DR. PATRICIA A. SORANNO (Orcid ID : 0000-0003-1668-9271)
DR. KENDRA SPENCE CHERUVELIL (Orcid ID : 0000-0003-1880-2880)
MRS. KATELYN B.S. KING (Orcid ID : 0000-0001-5471-842X)
DR. IAN M MCCULLOUGH (Orcid ID : 0000-0002-6832-674X)
MX. JOSEPH STACHELEK (Orcid ID : 0000-0002-5924-2464)
DR. MERIDITH BARTLEY (Orcid ID : 0000-0001-5896-2948)
MR. NOAH R LOTTIG (Orcid ID : 0000-0003-1599-8144)
DR. TYLER WAGNER (Orcid ID : 0000-0003-1726-016X)

Article type : Articles

**Journal: Ecological Applications**
**Manuscript type: Article**

**Running head:** Ecological prediction at macroscales

# Ecological prediction at macroscales using big data: Does sampling design matter?

**AUTHORS:** Patricia A. Soranno[1] and Kendra Spence Cheruvelil[1,2,10], Boyang Liu[3], Qi Wang[3], Pang-Ning Tan[3], Jiayu Zhou[3], Katelyn B.S. King[1], Ian M. McCullough[1], Joseph Stachelek[1], Meridith Bartley[4], Christopher T. Filstrup[5], Ephraim M. Hanks[4], Jean-Francois Lapierre[6], Noah R. Lottig[7], Erin M. Schliep[8], Tyler Wagner[9], Katherine E. Webster[1]

[1]Department of Fisheries and Wildlife, Michigan State University, 480 Wilson Road, East Lansing, Michigan 48824, USA

[2]Lyman Briggs College, Michigan State University, 919 East Shaw Lane, East Lansing, Michigan 48825, USA

[3]Department of Computer Science and Engineering, Michigan State University, 428 South Shaw Lane, East Lansing, Michigan 48824, USA

[4]Department of Statistics, The Pennsylvania State University, 324 Thomas Building, University Park,
Pennsylvania 16802, USA

[5]Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, USA

[6]Sciences Biologiques, Universite de Montreal, Pavillon Marie-Victorin, CP 6128, succursale Centre-Ville, Montreal, Quebec, H3C 3J7, CANADA

[7]Center for Limnology Trout Lake Station, University of Wisconsin Madison, Boulder Junction, Wisconsin, 54512, USA

[8]Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, Missouri 65211, USA

[9]U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, Forest Resources Building, University Park, Pennsylvania 16802, USA

[10]Corresponding author. E-mail: ksc@msu.edu

Manuscript received 9 July 2019; revised 13 December 2019; accepted 6 January 2020; final version received 9 February 2020.

**ABSTRACT**

Although ecosystems respond to global change at regional to continental scales (i.e., macroscales), model predictions of ecosystem responses often rely on data from targeted monitoring of a small proportion of sampled ecosystems within a particular geographic area. In this study, we examined how the sampling strategy used to collect data for such models influences predictive performance. We subsampled a large and spatially-extensive dataset to

investigate how macroscale sampling strategy affects prediction of ecosystem characteristics in 6,784 lakes across a 1.8 million km$^2$ area. We estimated model predictive performance for different subsets of the dataset to mimic three common sampling strategies for collecting observations of ecosystem characteristics: random sampling design, stratified random sampling design, and targeted sampling. We found that sampling strategy influenced model predictive performance such that (1) stratified random sampling designs did not improve predictive performance compared to simple random sampling designs and (2) although one of the scenarios that mimicked targeted (non-random) sampling had the poorest performing predictive models, the other targeted sampling scenarios resulted in models with similar predictive performance to that of the random sampling scenarios. Our results suggest that although potential biases in datasets from some forms of targeted sampling may limit predictive performance, compiling existing spatially-extensive datasets can result in models with good predictive performance that may inform a wide range of science questions and policy goals related to global change.

**KEYWORDS:** extrapolation, interpolation, lakes, prediction, sampling design, macroscale, data-intensive ecology, monitoring, sampling, ecological context

## INTRODUCTION

Scientific evidence from focused monitoring efforts has been used since the 1990's to inform environmental policy in response to broad-scale environmental stressors such as acid rain and lake eutrophication (Olsen et al. 1999), and there has been much interest in knowing how different strategies used to select sample ecosystems may affect inference (e.g., Janousek et al. 2019). Previous work has been conducted primarily at local to regional scales, often focusing on geographic areas containing the most sensitive ecosystems. In recent years, there has been a growing recognition of the need to predict ecosystem responses to global change over broader spatial extents that encompass scales from regions to continents (Miller et al. 2004, Dietze et al. 2018, Peters et al. 2018; hereafter referred to as macroscales *sensu* Heffernan et al. 2014). To date, it is unknown how sampling design affects our ability to understand and predict states and relationships in unsampled ecosystems at macroscales.

Prediction at macroscales is complicated because it requires integration of the multi-scaled spatial variation that underlies temporal responses to drivers of global change. Because spatial heterogeneity can be large and can exceed temporal variation (Soranno et al. 2019), it is a critical component to be accounted for when predicting ecosystem states and relationships at the macroscale. Further, prediction accuracy is strongly influenced by the spatial variation of the data used to generate models, which means that the strategy used to select sample ecosystems plays a large role in predictive modeling success (Thompson 2012).

There are two main ways to acquire data for predictive models at the macroscale – coordinated national monitoring programs and compilations of more localized (e.g., local or regional) and disparate datasets. Examples of the first approach include the U.S. Environmental Protection Agency's National Lakes Assessment program that samples approximately 1,000 lakes every five years, comprising ~1 % of lakes $\geq$ 1 ha (U.S. Environmental Protection Agency Office of Wetlands, Oceans and Watersheds Office of Research and Development 2017). Similarly, the U.S.D.A. Forest Service's Forest Health Monitoring Program samples approximately 12,500-25,000 plots annually, comprising 10-20% of all forest plots (Smith 2002). A recent example of the second approach is a macroscale dataset of lake observations created by compiling almost 90 disparate local and regional datasets across 17 U.S. states resulting in approximately 12,000 lakes with at least one observation, comprising 24% of lakes $\geq$ 4 ha (Soranno et al. 2017). In both approaches, a small proportion of ecosystems is sampled and the knowledge gleaned from them is consequently applied to unsampled ecosystems.

Various strategies have been used to select ecosystems for sampling in macroscale monitoring programs in the past, each with their strengths and weaknesses in terms of resources required, potential biases introduced, and predictive power (Urquhart et al. 1998, Olsen et al. 1999, Thompson 2012, Sauer et al. 2013). At the macroscale, sample ecosystems are rarely selected using a simple random design but are sometimes selected using a stratified random design. There has also been a long history of sampling ecosystems for purposes such as ecosystem management without using a probabilistic sampling design that allows representation of the entire population. In these cases, targeted sampling is conducted for subsets of ecosystems or landscapes that are of interest, such as regions that are of high conservation interest or ecosystems that are at high-risk of human perturbation (i.e., observational studies where there is little to no control over which ecosystems are sampled; Thompson 2012). None of these strategies result in a dataset that is a perfectly representative sample of the entire population, particularly when using the sample data for prediction of unsampled ecosystems. In practice, the majority of existing macroscale datasets are likely to be biased in different ways, with some datasets over- or others undersampling particular ecosystems or those with particular characteristics (e.g., Webb et al. 2013, Stanley et al. 2019, Zhao et al. 2019). For example, when multiple disparate datasets are compiled, the resulting datasets include data from a mixture of probabilistic sampling designs and targeted sampling efforts, the effects of which can only be quantified after the database has been created (e.g., GBIF, LAGOS-NE; Gaiji et al. 2013, Soranno et al. 2017).

When building a predictive model, it is a common practice to train the model using a subset of the dataset (training data) and then test the model using data that were withheld (out-of-sample or test data; Lohr 2019). A fundamental assumption behind most predictive empirical models is that the training and test data are generated from the same distributions (i.e., predictions made within the model space). Thus, the resulting predictions are thought of as interpolations. However, if the training and test data are from different distributions, then there is no guarantee that the model fitted to the training data will perform well on the test data (i.e., predictions made outside the model space). For example, predictions at unsampled ecosystems with predictor variables that exceed the range of predictors in the training data and/or comprise a novel combination of predictors may be unreliable and are commonly referred to as extrapolations (Conn et al. 2015). Encountering such novel settings may occur often in macroscale studies due to the broad spatial extent associated with them and the large gradients that exist at these extents for the many characteristics that make up an ecosystem's ecological context (e.g., land use/cover, geology, climate). Therefore, it is critical to assess how various sampling strategies with different purposes may introduce biases that affect distributions of training and test datasets and could change interpolations to extrapolations, thus influencing model predictive performance.

