LIMNOLOGY
and
OCEANOGRAPHY

Limnol. Oceanogr. 00, 2019, 1–14

doi: 10.1002/lno.11136

Biases in lake water quality sampling and implications for macroscale research

Emily H. Stanley,[1*] Sarah M. Collins,[2] Noah R. Lottig,[3] Samantha K. Oliver,[4] Katherine E. Webster,[5] Kendra S. Cheruvelil,[6] Patricia A. Soranno[5]

[1]Center for Limnology, University of Wisconsin, Madison, Wisconsin
[2]Department of Zoology and Physiology, University of Wyoming, Laramie, Wyoming
[3]Center for Limnology, Trout Lake Station, University of Wisconsin, Boulder Junction, Wisconsin
[4]Water Science Center, USGS Wisconsin, Middleton, Wisconsin
[5]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan
[6]Department of Fisheries and Wildlife, Lyman Briggs College, Michigan State University, East Lansing, Michigan

Abstract

Growth of macroscale limnological research has been accompanied by an increase in secondary datasets compiled from multiple sources. We examined patterns of data availability in LAGOS-NE, a dataset derived from 87 sources, to identify biases in availability of lake water quality data and to consider how such biases might affect perceived patterns at a subcontinental scale. Of eight common water quality parameters, variables indicative of trophic state (Secchi, chlorophyll, and total P) were most abundant in terms of total observations, lakes sampled, and long-term records, whereas carbon variables (true color and dissolved organic carbon) were scarcest. Most data were collected during summer from larger (≥ 20 ha) lakes over 1–3 yr. Approximately 80% of data for each variable is derived from ~ 20% of sampled lakes. Long-term (≥ 20 yr) records were rare and spatially clustered. Data availability is linked to major management challenges (eutrophication and acid rain), citizen science, and a few programs that quantify C and N variables. Resampling exercises suggested that correcting for the surface area sampling bias did not substantially change statistical distributions of the eight variables. Further, estimating a lake's long-term median Secchi, chlorophyll, and total P using average record lengths had high uncertainty, but modest increases in sample size to > 5 yr yielded estimates with manageable error. Although the specific nature of sampling biases may vary among regions, we expect that they are widespread. Thus, large integrated datasets can and should be used to identify tendencies in how lakes are studied and to address these biases as part broad-scale limnological investigations.

Environmental research in the 21st century increasingly includes investigations conducted at broad spatial scales. This growth has been motivated by a need to address environmental problems unfolding at regional and continental scales as well as by the growing availability of data at these scales (Heffernan et al. 2014; Estes et al. 2018). Datasets with spatial extents of thousands to millions of hectares are being generated via both remote sensing tools and ground-based measurements, allowing researchers and managers to see and understand patterns and processes in ways that had not previously been possible (Heffernan et al. 2014).

Emblematic of this trend in environmental research, the rise of broad-scale limnological studies has been rapid (Soranno et al. 2010; Seekell et al. 2018). Several hydrologic and water quality datasets have become publicly available over the last decade and are now routinely being used to ask questions about limnological patterns and their drivers (e.g., Filstrup et al. 2014; O'Reilly et al. 2015; Dugan et al. 2017*a*; Huser et al. 2018) or in upscaling exercises to estimate aquatic contributions to biogeochemical cycles at regional, continental, and global scales (e.g., Lapierre et al. 2017; Mendonça et al. 2017; DelSontro et al. 2018). While many broad-scale water quality datasets are generated by national agencies using standardized protocols (e.g., Hamill and Lew 2006; Fölster et al. 2014; US EPA 2016), others are compilations of smaller datasets that use a variety of sampling designs or analytical protocols (e.g., Sobek et al. 2007; Hartmann et al. 2014; Sharma et al. 2015; Dugan et al. 2017*b*; Read et al. 2017).

Motives for sampling and logistical realities dictate what data are generated by a research group or monitoring program, and these inevitably differ among projects (Hughes and Peck

[*]Correspondence: ehstanley@wisc.edu

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

2008; Behmel et al. 2017). There is an old joke that limnology is the study of eutrophic lakes within driving distance of major college campuses (derived from Vallentyne [1969]). This facetious definition provides an obvious example of how such a logistical constraint (or motive) can lead to a biased and limited dataset; in this case, extrapolating from a single productive lake to a set of surrounding lakes of varying trophic states would be patently unwise. Indeed, the possibility of arriving at inaccurate conclusions about broader scale patterns because of limits or biases in available data inspired the implementation of probability-based lake sampling designs by the U.S. Environmental Protection Agency (EPA). These statistically representative approaches are intended to overcome sampling biases and ensure accurate assessment of ecosystem conditions by programs responsible for monitoring status and trends of aquatic ecosystems (Paulsen et al. 1998; Peterson et al. 1999; Hughes and Peck 2008).

In contrast to systematic designs required for water quality assessments, integrated datasets by definition incorporate multiple sampling approaches. Yet, despite their haphazard composition, this latter type of dataset is increasingly being used as a means of expanding spatial and/or temporal extents of study and increasing overall data availability. This raises the question: Are we again opening ourselves up to drawing unreliable conclusions when using these integrated datasets? The answer to this question may be "yes" if there are consistent, unrecognized biases in how monitoring and research groups sample lakes. Thus, the goal of this study was to address two general questions: Are there sampling biases in the monitoring or studying of lakes? If so, what are they and how might they affect perceived patterns among lakes at broader spatial scales?

To consider these two questions, we examined data availability for eight common limnological variables in the LAGOS-NE database (Soranno et al. 2017). LAGOS-NE is a multiscaled lake and reservoir database that integrates 87 independent datasets from 17 northeastern and north–central U.S. states, and its construction included the resolution of practical challenges that arise when harmonizing diverse datasets (as described in Soranno et al. [2015] and Sprague et al. [2017]). We used this database as a representative of the growing collection of integrated lake datasets that span multiple regions, countries, or continents to search for collective tendencies in lake sampling. To identify biases in sampling that may persist or emerge, and how such biases might affect the determination of patterns or conditions across larger spatial extents, we decomposed the problem into four specific questions:

1. What do we measure? (How are data distributed among variables?)
2. When do we sample? (How do data vary within and among years?)
3. Where do we sample? (How are data distributed across the region and among lake types?)
4. What are the implications of these biases? (Do sampling tendencies influence estimates of lake state at large spatial scales and if so, how?)

