent. The scenario consisting of having in situ measurements for 25% of lakes within the LAGOS-NE study extent closely mimics a data set such as LAGOS-NE, where the proportion of lakes greater than 4 ha with in situ measurements is closer to 25% of the total population of lakes. The smallest sampling scenario, where 1% of lakes have in situ measurements, is similar to the current sample size of U.S. lakes sampled using a stratified design during the 2007 and 2012 National Lake Assessment.

To evaluate the potential predictive power gained by knowing water clarity for a given lake under each sample size scenario, we compared posterior marginal predictions to posterior conditional predictions of TP, TN, and CHL obtained at OOS locations. Let j denote an OOS lake used for OOS prediction, where $j = 1, \ldots, M$. The marginal predictions of $\mathbf{Y}_j$ were obtained for each OOS lake by sampling from the posterior predictive distribution of all nutrient productivity response variables. These predictions used predictor variables ($\mathbf{X}_i$) for each lake, but ignore any measured values of water clarity at lake j. Under the multivariate model specified in Eq. 1, the marginal predictive distributions are equivalent to the predictive distributions that would result from modeling each water quality variable independently. The conditional predictions are also obtained for each lake nutrient response variable, where we condition on the observed value of water clarity, for example, we predict TP, TN, and CHL conditionally at a lake given its observation of water clarity. These conditional predictions leverage the residual dependence between the lake nutrient response variables and water clarity captured by Σ. Given the multivariate normal model specification in Eq. 1, sampling from the posterior marginal and conditional predictive distributions is straightforward as both require sampling from multivariate normal distributions.

Using samples from both the posterior marginal and conditional predictive distributions, we compute the predictive root mean square error (RMSE) and median percent error (MPE) for each lake nutrient variable. Let $\hat{Y}_{jk}^{(m)}$ and $\hat{Y}_{jk}^{(c)}$ denote the mean estimates of the posterior marginal and conditional predictive distribution for nutrient k at lake j, respectively. The marginal and conditional MSE for nutrient k is computed

$$\mathrm{RMSE}_k^{(m)} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{Y}_{jk}^{(m)} - Y_{jk}\right)^2} \tag{3}$$

and

$$\mathrm{RMSE}_k^{(c)} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{Y}_{jk}^{(c)} - Y_{jk}\right)^2} \tag{4}$$

where the sum is over all M out of sample lakes. Similarly, the marginal and conditional MPE is computed

$$\mathrm{MPE}_k^{(m)} = \underset{j}{\mathrm{median}}\left|\frac{\hat{Y}_{jk}^{(m)} - Y_{jk}}{Y_{jk}}\right| \tag{5}$$

and

$$\mathrm{MPE}_k^{(c)} = \underset{j}{\mathrm{median}}\left|\frac{\hat{Y}_{jk}^{(c)} - Y_{jk}}{Y_{jk}}\right|. \tag{6}$$

Last, we calculated the percent decrease in MPE as ([$\mathrm{MPE_M} - \mathrm{MPE_C}$]/$\mathrm{MPE_M}$ * 100). Thus, the percent decrease in MPE reflects the reduction in MPE after conditioning predictions on observed water clarity. If there is very little dependence between water clarity and TP, TN, and CHL after accounting for the predictors in the model, the conditional and marginal predictions will be approximately equivalent. If conditional predictions of TP, TN, and CHL are better compared to marginal predictions, then we would infer that knowing information about a lake's water clarity is beneficial for making predictions about its nutrients and algal biomass. This process of randomly sampling lakes, fitting the JNPM, and making OOS predictions was repeated 10 times for each sampling scenario. We report the mean and standard deviation (SD) of the marginal and conditional RMSE, MPE, and percent decrease in MPE across the 10 iterations for each sample scenario. We evaluated the assumption of normality by examining quantile-quantile plots and histograms of the residuals. Plots of predicted vs. observed values for marginal and conditional predictions across all response variables and sample scenarios are in Supporting Information Fig. S2.