We used a large database compiled from local and regional disparate datasets to ask: what is the effect of sampling design on predictive models of ecosystem characteristics in unsampled ecosystems at the macroscale? We used 4,253-6,784 observations of lake nutrients and productivity from a dataset of 51,101 lakes and their ecological contexts within a spatial extent of 1,778,100 km$^2$ in the northeastern and midwestern U.S. to answer this question (Soranno et al. 2015, 2017). Although this database has its own inherent biases (e.g., undersampling of small lakes; Stanley et al. 2019), it includes a wide range of lake types with large gradients of ecosystem characteristics located across many regions with large gradients of ecological contexts. Therefore, it is an ideal database to create subsets of data that represent known degrees of bias in order to quantify the effects of sampling design on predictive models of ecosystem characteristics.

We developed scenarios (described below) that mimic three common strategies used for collecting observations on ecosystem characteristics at macroscales: random sampling design, stratified random sampling design, and targeted sampling. We used three measures of lake ecosystem characteristics, total phosphorus, total nitrogen, and chlorophyll *a*, to compare the predictive performance of models across these scenarios and strategies. We expected models to have highest predictive performance in cases of assumed interpolation and lowest in cases of assumed extrapolation (Conn et al. 2015). We also expected stratified random designs to increase predictive performance of the interpolation scenarios because the strata are chosen based on underlying ecological processes that are more likely to be related to spatial variation than strictly random sampling. Therefore, we expected predictive performance to be highest when using the stratified random designs, moderate when using the random designs, and lowest when using the targeted sampling. We also expected better predictive performance when using a relatively large proportion of lakes to train or build the predictive model. Finally, we expected nutrients, which are directly linked with landscape context variables, to be better-predicted than lake productivity. Our results will inform the design of macroscale ecosystem assessments, lead to more robust understanding of macroscale variation among ecosystems, and result in better predictions of unsampled ecosystems.

**Conceptualizing the effect of sampling design on predictive models of unsampled ecosystems**

We created seven scenarios that fall within one of the three common sampling strategies employed in macroscale studies, the data from which are used to develop models used to predict at unsampled ecosystems. Figure 1 depicts these strategies as columns with multiple scenarios under each strategy labeled (a-g) and with training data in orange and test data in blue. The scenarios depicted in the left panel of Figure 1 (a-b) illustrate the rare cases when ecosystems are selected at random at the macroscale, called *random sampling design*. Figure 1a depicts the best-case scenario whereby a large proportion of the data are used to train the model and a small proportion of data are used to test the model. We use this scenario as a predictive baseline to compare with the other scenarios that have smaller training datasets since having large datasets to build predictive models is extremely rare in ecology and those that are available often have been compiled from multiple (non-random) sources that are question- or problem-driven. Figure 1b shows the more common scenario in which a smaller dataset is available for model training and the test dataset is larger. If the sample size of the training dataset is sufficiently large, model predictions in these cases are assumed to be within the model space, resulting in *interpolation*.

A second set of scenarios demonstrate stratified random sampling designs that are commonly used in macroscale ecosystem assessments (Figure 1c–d). In these cases, factors that are thought to be important for driving ecosystem processes and patterns are used to first stratify the entire population of ecosystems and then ecosystems are randomly selected within each stratum. Figure 1c-d represents two common cases of *stratified random sampling design.* Figure 1c depicts sample selection using strata based on ecosystem type and Figure 1d depicts selection using strata based on spatial location of the ecosystem, i.e. region. Species presence and/or abundance are commonly estimated with stratified sampling designs whereby the landscape is stratified by ecologically important characteristics (e.g., moose surveys across vegetation types or high/low quality habitat or fish surveys in lakes stratified by depth and area; Ver Hoef 2008, Rask et al. 2010). Stratified random designs assume that the feature(s) used to define strata are ecologically relevant for the response variables being considered by the study (i.e., ecosystem type or regions drive variation among ecosystems). Therefore, as long as the sample size of the training dataset is sufficiently large, predictions for unsampled ecosystems are assumed to be *interpolations*.

The third common sampling strategy is targeted sampling that happens when assessments are question- or problem-driven. In these cases, particular ecosystem types or regions are targeted for sampling in order to answer specific questions or to assess specific populations of ecosystems. Two examples of this design are giant sequoia trees sampled to reconstruct regional fire histories (Swetnam 1993) and lakes in the U.S. sampled as part of the National Eutrophication Survey to study causes of eutrophication (US Environmental Protection Agency 1975). Figure 1e-g depict examples of *targeted sampling* that result in the training datasets being based on particular ecosystem types (Figure 1e), regions (Figure 1f), or regions with particular land uses (Figure 1g). In such cases of targeted sampling, when the data are used in predictive models of unsampled ecosystem types or regions, we assume that these ecosystems are not representative of all ecosystems. Therefore, we assume that these may represent cases of *extrapolation*.

## METHODS

### Study site and dataset

We used the LAGOS-NE database that spans the lake-rich regions of the northeastern and midwestern U.S. (Soranno et al. 2015, 2017) and includes 4,253 – 6,784 lakes depending on the response variable (from a total population of 51,101 lakes ≥ 4 ha). The study lakes include both shallow and deep lakes (interquartile range of maximum depth = 4.6-13.7 m), natural lakes and reservoirs, and lakes with watersheds that are entirely forested to entirely surrounded by agricultural land use. The lakes in this database cover broad gradients in climate, geology, land use/cover, hydrology, and topography. LAGOS-NE-$_{GEO}$ v1.05 includes lake, local, and regional ecological context (Soranno and Cheruvelil 2017a) and LAGOS-NE-$_{LIMNO}$ v1.087.1 includes at least one *in situ* observations of lake water quality for 10,173 lakes (Soranno and Cheruvelil 2017b). These lakes are nested within 65 regions defined by the level 4 hydrologic units regionalization (Seaber et al. 1987; hereafter referred to as regions and HU4s; Figure 1). These regions with an average area of 43,500 km$^2$, have been shown to account for regional variation in nutrients and productivity of this lake population (Cheruvelil et al. 2013). Data and code are available (see Data Availability).

### Lake response variables

We analyzed three ecosystem characteristics of lakes that represent major nutrients and primary productivity – total phosphorus (TP), total nitrogen (TN), and chlorophyll *a* (CHL). These variables are routinely measured by a wide range of academic, governmental, and non-governmental programs to assess water quality (Poisson et al. 2019). We selected lakes and observations using the following criteria. Lake nutrient and productivity observations were selected during the time of peak production in these lakes (i.e., the summer stratified period of 15 June through 15 September) during the years 1980 to 2011. For lakes with multiple observations within a summer or across multiple years, we selected a single sample that contained the most response variables. The resulting data came from lakes ranging from very nutrient-poor and low productivity systems to very nutrient-rich and high productivity systems that are distributed across our study area (Table 1; Figure 1).

### Local and regional ecological context predictor variables

We selected 18 predictor variables *a priori* that are consistently related to lake nutrients and productivity (Table 2; Read et al. 2015, Collins et al. 2017, Lapierre et al. 2018, Soranno et al. 2019). At the local scale, we included six lake-specific characteristics – lake connectivity type (defined as either lakes that have either no stream connections or only outflowing stream connections (Isolated), lakes with inflowing and outflowing stream connections (DR_Stream), or lakes with connections to upstream lakes (DR_LakeStream); lake water clarity (as measured by Secchi disk depth); maximum lake depth; lake complexity (a metric of lake shape that measures the deviation of the shoreline from a circular shape); and, lake elevation. Lake water clarity was included because it is available for nearly all lakes in our study sample (Figure 2) and model predictions are more accurate when they are conditional on water clarity (Wagner et al. 2020, Wagner and Schliep 2018). We also included five watershed-specific characteristics for the area of land draining directly into the lake as well as the area that drains into upstream-connected streams and lakes <10 ha (i.e., the inter-lake watershed; Soranno et al. 2017) – watershed wetland cover; watershed complexity (a metric of watershed shape that measures the deviation of the watershed boundary from a circular shape); watershed to lake area ratio; watershed stream density; watershed forest cover; and, watershed road density. Finally, seven regional-scale characteristics calculated for each HU4 were included in models – mean percent baseflow (an index of regional groundwater contribution); mean runoff; percent agricultural land use; mean annual temperature; mean annual precipitation; mean total nitrogen deposition in 1990; and, the difference in mean total nitrogen deposition from 1990 to 2010. Details on the data sources for these variables are provided in Soranno et al. (2017).