Our answers reveal the existence of sampling biases along the three axes of what, when, and where. Yet, they also suggest that not all biases are problematic, and there may be ways of managing biases to arrive at robust understanding of broad-scale limnological patterns.

Methods

We addressed our questions by examining water quality, morphometric, and geographic data for lakes and reservoirs in the "LIMNO," "LOCUS," and "GEO" modules of LAGOS-NE version 1.087.1 (Soranno and Cheruvelil 2017*a*,*b*,*c*; Soranno et al. 2017). Data in LAGOS-NE are derived from 87 different sources across a 17-state region of northeastern and north–central United States—an area that includes 141,265 lakes with surface areas > 1 ha. Database construction and content are described in detail by Soranno et al. (2015, 2017). Version 1.087.1 modules were acquired using the LAGOS R package (Stachelek et al. 2017), and data in the LAGOS-NE$_{\text{LIMNO}}$ module were supplemented with additional information from the State of New Hampshire. LAGOS-NE excludes chlorophyll uncorrected for phaeophytin and thus the majority of New Hampshire records for this variable. We chose not to discriminate between the two methods (following Stich and Brinker [2005]) and acquired chlorophyll data from the New Hampshire Department of Environmental Services. The updating process also brought in some additional information for other variables due to recent database improvements that led to greater data availability in 2016 when the request was made. Other modifications to the downloaded LAGOS-NE$_{\text{LIMNO}}$ data included removal of duplicate entries of Secchi disc depth for several lakes in Minnesota and correcting entries from one Wisconsin monitoring program in which concentrations of N species had not been converted to standard LAGOS-NE units. We also excluded lakes described as "out of county state" (the Laurentian Great Lakes and lakes that span the U.S.–Canada border). Finally, we truncated the dataset at 2010, because data availability reached a peak in 2010 then declined sharply in subsequent years, likely reflecting the time frame over which information was gathered from data sources during the construction of LAGOS-NE (Soranno et al. 2015).

LAGOS-NE$_{\text{LIMNO}}$ includes 17 in-lake variables (Soranno et al. 2017) derived from epilimnetic sample collection at a single point per lake. For the purposes of examining patterns of data availability, we narrowed our scope to focus on a core group of eight parameters: chlorophyll *a* (hereafter abbreviated as Chl *a*), true color (Color), dissolved organic carbon (DOC), ammonium-N (NH_4), nitrate+nitrite-N (NO_3), water clarity measured as Secchi disk depth (Secchi), total nitrogen (TN), and total phosphorus (TP). We included both direct and indirect (calculated as TKN + NO_3) measurements for TN, and used the direct measurement if both were available. We refer to these variables in three groups as: trophic (Secchi, Chl *a*, and TP), nitrogen (TN, NH_4, and NO_3), and carbon (Color and DOC) groups. In some cases, Chl *a*, TN, and DOC are used to represent each group in presentation of results, with details for all variables reported in the Supporting Information. To address Question 1 (*What do we measure:* How are data distributed among variables?), we determined data availability in terms of total data points and the types and amounts of data per lake for all variables. Temporal aspects of data availability (Question 2: *When do we sample:* How do data vary within and among years?) were examined in terms of within-year sampling, the changes in overall data availability through time, the length of sampling records, and the availability of long-term datasets. We used a criterion of data from 20 different years per lake to identify long-term records, although we did not require that sampling years be consecutive.

LAGOS-NE_{GEO} includes geographic information for all lakes and reservoirs > 1.0 ha (hereafter referred to as the census lakes or the census population), which allowed us to compare attributes of the subset of lakes that have been sampled to the entire population to identify sampling biases (Wagner et al. 2008) to address Question 3 (*Where do we sample:* How are data distribute across the region and among lake types?). For each water quality variable, we compared the percent of census lakes within each state to the percent of data each state contributed to LAGOS-NE and surface areas and depths of lakes with data relative to the census population.

For Question 4 (*What are the implications:* Do sampling tendencies influence estimates of lake state at large spatial scales?), we undertook two analyses to explore possible consequences of uneven data distribution. First, we considered the case of potential biases associated with lake surface area because of the availability of data for this attribute coupled with the expectation of a lake size sampling bias based on prior studies (Peterson et al. 1999; Wagner et al. 2008). Following a similar analysis by Hanson et al. (2007), we compared 1000 random samples drawn from LAGOS-NE_{LIMNO} for each water quality variable to a second set of 1000 samples selected using a stratified random approach that accounted for the distribution of lake surface areas in the census population. While this approach incorporates a range of random effects associated with multiple data sources and locations in LAGOS-NE, it provides a relatively straightforward and objective means of determining if conspicuous differences in averages or variances are associated with biased sampling of lakes based on their size. The proportion of lakes within each of 16 bins was determined from the log-normalized distribution of lake area of the census population. Lakes with data were randomly resampled within bins, and bins were sampled in proportion to the census lakes distribution to generate the set of 1000 samples. Observed (uncorrected for sampling bias) and corrected (for lake size bias) distributions were then compared using violin plots and associated summary statistics.

Because lakes are often not well-sampled or evenly sampled through time, our second resampling exercise considered how record length might influence estimation of a lake's long-term median value of a given variable. The general approach was to identify lake records composed of samples that were well-distributed through time, and then to resample these records to simulate scenarios of varying levels of sampling effort and to compare the results to the long-term ("true") median from the entire dataset. Originally, we attempted to use lakes with observations from all seasons of a year over 20 or more years. The within-year requirement proved to be too stringent relative to data availability, so we focused on summer months and stipulated the presence of observations in the second and third quarters (April–June and July–September) of each year. We also required that at least 100 lakes meeting these criteria must be available to include a variable in the analysis. Records from lakes that met these requirements were subsampled to generate secondary datasets by randomly selecting: a single weekly sample, a single monthly sample from the weekly samples, and finally, a single quarterly sample from the monthly samples for each year. This process produced a collection of secondary lake records, each of which was made up of 40 or more measurements that were well-distributed across the entire 20+ yr period. We then randomly resampled each secondary dataset 1000X using sample sizes of 1, 2, 3, 5, 10, 20, 30, and 40 and calculated a lake's median for each sample size. The difference between each sample size median and the lake's "true" median was then expressed as a relative "error":

$$\text{Percent error} = \left| \left(\frac{\text{sample median} - \text{true median}}{\text{true median}} \times 100 \right) \right|$$

so that results were comparable across all lakes and among all variables regardless of the absolute value of the medians. Finally, the distributions of the percent errors for the eight sample sizes were evaluated using violin plots and associated summary statistics. All analyses were conducted in R (R Core Team 2018); data files and code used to generate results are available through the Environmental Data Initiative (Stanley et al. 2019).