Results

Several of the predictor variables had credible intervals that did not overlap zero and posterior probabilities of a positive effect indicating that they were important for predicting nutrients and water clarity (Supporting Information Figs. S3, S4). Similar patterns in the relationships between response variables (nutrients and water clarity) and predictor variables were observed across all sampling scenarios (Supporting Information Fig. S3). In general, a greater proportion of predictor variables did not have credible intervals that intersected zero for the large sample scenarios (i.e., sampling 25% and 75% of lakes) than for the smallest sample size scenarios (i.e., sampling 1% and 5% of lakes).

Using water clarity data and leveraging its residual correlation with TP, TN, and CHL (Fig. 2; Supporting Information Fig. S5) resulted in substantial gains in predictive performance at OOS lakes compared to using landscape predictors alone. The improved predictive performance was observed for all three response variables and across all sampling scenarios with the exception of the 1% scenario where the gain in predictive performance was less and variability larger compared to other scenarios (Fig. 3). For example, the mean RMSE decreased for TP from 1.07 (SD = 0.13) to 0.92 (SD = 0.14) when conditioning predictions on water clarity and sampling 1% of lakes, while mean RMSE decreased for TP from 0.79 (SD = 0.02) to 0.65 (SD = 0.03) when conditioning predictions on water clarity and sampling 75% of lakes. MPE also decreased for TP, TN, and CHL when making predictions conditional on the observed water clarity data. For instance, mean MPE for TP under the 1% sample scenario dropped from 0.57 (SD = 0.07) to 0.48 (SD = 0.06), which means that median predictions went from being approximately 57–48% off of true (known) values after conditioning predictions on observed water clarity values—a 15% (SD = 6%) decrease in MPE. For the 75% scenario, mean MPE decreased from 0.43 (SD = 0.01) to 0.32 (SD = 0.007), a 25% decrease in MPE (Fig. 3). Similar to what was observed for RMSE, decreases in MPE were smaller and

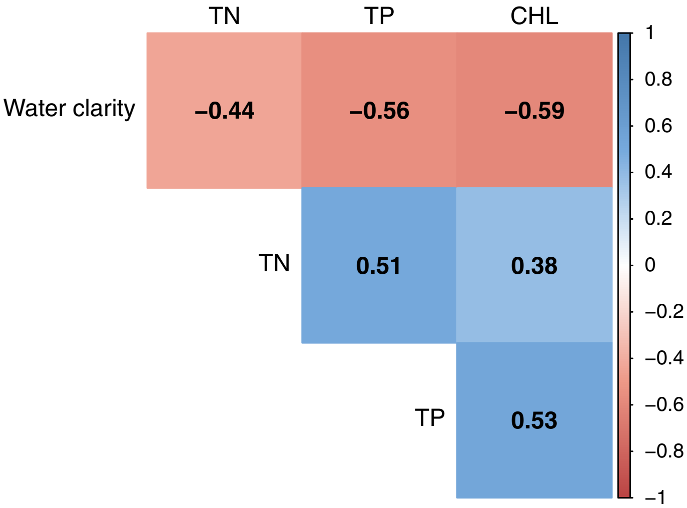

Fig. 2. Estimated residual correlations between pairs of nutrient-productivity variables for the sample scenario where 25% of the lakes had in situ measurements. See Supporting Information Fig. S5 for residual correlations estimated from all sampling scenarios.