**Macroscale sampling scenarios**

We created seven sampling scenarios that mimic common approaches used for collecting observations on ecosystem characteristics at the macroscale (Figure 1a-g). In these scenarios, we assumed that the population of LAGOS-NE lakes with TP (n = 5,896), TN (n = 4,253), or CHL (n = 6,784) represent the census population (but see Stanley et al. 2019), and that the training and test data were subsets of this population. We fitted models to each of these seven scenario datasets and compared predictive performance (see below for details) for modeling the state of 'unsampled' ecosystems in the test dataset.

*Random sampling designs*: In these two scenarios, we used a random sampling design and examined the effect of sample size on predictive model performance (Figure 1a-b). First, we created a scenario that represents an analytical predictive baseline with a large training dataset of 75% of sampled lakes (Figure 1a). As a contrast, we created a scenario that uses a small training dataset of 25% of sampled lakes (Figure 1b).

*Stratified random sampling designs:* In these two scenarios, we stratified the sampling based on ecological context measured at either the local scale (based on lake type) or the regional scale (based on the region membership of each lake) (Figure 1c-d). For the lake type strata, we created four clusters of lakes based on watershed and regional landscape context characteristics (Table 2) and using hierarchical clustering using Ward's method (Ward 1963). Cluster 1 was characterized by lakes in regions with above average number of and extent of upstream lakes. For the remaining three clusters that had below average regional upstream lake connectivity, lakes were characterized by either high stream density in the watershed (cluster 3), high percent of wetlands in the watershed and around the lake perimeter (cluster 4,) or by both low stream density and low wetland percent in the watershed (cluster 2). For both stratified random scenarios, we selected 25% of lakes within each strata (lake type or region) to build the predictive models, and then predicted the values for the remaining 75% of lake ecosystems as we did for the random sampling design scenario that had a small training dataset (described above).

*Targeted sampling:* We created three targeted sampling scenarios by selecting lakes of particular types, particular regions, or particular types of regions. First, using the above four lake type clusters, we selected all lakes in two of the four clusters to form the training dataset, and tested the model on the lakes in the remaining two of the four clusters. Second, we selected all lakes in half of the regions to form the training dataset and tested the model on lakes in the remaining half of the regions. For these two targeted scenarios, we split the sampled lake data approximately 50:50 and randomly selected the lake type clusters or regions for training the models. For the third targeted scenario, we deliberately selected half of the regions with the lowest proportion of agricultural land to form the training dataset and tested the model on lakes in the remaining regions with the highest proportion of agricultural land. However, because lakes are not distributed equally across regions, the number of lakes was not 50:50 in the training:test datasets for this scenario. The high-agriculture regions contain only 25% of the sampled lakes in the study area, whereas the low-agricultural regions are very lake-rich and contain 75% of the sampled lakes in the study area.

**Predictive models of ecosystem characteristics**

We used random forest models (Breiman 2001, Liaw and Wiener 2002) to predict each of the three response variables (TP, TN, CHL) based on the 18 local (lake-specific and watershed) and regional predictor variables described above that are related to lake nutrients and productivity in the LAGOS-NE lakes (Table 1, 2). Random forest is an ensemble learning method that generates its prediction by averaging the outputs produced by a set of regression trees; and, each regression tree is created via bootstrapping by applying sampling with replacement on the training data (Breiman 2001, Zhou 2012). Although there are no distributional assumptions for random forests, the algorithm determines the best model based on squared error between predictions and true, out of sample, data (Breiman 2001).

We log-transformed the response data after adding 0.1 to the values to down-weight errors on lakes with large data values so that our error terms are closer to percent error than to absolute error. For predictor variables, there were a few cases of missing values (1.97% of values). Those values were imputed with the mean value for that variable so that all observations could be used in the random forest models. The predictor variables were standardized by subtracting the mean and dividing by the standard deviation.

After these pre-processing steps, the dataset was split into training and test datasets based on the seven scenarios depicted and described above (Figure 1). To minimize the likelihood of chance selection affecting modeling results, we randomly split the dataset into training and testing datasets 10 times for each scenario possible (four of the seven scenarios). For the two scenarios that used four lake types, we could only create six sets of training and testing datasets (all possible combinations of four types). For the targeted sampling scenario that used regional land use, only one train/test dataset could be created.

We then used the random forest method, building 189 total independent models, one for each combination of response variable (3), sampling design scenario (7), and train/test dataset combination (1, 6, or 10 as described above). Random forest has several hyperparameters that need to be tuned with the training data, including maximum tree depth and number of trees. We conducted a grid search, with both of these hyperparameters allowed to range from 50 to 200. We performed 5-fold cross validation (Stone 1974) on the training data to determine the optimal hyperparameter setting. Specifically, we iteratively reserved four of the five folds for model building and used the remaining fifth fold as a validation set to select the best hyperparameters. We then re-trained the random forest model on the entire training set using the best hyperparameter values and applied the resulting model to the test dataset to predict the response variables. We trained our random forest model using the Python scikit-learn RandomForest package with Gini impurity as the splitting criterion of the tree (Pedregosa et al. 2011).

*Predictive performance:* We quantified model predictive performance three ways for each of the 189 independent models to compare the effect of sampling scenarios on model performance. First, we calculated the root mean squared error (RMSE), which is a measure of average prediction error that is in the units of the log-transformed response variable. Second, we calculated the median relative absolute error (MRAE; $\varepsilon = median(\frac{|\hat{y} - y|}{y})$), which is a unitless measure of relative error that can be useful for comparing model performance across response variables. Third, we calculated the predictive $R^2$, which is a bounded measure of model relative accuracy whereby 0 indicates that model prediction is no better than using the mean value of the response variable and 1 indicates perfectly accurate model prediction. For the five scenarios with multiple train/test dataset combinations, we calculated average predictive performance and corresponding standard error over the multiple train/test dataset combinations.

## RESULTS

The predictive models accounted for 34-63% of the variation in lake nutrients and productivity across the seven scenarios that mimic three common macroscale sampling strategies (Figure 3A). In general, $R^2$ values decreased from larger to smaller training datasets, from random sampling design (stratified or not) to targeted sampling, and from TP to TN and to CHL. $R^2$ was > 0.5 for all of the response variables and sampling scenarios, except for when modeling TN using the regional land use targeted sampling scenario (Figure 3A (g)). The predictor variables that accounted for most of the variation in responses were lake and watershed landscape characteristics such as lake maximum depth, watershed percent forest, and water clarity (Appendix S1: Table S1-S3).

The two random sampling design scenarios are (a) and (b) in Figure 3. The scenario that used 75% of lakes to build the model (a) resulted in predictions of nutrients and productivity with the lowest error as measured by RMSE and MRAE (Figure 3B-C). Although we considered this scenario to be somewhat unrealistic in practice due to the large sample size (e.g., n = 4,422 for TP), when we decreased that sample size to 25% of sampled lakes (e.g., n = 1,474 for TP; (b)), the effect on predictive performance was negligible (change in RMSE of 0.02-0.03; Figure 3B, Table 3).

The two stratified random sampling design scenarios are (c) and (d) in Figure 3. When comparing the simple random sampling design scenarios with the smaller training dataset (b) to these two stratified random sampling designs, we found small differences in predictive performance (Figure 3, Table 3). In fact, the differences among these three random sampling design scenarios (stratified or not) were nonexistent to negligible. Therefore, the three assumed interpolation scenarios with smaller training datasets (b – d) were similarly able to predict lake nutrients and productivity.

The three scenarios that represent targeted sampling are (e) - (g) in Figure 3. Targeted sampling based on lake type (e) or region (f), resulted in slightly lower predictive performance and higher variation across simulated datasets compared to the random sampling design scenario that uses the smaller training dataset (b) (Figure 3). However, the scenario that mimicked targeted sampling of regions with high agriculture (g) resulted in the poorest performance of any scenario, particularly for TN. This poor performance is likely due to this scenario being a case of extrapolation as demonstrated by differences in the distributions of the response variables and important predictor variables between the training and testing datasets for (g) that was not apparent for the other scenarios (Figure 4).