Results

Q1—what do we measure?

Trophic variables—Secchi, Chl *a*, and TP—dominate the set of eight variables considered here, both in terms of number of lakes sampled and total data points (Table 1). Secchi is most abundant, representing over 50% of all data in our LAGOS-NE subset. N species, led by NO_3, were quantified more frequently and in more lakes than carbon variables (Color and DOC). Each of the eight variables were sampled in an average of 5.8% of the census lakes, and at least one variable has been quantified in just over 10% of the region's 141,265 lakes. Most lakes that have been sampled have information on multiple

Table 1. Distribution of data among variables in terms of the percent (and number) of lakes > 1.0 ha within the LAGOS-NE region (i.e., the "census" population of lakes) with data for each variable and the total number of data points for the dataset considered in this study.

Group	Variable	% of census lakes with data (*n*)	% of all data (*n* data points)
Trophic	Secchi	9 (12,377)	54 (708,172)
	Chl *a*	6 (8525)	15 (197,868)
	TP	7 (10,490)	11 (148,759)
Nitrogen	TN	5 (6553)	4 (56,970)
	NH_4	5 (6502)	4 (47,525)
	NO_3	6 (8173)	5 (68,478)
Carbon	Color	4 (5636)	3 (43,542)
	DOC	4 (4997)	2 (29,287)

variables; 13.6% of lakes have only one variable quantified, whereas > 55% have data for five or more variables.

Q2—when do we sample?

On the annual time scale, limnological data are dominated by measurements made between 01 June (day 152) and 30 September (day 273), with an average of 77%, 61%, and 56% of data collected in this window for trophic, N, and C groups, respectively (Fig. 1). A second peak is apparent for the C and N groups during the fall (October–December). Data for DOC are the most evenly distributed throughout the year, and along with Color, are characterized by a regular monthly pattern of data production outside the summer window.

Although the earliest record in LAGOS-NE is from 1933, proliferation of water quality data began in earnest in the 1970s (Fig. 2). The notable peak for most variables in 1984 corresponds to the EPA's Eastern Lakes Survey, which included over 1500 lakes. A similar, though much smaller, peak can also be seen for EPA's 2007 National Lakes Assessment. The Adirondack Long-Term Monitoring Program also began in 1984, adding an additional ~ 350 lakes. Despite the striking accumulation of data over the past five decades, the number of lakes sampled per year has increased slowly or not at all for variables in the N and C groups since the early-mid 1990s (Fig. 2).

Long-term datasets are relatively sparse, ranging from 1229 lakes with 20 or more years of data for Secchi to a low of only 53 lakes for TN (Table 2). Not surprisingly, most long-term datasets fall within the trophic group and are much scarcer within the N and C groups. At the other end of the spectrum, most lakes with water quality data were sampled over a period of 1–3 yr, leading to few observations per lake (Table 2). The percentage of lakes that have data from only a single date over the entire 77-yr span of LAGOS-NE ranged from 17% (Chl *a*) to 65% (DOC; Table 2). Overall, well-sampled variables are well-sampled both within and among lakes, whereas poorly

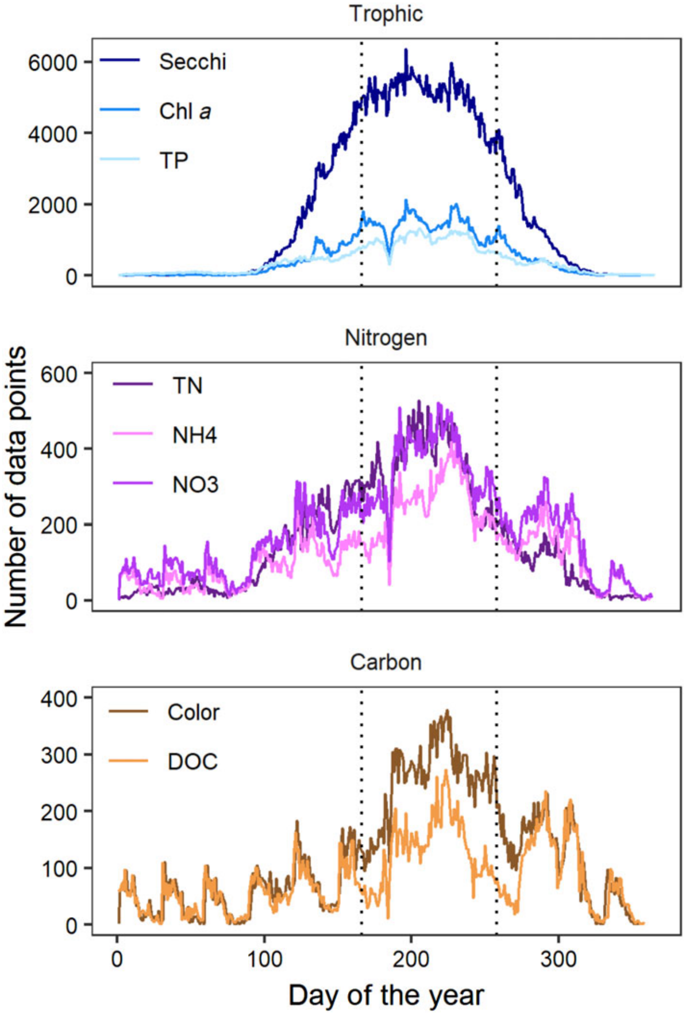

Fig. 1. Within-year data availability for the trophic (top), nitrogen (middle), and carbon (bottom) variables. Dashed vertical lines represent 01 June and 30 September. Note difference in *y*-axis scales.