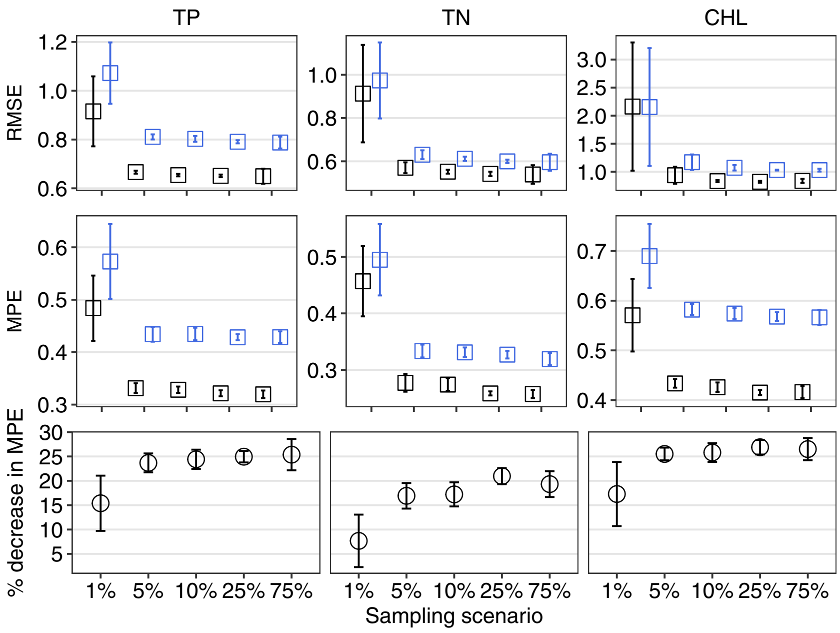

Fig. 3. Marginal (blue squares) and conditional (black squares) RMSE, MPE, and the percent decrease in MPE. Marginal predictions of TP, TN, and CHL concentrations are made without reference to the values of other water quality variables. Conditional predictions are made conditional on the value of observed water clarity (Secchi disk depth) information for a given lake. Within a sample size scenario, percent decrease in MPE reflects the reduction in MPE after conditioning predictions on observed water clarity. Note differences in Y-axis scales for RMSE and MPE.

more variable for the 1% sample size scenario compared to other sample scenarios and TN was more difficult to predict overall compared to TP and CHL. Contrary to expectations, the improvements in predictive performance were not very sensitive to sampling scenario for either conditional or marginal predictions, with improvements only observed when sampling more than 1% of lakes. This suggests that increasing sample size above 5% of the lakes, for either approach, does not greatly improve predictive capacity in OOS lakes. Therefore, conditional modeling of TP, TN, and CHL on water clarity was better able to reduce predictive uncertainty than was greatly increasing sample size.

Discussion

Decision-makers focus on parameters of TP, TN, and CHL when setting lake standards and writing environmental policy. However, water clarity is an easier parameter to measure and has a long history of being measured by volunteers, which results in there being many observations of water clarity in existing databases. Here, we draw on the coupled nature of nutrients in lake ecosystems (Wagner and Schliep 2018) to demonstrate the significance of leveraging water clarity data for predicting nutrient concentrations and indicators of algal biomass in lakes. Our results highlight the substantial benefit of conditioning nutrient and productivity models on water clarity data by reducing the error (8–25% depending on sampling scenario, median = 23% across response variables and scenarios) in nutrient predictions in lakes with water clarity data that are lacking nutrient and productivity data. Additionally, our results were consistent across different sample sizes. Except for the 1% scenario, increasing the training data set sample size did not result in measurable gains in mean prediction error for both marginal and conditional predictions. Conditionally modeling correlations between water clarity and lake nutrients or productivity variables provided larger increases in predictive power than increasing the data size without accounting for this correlation. This result highlights the usefulness of water clarity for prediction and provides a cost-effective path for increasing predictive accuracy of lake nutrients and productivity when these parameters are difficult to obtain.

There are, of course, other frameworks in which the relationship between water clarity and lake nutrients could be modeled. For example, one could construct a univariate regression model with a lake nutrient as the response and water clarity measurements (and other lake characteristics) as predictor variables. Prediction of the nutrient at an unobserved lake using the fitted regression model requires that all predictor variables are observed. In this case, lake nutrient predictions would be limited to only those lakes where water clarity was observed. In our analysis, this approach would result in nearly 700 lakes with no nutrient predictions. In contrast, jointly modeling nutrients and conditioning predictions on water clarity allows for model-based predictions at lakes where water clarity is not observed, with predictions similar to those resulting from our marginal approach. When water clarity data are available, predictions of nutrient levels can take advantage of it. Thus, jointly modeling nutrient and productivity data facilitates nutrient predictions in lakes across large spatial extents, can improve our understanding of the effects of global change on lake ecosystems, and can inform lake water quality management.