## DISCUSSION

We studied 6,784 lakes across a spatial extent of 1.8 million $km^2$ to understand how different sampling strategies may affect model predictions of commonly measured ecosystem characteristics in unsampled ecosystems at macroscales. We found that although the sampling strategy used is likely to influence model predictive performance, the differences may not always be as large or as expected based solely on sample sizes and whether the strategy results in interpolation or extrapolation. We have two specific take home messages from this research. First, sampling designs based on two commonly used stratified random approaches (i.e., by region or by ecosystem type) did not result in better predictions of lake nutrients and productivity compared to a simple random sampling design, suggesting that at the macroscale, stratified random sampling designs may not *always* be better than simple random sampling designs. Second, models trained with data from targeted sampling were not always the poorest performing models. However, the predictive performance varied across the three targeted sampling scenarios and three response variables. This fact suggests that data from some targeted sampling may result in extrapolation and poor model performance, and thus should be examined for potential biases before use. Below, we discuss the effects of sampling strategies on predictive model performance and interpret these effects within the context of LAGOS-NE, the database that was used to create the seven sampling scenarios. Then, we discuss the implications of our results for optimizing macroscale sampling designs.

**Effects of sampling strategies on predictive performance**

We anticipated that random and stratified random sampling designs would outperform targeted sampling. This expectation was based mainly on the assumption that targeted sampling designs would result in the training and test data having different distributions, meaning predictions would be made outside of the model space (i.e., extrapolation). However, our results demonstrated that this assumption does not always hold true. For example, the distributions of the training and testing datasets for the response and predictor variables were very similar for the Targeted-Type scenario (Figure 4). Recent work on identifying when predictions will be extrapolation or interpolation suggests that this can be done by either examining distributions of predictor variables or comparing predictive variance at out-of-sample locations to a threshold (e.g. maximum predictive variance) based on in-sample locations (Bartley et al. In Press, Conn et al. 2015). As our example shows, not all targeted sampling designs will result in extrapolation and it may be acceptable to include data from such targeted efforts in larger, compiled datasets.

We also anticipated that stratified random sampling would result in better predictive performance than random sampling. There is intuitive appeal to stratified random sampling designs, particularly given the large amounts of ecological variation that exist at the regional scale (e.g., Cheruvelil et al. 2013, Lapierre et al. 2018). However, we did not find this to be the case, perhaps because LAGOS-NE includes a large sample size of 6,784 lakes (Table 3) that are relatively evenly distributed such that they capture the large geographic gradients that are present across the study area. It is also important to recognize that stratified random sampling design is better than random sampling design only if the strata used are ecologically relevant. For example, some sampling designs stratify by ecosystem size or area (e.g., U.S. EPA 2017). However, we did not include lake area as a stratum because it is not related to lake characteristics in LAGOS-NE (Stanley et al. 2019). Although we did not find stratified random designs to improve predictions over simple random designs, there may be other ways to stratify lakes that we did not consider here; further, there may be other ecosystem types, locations, or uses of macroscale monitoring data that require stratification.

Finally, we expected that lake nutrients would be better predicted than productivity because nutrients are more directly related to landscape context characteristics (Wagner and Schliep 2018) and exhibit a stronger spatial structure than CHL in our study area (Lapierre et al. 2018). This expectation was supported by lower $R^2$s and higher RMSEs for CHL than for the nutrients. In fact, the errors may be enough to suggest an alternate trophic state (e.g., the predicted value could be beyond a trophic threshold between mesotrophic and eutrophic). These results suggest that there is no one best sampling design for all response variables and that multiple metrics should be used when evaluating model predictive performance. The different diagnostic metrics also suggest that there are some subtle differences depending on which metric is used, and caution should be made in interpreting the results when selecting a sampling design based on one model performance metric alone.

Our conclusions should be interpreted within the context of the data used to conduct the research, specifically regarding the type of database, the study area, and the sample sizes used. This research was conducted using a compiled database of 87 disparate lake water quality datasets, many of which were sampled by individual U.S. state agencies (LAGOS-NE; Soranno et al. 2015, 2017). Consequently, sample lakes in LAGOS-NE were selected using a variety of different sampling strategies. In particular, sampled lakes tend to be larger and more connected than all lakes in the study area (Stanley et al. 2019). Therefore, lakes with *in situ* measurements in LAGOS-NE may not completely represent all lakes within the study area and the mimicked random sampling designs are not truly random (i.e., random selection from ~4,000 to 8,000 lakes with lake nutrients and productivity data rather than random selection from ~51,000 lakes in the census population). However, we believe that the sampled lakes in LAGOS-NE can provide a good approximation of all lakes in this geographic extent for three important reasons: (1) Because LAGOS-NE contains sampled lakes that vary widely by lake type, region, and ecological contexts, it contains sufficient variation in predictors and responses to effectively build predictive models; (2) a prior resampling exercise that corrected for the surface area sampling bias that we know is present in LAGOS-NE did not substantially change the statistical distributions of total nutrients and productivity (Stanley et al. 2019); and, (3) the combined sample sizes are likely large enough that any existing biases due to individual program sampling designs would have only minor effect on model performance.

**Implications for macroscale sampling designs**

Macroscale monitoring programs often use either a stratified random design or targeted sampling. Our results from LAGOS-NE, which includes a variety of sampling designs, suggest that predictions from targeted sampling designs may sometimes perform similarly to those from random sampling designs. Thus, there is potential to include these datasets that were created to answer particular questions or to address specific environmental problems in compiled datasets because the bias associated with these data have only minimal effect on prediction errors. This fact will be especially true when data from targeted sampling designs make up a small proportion of the total compiled ecosystem data, resulting in differences between the distributions of training and test datasets (i.e., extrapolation). Moreover, because it is unlikely that a large number of datasets will have exactly the same sample biases, assembling multiple datasets should tend to minimize the impact of any one dataset collected for one particular reason on prediction. Therefore, the use of such secondary datasets compiled from multiple sources, as was done for LAGOS-NE, is useful for macroscale prediction of ecosystem characteristics. Based on our results using lake nutrients and productivity, we make two specific suggestions for optimizing sampling designs at macroscales.

*To stratify or not?* Our results suggest that it may not be necessary to stratify when a relatively large sample size is feasible and relevant strata for prediction are either not present or unknown. Macroscale monitoring programs generally sample less than 20% of ecosystems, and sometimes as little as 1%. In comparison, LAGOS-NE includes 8 to 13% of all lakes ≥ 4 ha, depending on the response variable. And, for a stratified random design to be effective, the stratification must account for some variation in the ecosystem characteristics of interest such that resulting predictions are interpolations (predictions within model space) as opposed to extrapolation (predictions outside of model space). We tested two commonly used approaches for stratification that have been shown to capture variation in lake nutrients and productivity - regions (e.g., Cheruvelil et al. 2013) and local lake and watershed characteristics (e.g., Collins et al. 2018) - but were unable to document substantial improvements in model performance over simple random designs. This fact is likely because the relatively large number of lakes spread across wide environmental gradients in LAGOS-NE resulted in similar distributions of training and testing data (Figure 4) such that strata were not necessary to effectively capture variation in predictors and responses. Therefore, predictive performance was not substantially improved by adding stratification to a simple random sampling design.

Further, it is not likely that a single stratification design would adequately capture the complexity of all ecosystem characteristics, particularly when one considers biological, physical, and chemical characteristics in diverse ecosystems. Because the key characteristics that are most beneficial to use as strata will vary by response variable, it may be more effective to increase the total sample size across the study area rather than to spread samples across strata. Such relatively large and distributed sampling should help to increase predictive performance. The relative performance of simple random versus stratified random designs warrants testing in other settings, for other macroscale datasets, and for other ecosystem characteristics to test the generality of our results. For example, the need for stratification may become more important as landscapes become more heterogeneous or vary across strata and as sample sizes drastically decrease, resulting in sampled ecosystems being less likely to represent a large proportion of the total landscapes or ecosystems within a study area.