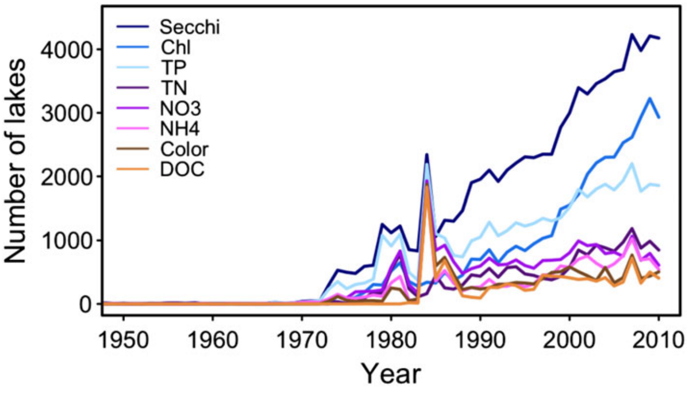

Fig. 2. The number of lakes with data in each year for the eight focal variables. The timeline was truncated at 1945 to highlight major trends, although earlier records are present in LAGOS-NE.

Table 2. Analysis of sampling intensity per lake for the eight focal variables, including number and percent of lakes with single site visits, lakes with long-term data records, average (and maximum) number of observations (*n*) per lake, and the average (and maximum) number of distinct years individual lakes were sampled.

Group	Variable	Number of lakes sampled 1X (% of all sampled lakes)	Number of lakes with 20+ yr of data (% of all sampled lakes)	Median *n*/ lake (max)	Median number of years with observations (max)
Trophic	Secchi	3148 (25)	1229 (10)	7 (3434)	3 (41)
	Chl *a*	1434 (17)	271 (3)	3 (3075)	3 (33)
	TP	3662 (35)	386 (4)	3 (2861)	2 (34)
Nitrogen	TN	1920 (29)	53 (1)	3 (489)	2 (27)
	NH_4	3080 (47)	60 (1)	2 (395)	1 (29)
	NO_3	3507 (43)	109 (1)	2 (1055)	1 (27)
Carbon	Color	3576 (63)	95 (2)	1 (432)	1 (30)
	DOC	3256 (65)	62 (1)	1 (361)	1 (27)

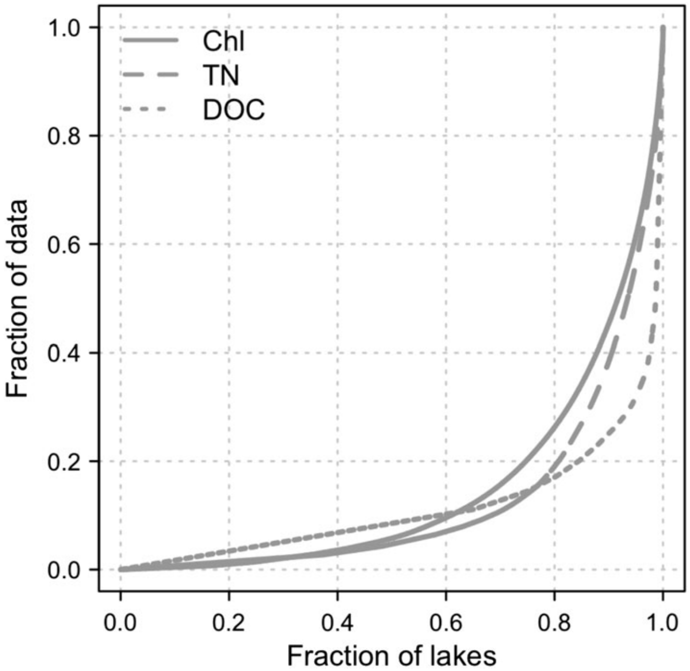

Fig. 3. Cumulative frequencies of data accumulation as a function of the percent of lakes contributing to the dataset for variables representing trophic (Chl *a*), nitrogen (TN), and carbon (DOC) groups.

sampled variables typically have one or a few measurements per lake drawn from a limited number of lakes. The effect of sampling many lakes rarely and few lakes consistently over time is that roughly 80% of LAGOS-NE_{LIMNO} data for each variable are derived from ~ 20% of sampled lakes (Fig. 3; Supporting Information Fig. S1). Color represents the extreme case with just 9% of the sampled lakes accounting for 80% of the data.

Q3—where do we sample?

Limnological data are unevenly distributed across the LAGOS domain. For trophic variables, there is a reasonable concordance between the percent of lakes in any one state and the percent of sampled lakes from that state that are in the database (Fig. 4). For example, Minnesota, which contains more lakes than any other state (21% of all lakes in the 17-state area), also has the largest fraction of lakes with Chl *a* and Secchi data in LAGOS-NE (29–32% respectively). This concordance is weaker for the N and C groups, however. New Hampshire, Maine, and New York are responsible for a disproportionately large percentage of the lakes with N data, whereas Wisconsin, New York, and Maine similarly overcontribute to the carbon group. Individual research and monitoring programs can have a major influence on these patterns, as was the case for N and C groups, in which a substantial fraction of data was generated by the Adirondack Long-Term Monitoring Program in New York.

The geographic unevenness of data is particularly acute for long-term records. As of 2010, 7 of the 17 states (Connecticut, Illinois, Indiana, Iowa, Massachusetts, New Jersey, and Ohio) had no lakes with 20 yr of data for any of the eight variables, whereas Pennsylvania had one lake with long-term data (Fig. 5; Supporting Information Table S1). Trophic records are most abundant in Missouri and the northern half of the study region. Long-term data availability is progressively rarer for N and C variables; such records for any of the three N variables are present in nine states and in just five states for C variables. DOC represents the extreme case; long-term lake data occur in three geographic clusters across four states (Maine, New York, Wisconsin, and Michigan), with approximately 75% of all records derived from lakes in the Adirondack region of New York.