Our results have significant implications for predictions of lake nutrients and productivity at regional to continental scales because of the widespread availability of water clarity data. For example, across the LAGOS-NE study region (Fig. 1), 3407, 6122, and 8525 lakes have water clarity information but lack measured TN, TP, and CHL, respectively. Leveraging the water clarity data in the LAGOS-NE data set to estimate nutrient concentrations in these lakes could increase the number of lakes with either measured or predicted (more accurate predictions compared to using only a univariate landscape-based model) nutrient values by 32%, 93%, and 47%, respectively.

Given the prevalence of citizen lake monitoring programs that collect water clarity data (e.g., Poisson et al. 2019) that often span multiple decades for individual lakes, it is likely that similar opportunities exist beyond our focal study region to leverage water clarity information for conditioning nutrient and productivity predictions. Additionally, data sets of satellite-derived water clarity are becoming more prevalent and increasingly being used to assess water quality patterns and trends in lake ecosystems (e.g., Olmanson et al. 2014; Lee et al. 2016; Rose et al. 2017). Consequently, new opportunities exist to not only leverage directly measured water clarity, but also to rely on remotely sensed values to increase the accuracy and availability of nutrient and productivity data in lakes that lack monitoring programs.

When conditioning the predictions on water clarity data, we increased the accuracy of predictive models by 8–27% (median = 23%) for predictions of TN, TP, and CHL. However, there was still a substantial amount of unexplained error in predictions of TP, TN, and CHL (conditional MPE ranged from 0.32 to 0.48, 0.26 to 0.46, 0.42 to 0.57, respectively). Spatial variation is likely one reason for some of the unexplained variability. For example, the JNPM we used in this study did not account for potential spatial dependency between lakes and it assumed that the relationships between predictors and response variables were constant across space. Future model development could address these issues by allowing for spatially varying coefficients, which may further improve model predictions (Fergus et al. 2016).

Ecosystems are inherently complex and characterized by nonlinear relationships and complex interactions acting at multiple different spatial scales (Peters et al. 2007; Evans et al. 2013). However, simple models often are more effective than their more complex counterparts for generating predictions in a variety of situations (Downing et al. 2001; Peters and Herrick 2004). The most significant advantages of the modeling approach presented here for predicting nutrients and productivity in lakes are that the prediction models are simple, easily allow for missing lake nutrient and productivity data, and are based on widely available geographic information system databases across the U.S.

References

Canfield, D. E., J. V. Shireman, D. E. Colle, W. T. Haller, C. E. Watkins II, and M. J. Maceina. 1984. Prediction of chlorophyll a concentrations in Florida lakes: Importance of aquatic macrophytes. Can. J. Fish. Aquat. Sci. **41**: 497–501. doi:10.1139/f84-059

Canfield, D. E., C. D. Brown, R. W. Bachmann, and M. V. Hoyer. 2002. Volunteer lake monitoring: Testing the reliability of data collected by the Florida Lakewatch program. Lake Reserv. Manage. **18**: 1–9. doi:10.1080/07438140209353924

Carlson, R. E. 1977. A trophic state index for lakes. Limnol. Oceanogr. **22**: 361–369. doi:10.4319/lo.1977.22.2.0361

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

Dodds, W. K., W. W. Bouska, J. L. Eitzmann, T. J. Pilger, K. L. Pitts, A. J. Riley, J. T. Schloesser, and D. J. Thornbrugh. 2009. Eutrophication of US freshwaters: Analysis of potential economic damages. Environ. Sci. Technol. **43**: 12–19. doi:10.1021/es801217q

Downing, J. A., S. B. Watson, and E. McCauley. 2001. Predicting cyanobacteria dominance in lakes. Can. J. Fish. Aquat. Sci. **58**: 1905–1908. doi:10.1139/f01-143