*Space or time?* Our study examined the macroscale spatial predictions of lake nutrients and productivity by leveraging the broad spatial gradients in the LAGOS-NE database. In fact, an analysis of LAGOS-NE data using annual time scales across several decades found that spatial variation of lake nutrients and productivity far exceeded temporal variation (Soranno et al. 2019). However, if the goal is to predict responses of all ecosystems across regions and continents to a range of global change stressors, then making predictions across both space and time is essential (Janousek et al. 2019). Unfortunately, there are few spatially- and temporally-extensive datasets and the costs of long-term monitoring through both time and space are untenable. Thus, for new macroscale sampling programs, we recommend first capturing the existing spatial variation in predictor and response variables by sampling across the full range of ecological contexts present across a study area. Then, once sufficient spatial variation is captured, resources could be directed towards a smaller number of systems that are repeatedly sampled to capture temporal variation. By combining the use of secondary datasets that have excellent spatial coverage across a range of ecological context settings with sampling designs focused on filling in gaps in the temporal domain, macroscale studies will be able to inform a wide range of science questions and policy goals related to forecasting the effects of global change on ecosystem characteristics.

**ACKNOWLEDGMENTS**

Author contributions are as follows. PAS and KSC are co-lead authors and contributed equally to the manuscript by leading the conceptualization and writing of the manuscript. After the co-leads, there are 4 groups of authors in decreasing level of contribution, with authors listed in alphabetical order within each group. 1) QW and BL performed the analysis, with 2) P-NT and JZ as supervisors. 3) KBSK, IMM, and JS performed database queries, summaries, created tables and figures, and the code repository. 4) The remaining authors, in addition to those in groups 1-3, contributed to the development, editing, and writing of the paper. The authors declare that they have no conflict of interest. Further, we wish to thank Autumn Poisson and all participants from the 2018 Continental Limnology Project Workshop at Pennsylvania State University, including Emily Stanley, Nicole Smith, Nathan Wikle, Sarah Collins, and Claire Boudreau. Thanks to Meredith and Justin Holgerson for their feedback and providing information about the FIA program, to Allie Shoffner for her editorial suggestions, and to the anonymous reviewers whose suggestions improved this manuscript. Funding was provided by the US NSF Macrosystems Biology Program grants, DEB-1638679; DEB-1638550, DEB-1638539, DEB-1638554. PAS was also supported by USDA National Institute of Food and Agriculture Hatch Project, Grant Number: 176820. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.



## LITERATURE CITED

Bartley, M. L., E. M. Hanks, E. M. Schliep, P. A. Soranno, and T. Wagner. In Press. Identifying and characterizing extrapolation in multivariate response data. PLoS ONE.

Breiman, L. 2001. Random Forests. Machine Learning 45:5–32.

Cheruvelil, K. S., P. A. Soranno, K. E. Webster, and M. T. Bremigan. 2013. Multi-scaled drivers of ecosystem state: quantifying the importance of the regional spatial scale. Ecological Applications 23:1603–1618.

Collins, S. M., S. K. Oliver, J. 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. Ecological Applications 27:1529–1540.

Conn, P. B., D. S. Johnson, and P. L. Boveng. 2015. On Extrapolating Past the Range of Observed Data When Making Statistical Predictions in Ecology. PLoS ONE 10:e0141416.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, H. W. Loescher, C. K. Lunch, B. C. Pijanowski, J. T. Randerson, E. K. Read, A. T. Tredennick, R. Vargas, K. C. Weathers, and E. P. White. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences 115:1424–1432.

Gaiji, S., V. Chavan, A. H. Ariño, J. Otegui, D. Hobern, R. Sood, and E. Robles. 2013. Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. Biodiversity Informatics 8.

Heffernan, J. B., P. A. Soranno, M. J. Angilletta, L. B. Buckley, D. S. Gruner, T. H. Keitt, J. R. Kellner, J. S. Kominoski, A. V. Rocha, J. Xiao, T. K. Harms, S. J. Goring, L. E. Koenig, W. H. McDowell, H. Powell, A. D. Richardson, C. A. Stow, R. Vargas, and K. C. Weathers. 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. Frontiers in Ecology and the Environment 12:5–14.

Janousek, W. M., B. A. Hahn, and V. J. Dreitz. 2019. Disentangling monitoring programs: design, analysis, and application considerations. Ecological Applications 0:e01922.

Lapierre, J.-F., S. M. Collins, D. A. Seekell, K. S. Cheruvelil, P.-N. Tan, N. K. Skaff, Z. E. Taranu, C. E. Fergus, and P. A. Soranno. 2018. Similarity in spatial structure constrains ecosystem relationships: Building a macroscale understanding of lakes. Global Ecology and Biogeography 27:1251–1263.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:5.

Lohr, S. L. 2019. Sampling: Design and Analysis. 2 edition. Routledge. New York, NY.

Miller, J. R., M. G. Turner, E. A. H. Smithwick, C. L. Dent, and E. H. Stanley. 2004. Spatial extrapolation: The science of predicting ecological patterns and processes. BioScience 54:310–320.

Olsen, A. R., J. Sedransk, D. Edwards, C. A. Gotway, W. Liggett, S. Rathbun, K. H. Reckhow, and L. J. Young. 1999. Statistical Issues for Monitoring Ecological and Natural Resources in the United States. Environmental Monitoring and Assessment 54:1–45.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:2825–2830.

Peters, D. P. C., N. D. Burruss, L. L. Rodriguez, D. S. McVey, E. H. Elias, A. M. Pelzel-McCluskey, J. D. Derner, T. S. Schrader, J. Yao, S. J. Pauszek, J. Lombard, S. R. Archer, B. T. Bestelmeyer, D. M. Browning, C. W. Brungard, J. L. Hatfield, N. P. Hanan, J. E. Herrick, G. S. Okin, O. E. Sala, H. Savoy, and E. R. Vivoni. 2018. An integrated view of complex landscapes: A big data-model integration approach to transdisciplinary science. BioScience 68:653–669.

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. Frontiers in Ecology and the Environment. https://doi.org/10.1002/fee.2128

Rask, M., M. Olin, and J. Ruuhijärvi. 2010. Fish-based assessment of ecological status of Finnish lakes loaded by diffuse nutrient pollution from agriculture. Fisheries Management and Ecology 17:126–133.

Read, E. K., V. P. Patil, S. K. Oliver, A. L. Hetherington, J. A. Brentrup, J. A. Zwart, K. M. Winters, J. R. Corman, E. R. Nodine, R. I. Woolway, H. A. Dugan, A. Jaimes, A. B. Santoso, G. S. Hong, L. A. Winslow, P. C. Hanson, and K. C. Weathers. 2015. The importance of lake-specific characteristics for water quality across the continental United States. Ecological Applications 25:943–955.

Sauer, J. R., W. A. Link, J. E. Fallon, K. L. Pardieck, and D. J. Ziolkowski. 2013. The North American Breeding Bird Survey 1966–2011: Summary analysis and species accounts. North American Fauna:1–32.

Seaber, P. R., F. P. Kapinos, and G. L. Knapp. 1987. Hydrologic unit maps. USGS Numbered Series, U.S. G.P.O.,.

Soranno, P. A., 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. 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.

Soranno, P. A., E. G. Bissell, K. S. Cheruvelil, S. T. Christel, S. M. Collins, C. E. Fergus, C. T. Filstrup, J.-F. Lapierre, N. R. Lottig, S. K. Oliver, C. E. Scott, N. J. Smith, S. Stopyak, S. Yuan, M. T. Bremigan, J. A. Downing, C. Gries, E. N. Henry, N. K. Skaff, E. H. Stanley, C. A. Stow, P.-N. Tan, T. Wagner, and K. E. Webster. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience 4:28.

Soranno, P. A., and K. S. Cheruvelil. 2017a. 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.

Soranno, P. A., and K. S. Cheruvelil. 2017b. 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. DOI: 10.6073/pasta/b1b93ccf3354a7471b93eccca484d506.

Soranno, P. A., T. Wagner, S. M. Collins, J.-F. Lapierre, N. R. Lottig, and S. K. Oliver. 2019. Spatial and temporal variation of ecosystem properties at macroscales. Ecology Letters 22:1587–1598.

Stanley, E. H., S. M. Collins, N. R. Lottig, S. K. Oliver, K. E. Webster, K. S. Cheruvelil, and P. A. Soranno. 2019. Biases in lake water quality sampling and implications for macroscale research. Limnology and Oceanography 64:1572–1585.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) 36:111–147.