With respect to lake attributes, our analysis confirmed the preferential sampling of larger lakes, with the extent of this departure varying by both state and variable type (Fig. 6; Supporting Information Fig. S2). Surface area is positively associated with an increased likelihood that a lake is sampled, reaching the point where 40–87% of all census lakes between 1,000 and 10,000 ha have data for all eight variables

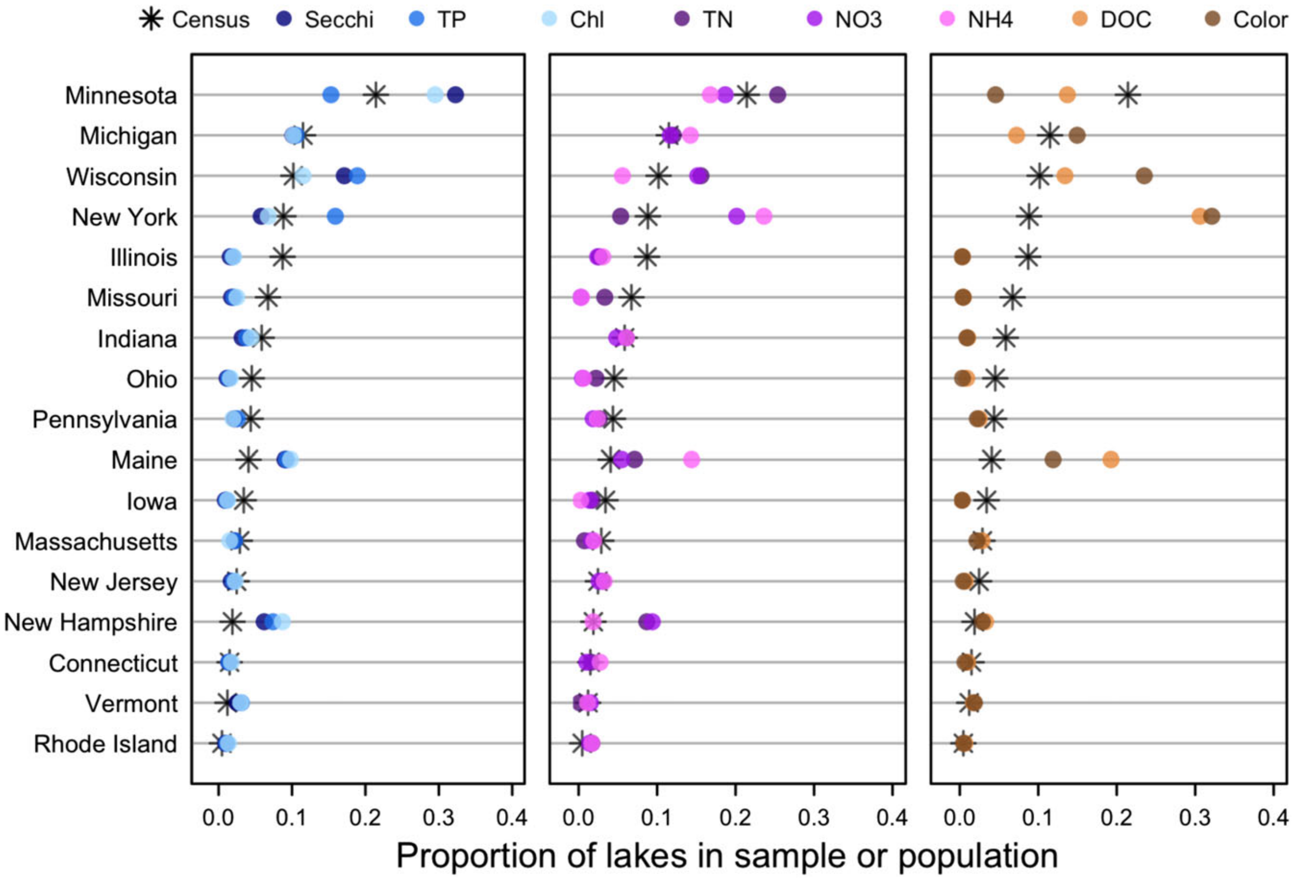

Fig. 4. The proportion of all lakes present in each state (census) present in the census population and the proportional contribution of lakes in each state to LAGOS-NE (i.e., proportion of lakes in the sample) for each of the eight selected variables.

(Supporting Information Fig. S3). The tendency to sample larger lakes was related to an undersampling of abundant hydrologically isolated lakes and oversampling of drainage lakes low in the landscape (Supporting Information Fig. S4). Finally, we were unable to make a reasonable comparison between lake depths of the subset of sites that were sampled for water chemistry and depths for the entire population of census lakes. Depth data were available for < 10% of all lakes (10,363 lakes), and most lakes with depth information (93%) have also been sampled for chemistry. Thus, the two datasets were virtually identical.

Q4—how do sampling tendencies influence estimates of lake conditions?

Resampling the LAGOS-NE data in a stratified random fashion to correct for the surface area bias led to relatively small changes in data distribution for most variables. For trophic variables, the overall shapes of observed and corrected sample distributions had varying degrees of divergence but did not result in significant changes (Fig. 7; Supporting Information - Table S2). Median Secchi for corrected samples was slightly shallower (1.8 m) relative to observed data (2.3 m), which is marginally larger than reported differences in simultaneous measurement of Secchi by two different samplers (15% [Häkanson 1992] or 0.2 m [Obrecht et al. 1998]). Chl *a* and TP medians were just slightly higher for the corrected dataset (1.3 and 3 μg L^{-1} for Chl *a* and TP, respectively) although there was substantial overlap in the first to third quartile ranges of observed and corrected datasets for all three variables. The Chl *a* and TP differences are smaller than those reported in an intercomparison of simultaneous sampling by university researchers and citizen scientists (3.8 μg L^{-1} for Chl *a* and 5 μg L^{-1} for TP; Obrecht et al. 1998). There was a small shift toward higher Color and DOC values among corrected samples, although this effect was modest for DOC (corrected median of 6.70 mg L^{-1} with first to third quartiles of 4.5–9.9 vs. observed median of 5.60 mg L^{-1} and first to third quartiles of 3.9–8.4; Supporting Information Table S2) compared to the corrected Color data (corrected median of 35 Platinum Cobalt Units [PCU], with first to third quartiles of 16–80 vs. observed median and interquartile range of 20 PCU and 10–40). The slight shift toward higher concentrations for corrected samples was absent for all three N variables, as medians for NH_4 were identical (18 μg L^{-1}) and the median and maximum values were lower for corrected TN and NO_3 datasets relative to observed (Supporting Information Table S2).