Elliott, J. A. 2012. Is the future blue-green? A review of the current model predictions of how climate change could affect pelagic freshwater cyanobacteria. Water Res. **46**: 1364–1371. doi:10.1016/j.watres.2011.12.018

Evans, M. R., and others. 2013. Predictive systems ecology. Proc. R. Soc. B Biol. Sci. **280**: 20131452. doi:10.1098/rspb.2013.1452

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**. doi:10.1371/journal.pone.0164592

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. Ecology **84**: 2809–2815. doi:10.1890/02-3114

Lee, Z., S. Shang, L. Qi, J. Yan, and G. Lin. 2016. A semi-analytical scheme to estimate Secchi-disk depth from Landsat-8 measurements. Remote Sens. Environ. **177**: 101–106. doi:10.1016/j.rse.2016.02.033

Olmanson, L. G., P. L. Brezonik, and M. E. Bauer. 2014. Geospatial and temporal analysis of a 20-year record of landsat-based water clarity in Minnesota's 10,000 lakes.

J. Am. Water Resour. Assoc. **50**: 748–761. doi:10.1111/jawr.12138

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

Paerl, H. W., and V. J. Paul. 2012. Climate change: Links to global expansion of harmful cyanobacteria. Water Res. **46**: 1349–1363. doi:10.1016/j.watres.2011.08.002

Peters, D. P., and J. E. Herrick. 2004. Strategies for ecological extrapolation. Oikos **106**: 627–636. doi:10.1111/j.0030-1299.2004.12869.x

Peters, D. P., B. T. Bestelmeyer, and M. G. Turner. 2007. Cross–scale interactions and changing pattern–process relationships: Consequences for system dynamics. Ecosystems **10**: 790–796. doi:10.1007/s10021-007-9055-6

Poisson, A. C., I. M. McCullough, K. S. Cheruvelil, K. C. Elliott, J. A. Latimore, and P. A. Soranno. 2019. Quantifying the contribution of citizen science to broad-scale ecological databases. Front. Ecol. Environ. doi:10.1002/fee.2128

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

Rose, K. C., S. R. Greb, M. Diebel, and M. G. Turner. 2017. Annual precipitation regulates spatial and temporal drivers of lake water clarity. Ecol. Appl. **27**: 632–643. doi:10.1002/eap.1471

Schindler, D. W. 2012. The dilemma of controlling cultural eutrophication of lakes. Proc. R. Soc. B **279**: 4322–4333. doi:10.1098/rspb.2012.1032

Smith, V. H., and D. W. Schindler. 2009. Eutrophication science: Where do we go from here? Trends Ecol. Evol. **24**: 201–207. doi:10.1016/j.tree.2008.11.009

Soranno, P. A., and K. S. Cheruvelil. 2017*a*. 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; [accessed 2018 May]. Available from https://doi.org/10.6073/pasta/b1b93ccf3354a7471b93eccca484d506

Soranno, P. A., and K. S. Cheruvelil. 2017*b*. 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. Gigascience **6**: 1–22. doi:10.1093/gigascience/gix101

Stachelek, J., and S. K. Oliver. 2017. LAGOS: R interface to the LAke multi-scaled GeOSpatial & temporal database; [accessed 2018 May]. Available from https://github.com/cont-limno/LAGOS

The EU Water Framework Directive. 2014. Available online: doi:10.2779/75229 (Accessed November 2018).

USEPA. 2009. National lakes assessment: A collaborative survey of the nation's lakes. EPA 841-R-09-001. U.S. Environmental Protection Agency, Office of Water and Office of Research and Development. Available from https://www.epa.gov/sites/production/files/2013-11/documents/nla_newlowres_fullrpt.pdf

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

Wagner, T., and E. M. Schliep. 2018. Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales. Limnol. Oceanogr. **63**: 2372–2383. doi:10.1002/lno.10944

Acknowledgments

This research was funded by the National Science Foundation (EF-1638679, EF-1638554, EF-1638539, and EF-1638550). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of Interest

None declared.

Submitted 12 June 2019
Revised 14 October 2019
Accepted 21 November 2019