Swetnam, T. W. 1993. Fire history and climate change in giant Sequoia groves. Science 262:885–889.

Thompson, S. K. 2012. Sampling. 3 edition. Wiley, Hoboken, N.J.

Urquhart, N. S., S. G. Paulsen, and D. P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8:246–257.

US Environmental Protection Agency. 1975. National Eutrophication Survey Methods 1973–1976 (Working Paper No. 175), Tech. rep. United States Environmental Protection Agency, Office of Research and Development, Corvallis, OR, USA.

U.S. Environmental Protection Agency Office of Wetlands, Oceans and Watersheds Office of Research and Development. 2017. National Lakes Assessment 2012: Technical Report. U.S. Environmental Protection Agency, Washington DC.

Ver Hoef, J. M. 2008. Spatial methods for plot-based sampling of wildlife populations. Environmental and Ecological Statistics 15:3–13.

Wagner, T., N. R. Lottig, E. M. Schliep, J. J. Stachelek, K. S. Cheruvelil, and S. M. Collins. 2020. Increasing accuracy of nutrient predictions in thousands of lakes by leveraging water clarity data by citizen scientists. Limnology and Oceanography Letters. https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lol2.10134

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. Limnology and Oceanography 63:2372–2383.

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58:236–244.

Webb, E. L., D. A. Friess, K. W. Krauss, D. R. Cahoon, G. R. Guntenspergen, and J. Phelps. 2013. A global standard for monitoring coastal wetland vulnerability to accelerated sea-level rise. Nature Climate Change 3:458–465.

Zhao, S., N. Pederson, L. D’Orangeville, J. HilleRisLambers, E. Boose, C. Penone, B. Bauer, Y. Jiang, and R. D. Manzanedo. 2019. The International Tree-Ring Data Bank (ITRDB) revisited: Data availability and global ecological representativity. Journal of Biogeography 46:355–368.

Zhou, Z.-H. 2012. Ensemble methods: Foundations and algorithms. Chapman & Hall Inc, Boca Raton, FL.

## DATA AVAILABILITY

Data and code are available in Zenodo at: http://doi.org/10.5281/zenodo.3606832

## TABLES

**Table 1:** Summary of response variables (minimum, maximum, median, mean, 25th and 75th percentiles.

| Response Variable | Units | n | Min | 25th | Median | Mean | 75th | Max |
|---|---|---|---|---|---|---|---|---|
| Total Phosphorus | µg/l | 5896 | 0 | 10 | 16 | 39.9 | 34 | 1184 |
| Total Nitrogen | µg/l | 4253 | 0 | 380 | 600 | 944.3 | 990 | 20574 |
| Chlorophyll-a | µg/l | 6784 | 0 | 2.6 | 5 | 16.28 | 13 | 553.4 |

**Table 2:** Summary of the predictor variables (minimum, maximum, median, mean, 25$^{th}$ and 75$^{th}$ percentiles) at the local (lake and watershed) and regional scales. Note, lake connectivity is a categorical variable with 3 categories (Isolated, DR_Stream, and DR_LakeStream). Lake and watershed complexity refer to lake and watershed boundary complexity factor, respectively, which are measures of reticulation, N dep refers to nitrogen deposition in 1990, and N dep difference refers to N deposition in1990 minus that in 2010.

| **Predictor Variable** | **Units** | **Min** | **25th** | **Median** | **Mean** | **75th** | **Max** |
|---|---|---|---|---|---|---|---|
| **Local** | | | | | | | |
| lake connectivity[1] | NA | NA | NA | NA | NA | NA | NA |
| lake water clarity[1] | m | 0 | 1.30 | 2.40 | 2.75 | 3.80 | 18.25 |
| lake max depth[1] | m | 0.30 | 4.60 | 8.53 | 10.84 | 14.02 | 198.4 |
| lake complexity[1] | NA | 1.00 | 1.40 | 1.75 | 2.11 | 2.35 | 30.27 |
| lake elevation[2] | m | 0 | 241.1 | 323.9 | 316.5 | 412.1 | 1038.6 |
| watershed wetland[3] | % | 0.00 | 2.42 | 7.23 | 12.28 | 17.77 | 93.08 |
| watershed complexity[1] | NA | 1.21 | 2.02 | 2.37 | 2.59 | 2.85 | 25.49 |
| watershed lake ratio[1] | NA | 0.01 | 3.88 | 8.31 | 42.63 | 21.03 | 53517.4 |
| watershed stream density[1] | m/ha | 0 | 0 | 3.08 | 4.54 | 7.57 | 71.77 |
| watershed forest[3] | % | 0 | 23.70 | 53.8 | 49.91 | 75.05 | 100 |
| watershed road density[4] | m/ha | 0 | 14.50 | 24.35 | 30.96 | 39.36 | 262.66 |
| **Regional** | | | | | | | |
| baseflow mean[5] | % | 14.18 | 47.92 | 52.62 | 52.08 | 58.44 | 78.83 |
| runoff mean[5] | in/year | 2.80 | 7.26 | 10.65 | 13.21 | 22.59 | 26.95 |
| agriculture[3] | % | 1.79 | 5.67 | 26.33 | 28.64 | 34.07 | 78.66 |
| temperature mean[6] | $^{o}$C | 3.46 | 5.44 | 6.15 | 6.83 | 8.17 | 15.40 |
| precipitation mean[6] | mm | 606.60 | 714 | 839.3 | 910.3 | 1106.8 | 1282.7 |
| N dep mean[7] | kg/ha | 2.68 | 4.37 | 5.36 | 5.27 | 5.99 | 8.67 |
| N dep difference[7] | kg/ha | -1.49 | -0.11 | 1.47 | 1.30 | 2.47 | 4.66 |

[1] National Hydrography Dataset (NHD) 2013; and Soranno et al. 2015

[2] USGS National Elevation Dataset (NED); 2013

[3] National Land Cover Database (NLCD); 2006

[4] United States Census TIGER roads data; 2013

[5] United States Geological Survey (USGS); 1951-1980
[6] PRISM climate group 30-year normal; 1981-2010
[7] National Atmospheric Deposition Program; 1990-2010

**Table 3.** The number of lakes in the training and testing datasets for each of the seven sampling scenarios and three response variables. For all scenarios except (g), these are average numbers of lakes over multiple subsets of training and testing data. The numbers in parentheses are the percent of the total lake population ≥ 4 ha comprised for each scenario and response variable combination.

| | **TP** | | **TN** | | **CHL** | |
|---|---|---|---|---|---|---|
| **Sampling Scenario** | Training | Testing | Training | Testing | Training | Testing |
| (a) Random-Large | 4,422 | 1,474 | 3,190 | 1,063 | 5,088 | 1,696 |
| | *(9 %)* | *(3 %)* | *(6 %)* | *(2 %)* | *(10 %)* | *(3 %)* |
| (b) Random-Small | 1,474 | 4,422 | 1,063 | 3,190 | 1,696 | 5,088 |
| | *(3 %)* | *(9 %)* | *(2 %)* | *(6 %)* | *(3 %)* | *(10 %)* |
| (c) Stratified-Type | 1,474 | 4,422 | 1,063 | 3,190 | 1,696 | 5,088 |
| | *(3 %)* | *(9 %)* | *(2 %)* | *(6 %)* | *(3 %)* | *(10 %)* |
| (d) Stratified-Region | 1,474 | 4,422 | 1,063 | 3,190 | 1,696 | 5,088 |
| | *(3 %)* | *(9 %)* | *(2 %)* | *(6 %)* | *(3 %)* | *(10 %)* |
| (e) Targeted-Type | 2,869 | 3,024 | 2,079 | 2,171 | 3,393 | 3,379 |
| | (6 %) | (6 %) | (4 %) | (4 %) | (7 %) | (7 %) |
| (f) Targeted-Region | 2,927 | 2,927 | 2,127 | 2,127 | 3,392 | 3,392 |
| | (6 %) | (6 %) | (4 %) | (4 %) | (7 %) | (7 %) |
| (g) Targeted-AgRegion | 4,422 | 1,474 | 3,190 | 1,063 | 5,088 | 1,696 |
| | *(9 %)* | *(3 %)* | *(6 %)* | *(2 %)* | *(10 %)* | *(3 %)* |