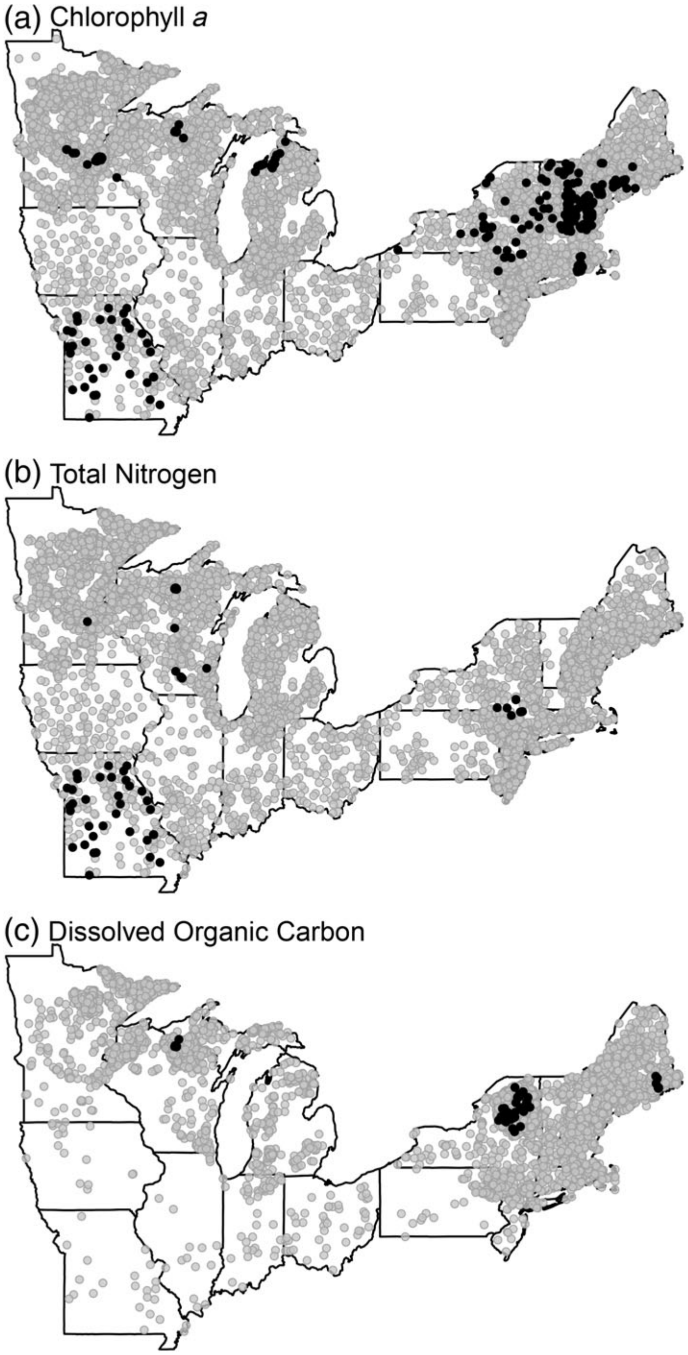

Fig. 5. Maps of the LAGOS-NE domain illustrating the location of all lakes (gray) and lakes with 20 or more years of data (black) for representative variables. See Supporting Information Table S1 for counts of long-term records for each state.

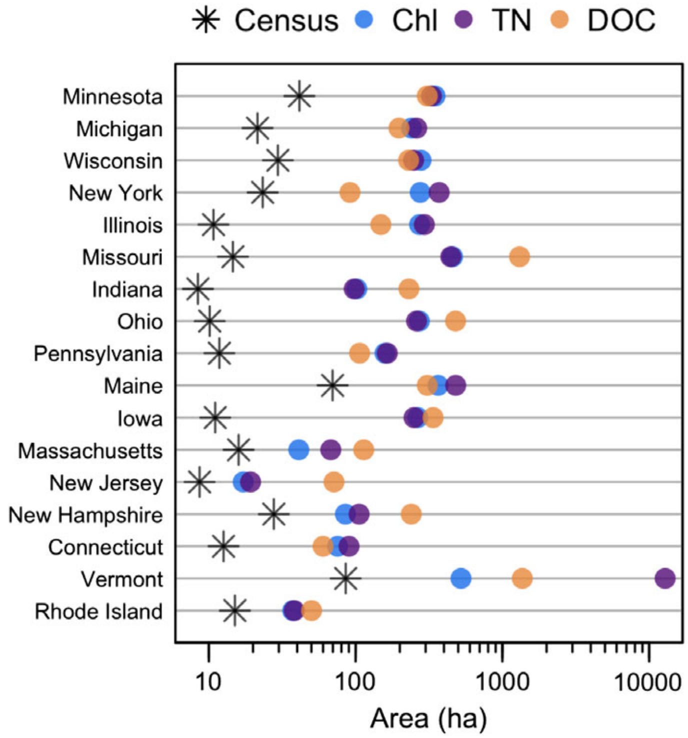

Fig. 6. Comparison of median surface area for all lakes present in each state (census) and median surface areas of lakes with data for representative variables (Chl *a*, TN, and DOC) within each state in the LAGOS-NE database.

With respect to effects of sample size on estimates of "true" lake medians, only the three trophic variables met our criteria for inclusion in the analysis (100 or more lakes with observations from the second and third quarters of the year from 20 or more years). As expected, increasing the sample size decreased both the range and magnitude of the average percent error of estimated medians for all variables (Fig. 8). The average error of the "true" median based on a single sample varied from 18% (Secchi) to 38% (Chl *a*; Supporting Information Table S3), with the average percent error dropping below 10% for sample sizes of 20, 5, and 10 for Chl *a*, Secchi, and TP, respectively. However, even as the average percent error declined below 10%, the associated ranges varied from 2.4–28% for Chl *a* to 0–21% for Secchi. The median number of observations per lake for Secchi is 7 (Table 2), and this corresponded to a median error of 7.5% (range = 0–34%). Similarly, three observations per lake (the median sample size for both Chl *a* and TP) corresponded to average errors of 24% (11–66%) for Chl *a* and 16% (0–33%) for TP.

Discussion

Our examination of limnological data compiled from multiple sources demonstrated a collective water quality sampling bias toward trophic state variables (Secchi, Chl *a*, and TP), larger lakes positioned lower in drainage networks, summer, and "snapshot data" (i.e., derived from one or a few lake visits). Conversely, these tendencies are coupled with a scarcity of multiyear records, limited data availability for nitrogen and carbon variables, and spatial variation in the nature or degree of all the above tendencies.

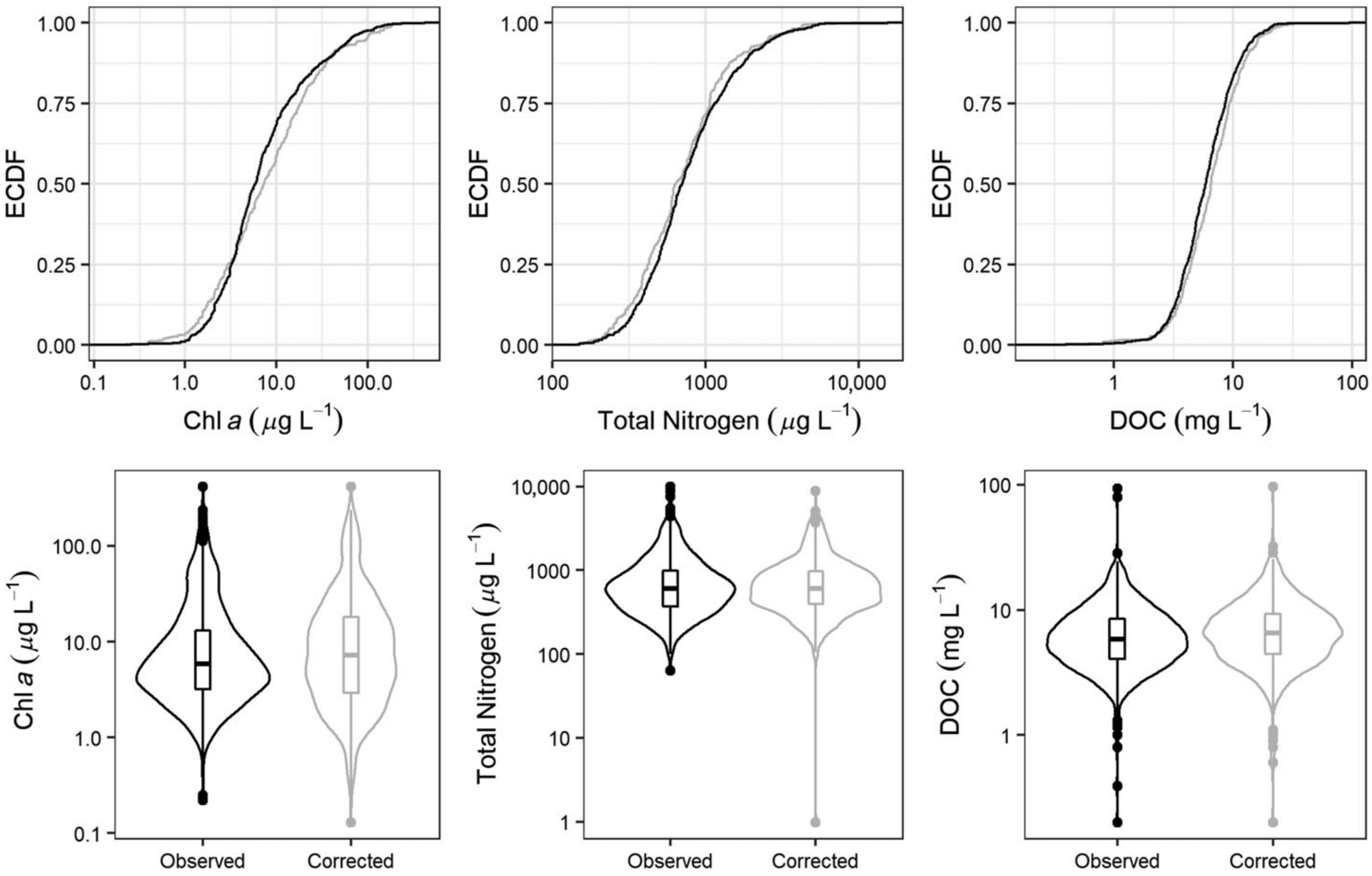

Fig. 7. Empirical cumulative frequency distributions (ECDF) and violin plots of observed data (black) and data corrected for the actual distribution of lake areas (gray) for representative trophic (Chl *a*), nitrogen (TN), and carbon (DOC) variables. See Supporting Information Table S2 for statistical summaries of distributions for all variables.

What we measure

The dominance of trophic data, especially Secchi, is not surprising given widespread and persistent concerns about eutrophication in the United States and beyond (Dodds et al. 2009; Smith and Schindler 2009) and the ease of measurement and extensive citizen science programs that collect water clarity data (Bigham Stephens et al. 2015). Similarly, the strong preference for quantifying phosphorus but not nitrogen (or carbon) reflects the long-standing identification of phosphorus as the nutrient that limits primary production and dictates trophic state (e.g., Dillon and Rigler 1974; Schindler 1974; Correll 1998). We also saw an influence of acid precipitation on variable choices. Among the N and C groups, NO_3 was the best quantified because of its inclusion in research and monitoring programs that emphasized acid–base chemistry. Some of these programs also measured Color and/or DOC, and most acid rain monitoring focused on regions with low acid neutralizing capacity (e.g., EPA's Eastern Lakes Survey, Adirondacks Long-Term Monitoring Program in northeastern states). To a very large degree, acid rain has shaped the geography of C and N data for lakes.