## FIGURE LEGENDS

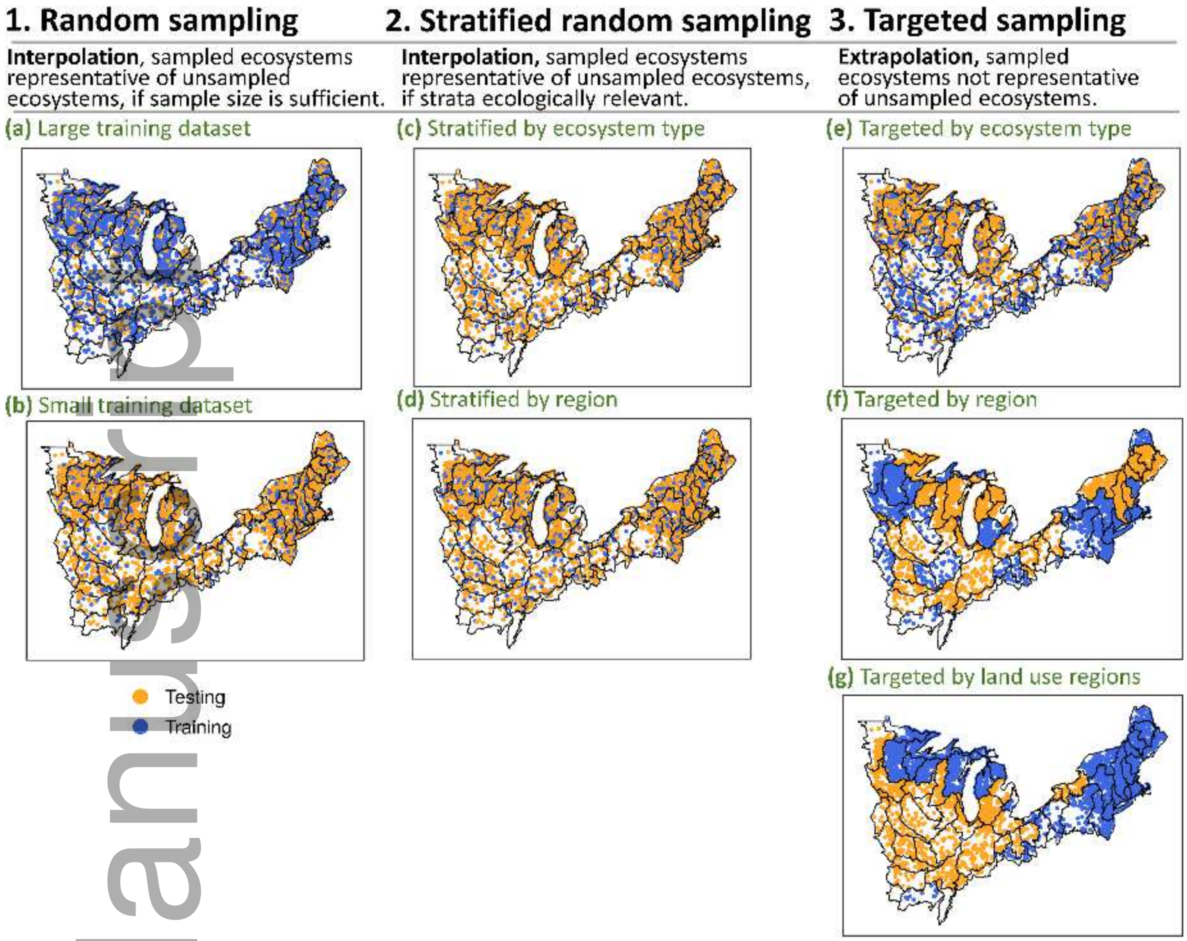
**Figure 1.** Conceptual figure depicting three types of sampling strategies (1-3) used for selecting ecosystems to sample at macroscales. Underneath each type is a description of the assumptions underlying the resulting models. In all seven depictions (a-g), there are ecosystems that are used

to build predictive models (training dataset; blue circles) and ecosystems that are used to test the predictive models (test, orange circles). From left to right: 1. Random sampling designs whereby ecosystems are chosen completely randomly from the sample extent; predictive models for unsampled ecosystems are assumed to be interpolation, if sample size is sufficient. 2. Stratified random sampling designs whereby ecosystems are first stratified by ecosystem type (top) or their location within ecological regions (regions depicted by dark lines, bottom) that are thought to drive variation among ecosystems and second, ecosystems are selected randomly within those strata; predictive models for unsampled ecosystems are assumed to be interpolation, if the strata are ecologically relevant and sample size is sufficient. 3. Targeted sampling whereby particular types of ecosystems (top), particular ecological regions (middle), or regions with particular land uses (bottom) are targeted for sampling in order to answer a particular question; predictive models for unsampled ecosystems are assumed to be extrapolation. Black lower-case letters relate to the seven scenarios used in this study that are described and depicted throughout.

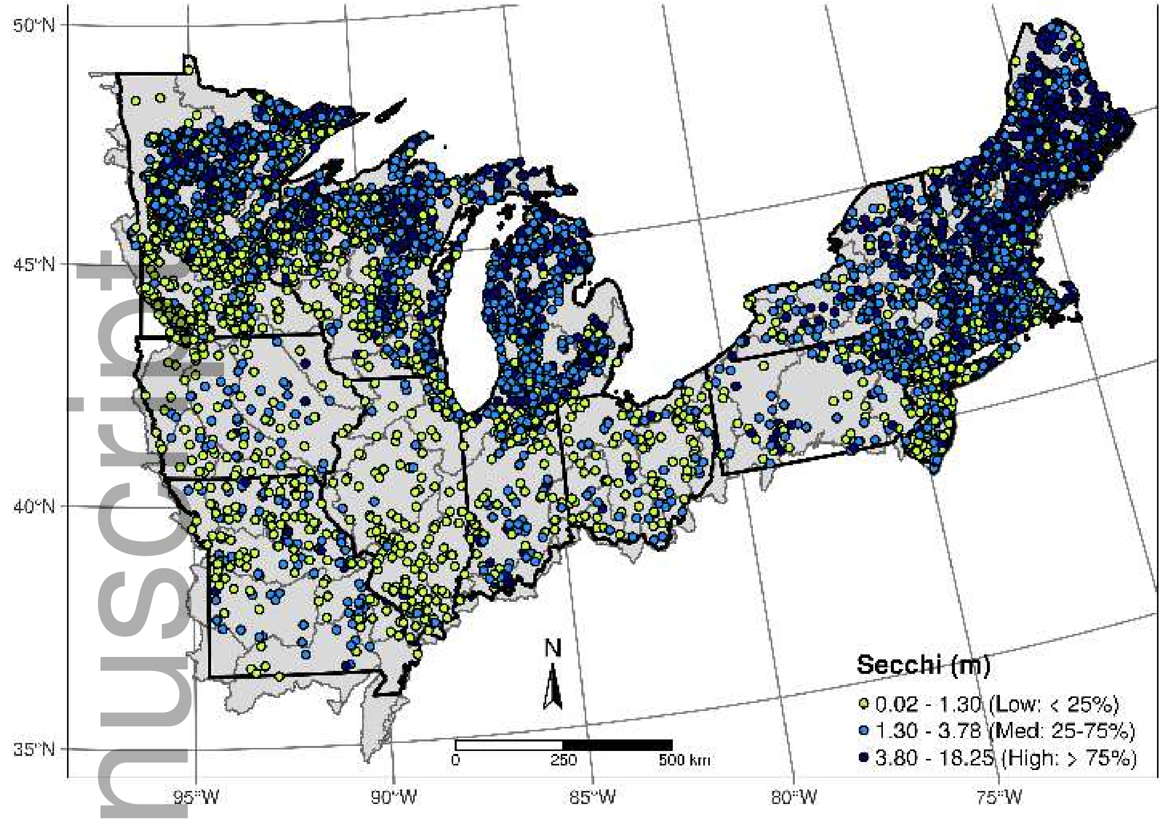
**Figure 2.** Map of lakes color coded by water clarity measured as Secchi disk depth (m), colored by percentile. Gray lines delineate regions.

**Figure 3.** Boxplots showing the model predictive performance of each scenario indicated by letters (X axis labels, letters as per Fig. 1) as measured by predictive $R^2$ (A), root mean square error (RMSE; B), and median relative absolute error (MRAE; C). The colors signify the different types of sampling strategies: random (yellow), stratified (green), and targeted (blue). Y-axis scales are truncated for better visualization.

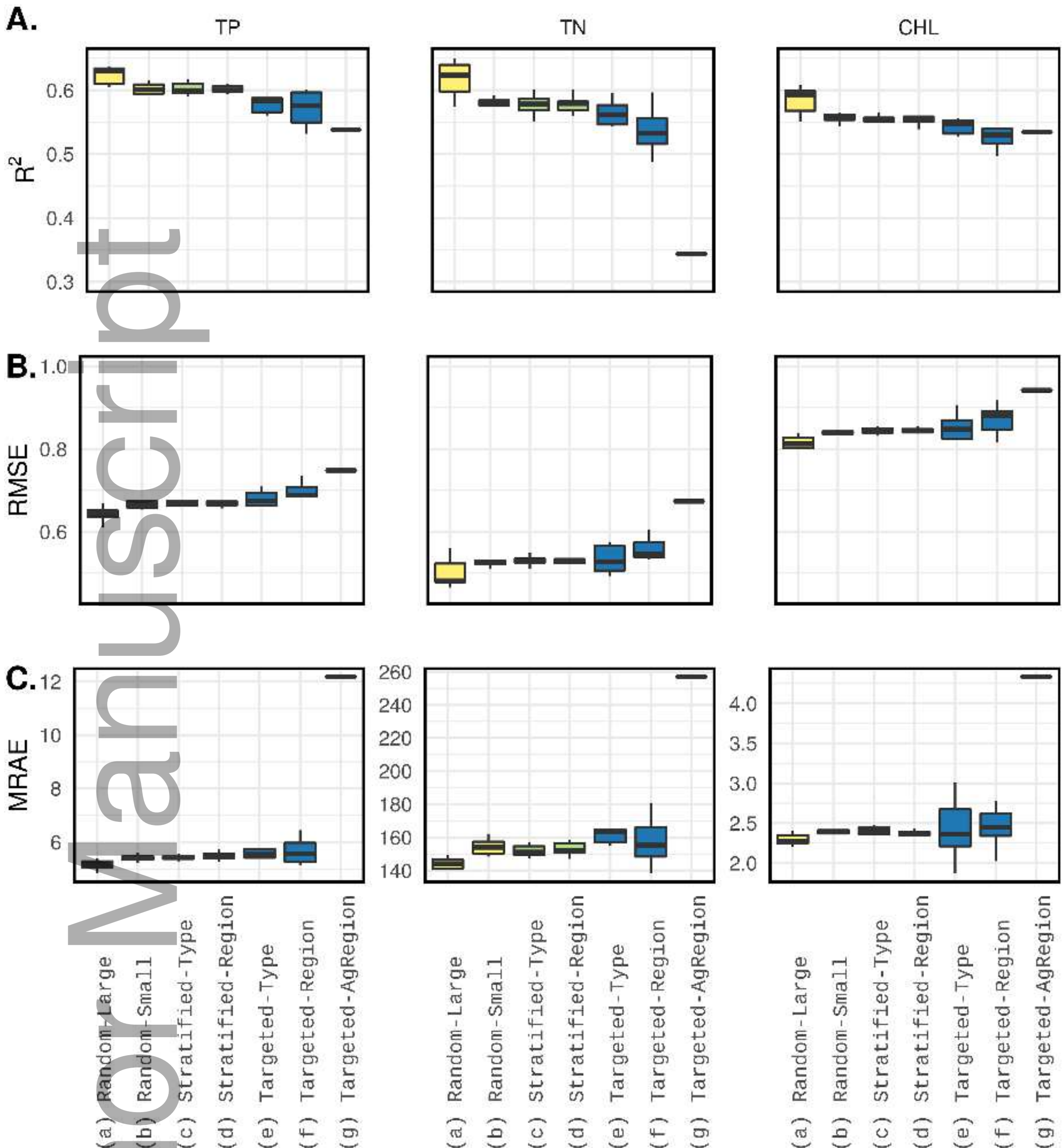
**Figure 4.** Density plots showing the distribution of data in the training (blue) and testing (orange) dataset for each sample design scenario (a-g as per Fig. 1) and for left to right: TP (μg/L), TN (μg/L), CHL (μg/L), water clarity (m), lake maximum depth (m), and watershed percent forested. One randomly selected dataset for each sample design scenario is portrayed in this figure. The X-axis is truncated and axis labels are not shown to better visualize the majority of the data for best visual comparison of the training vs test datasets. The letters are as for Figure 1.



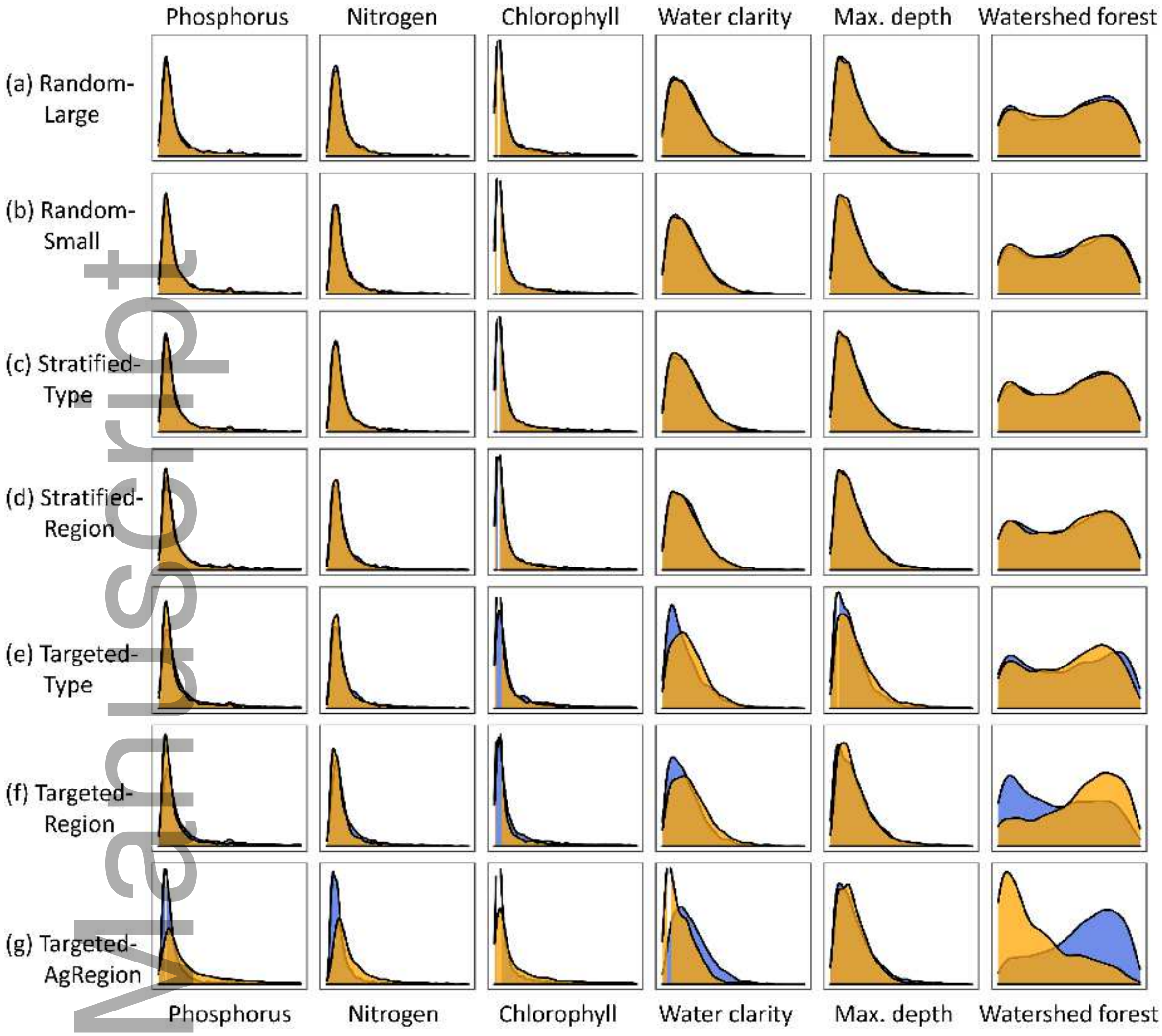

eap_2123_f1.tif




eap_2123_f2.tif




eap_2123_f3.tif




eap_2123_f4.tif