Unfortunately, the focus on trophic variables combined with the relatively narrow geography of acid rain impacts and associated C and NO_3 monitoring has put researchers and managers in a weak position for addressing other known or emerging water quality issues. As a case in point, it is now clear that nitrogen can play a role in the occurrence of harmful cyanobacterial blooms (Gobler et al. 2016) and can limit or colimit phytoplankton growth in some places (Bergström and Jansson 2006) or at some times of the year (Søndergaard et al. 2017). Similarly, although acid precipitation has abated in many parts of northeastern United States and northern Europe, N deposition has continued and expanded into new areas, and some highly agricultural regions are now receiving high atmospheric inputs of N in the form of NH_4 rather than NO_3 (Du et al. 2014). Yet, in areas that were not historically exposed to acid rain, there is little limnological context for considering the consequences of such shifts in form and/or amount of N inputs to lakes.

When we sample

Water quality data display conspicuous temporal biases both within and among years. At the annual scale, the bias

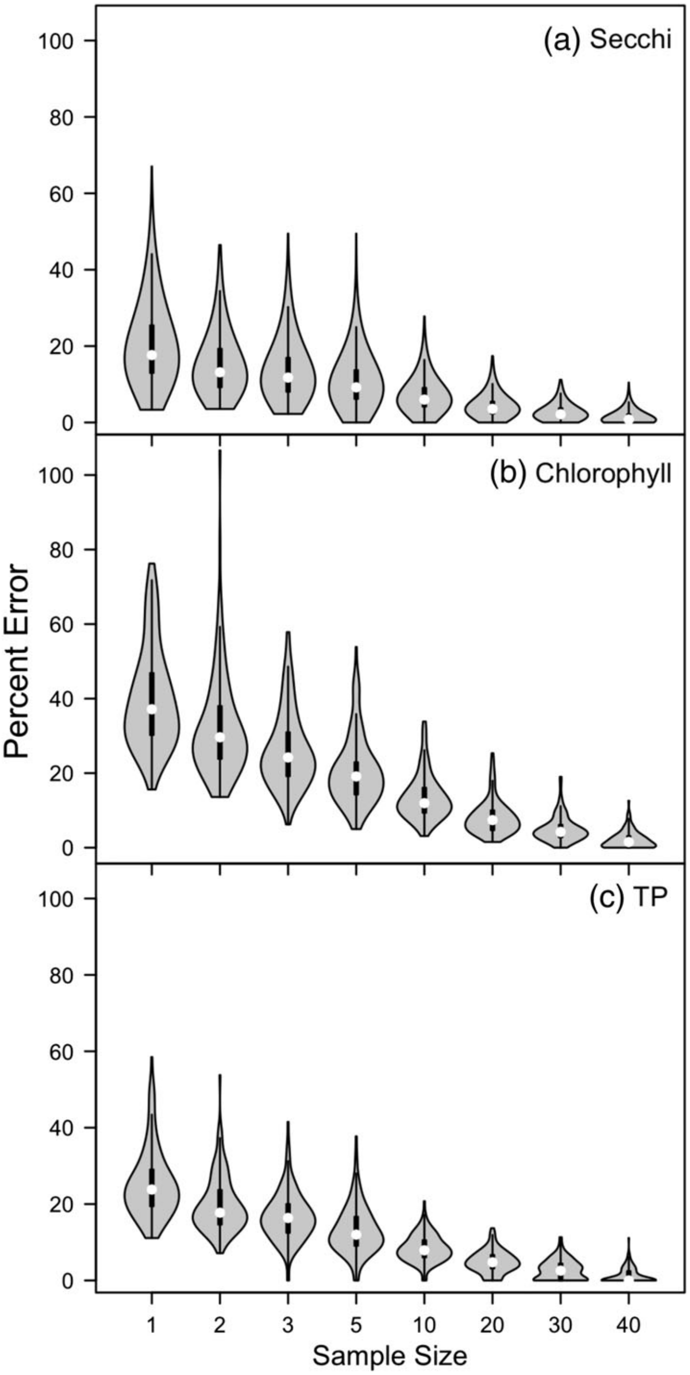

Fig. 8. Violin plots of the percent error of medians of Secchi (a), Chl *a* (b), and TP (c) estimated from > 100 lakes using different sample sizes relative to a lake's "true"(long-term) median. See Supporting Information Table S3 for statistical summaries of distributions.

toward summer sampling is widespread and is particularly pronounced for trophic variables. The more even within-year distributions of Color and DOC data result from the regular sampling regime of the New York Adirondack Long-Term Monitoring Program and, to a lesser degree, the North Temperate Lakes Long-Term Ecological Research program (Wisconsin), which account for 83% and 13%, respectively, of all DOC data collected during the typically ice-covered months of January, February, and March. Summer biases undoubtedly reflect easier access to lakes and an expectation of reduced productivity in colder months, especially for lakes that freeze. However, recent meta-analyses have illustrated that lakes remain metabolically active while ice covered and argued that the focus on warm-season sampling limits both the general understanding of winter dynamics and the capacity to predict how lakes are likely to respond to unfolding changes in climate (e.g., Bertilsson et al. 2013; Denfeld et al. 2016; Hampton et al. 2017).

The second conspicuous result regarding the temporal distribution of data is the surprisingly short duration of most lake records. This brevity is evident from medians of 1–3 yr of data per lake for all variables, as well as by the high percentages of lakes with just a single observation per variable. The complement to very short data records is the small number and percent of lakes with several years of observations. Even for Secchi, the number of lakes with long-term (20+ yr) records is surprisingly low (*n* = 1229). While this number represents a healthy 10% of all lakes with Secchi data, it translates to < 1% of census lakes in the LAGOS-NE region. Optimistically, our accounting of long-term records should now be conservative, as continued sampling since the 2010 cut-off will have increased the number of sites with multiyear records. However, support for long-term monitoring has waned in recent years (National Research Council 2004; Hughes et al. 2017), including for water quality programs (Sprague et al. 2017), and so the number of additional lakes that now meet this 20-yr criterion may be smaller than expected.

Our interest in the availability of long-term data is motivated by the unique value of these records for documenting and understanding change over time scales of years to decade, for reconstructing large-scale historic changes in lake conditions in concert with remote sensing imagery (Boucher et al. 2018), and the disproportionately large contribution that long-term research and monitoring programs have made to development of ecological ideas, natural resource management activities, and environmental policy (Lindenmayer et al. 2010; Hughes et al. 2017). Unfortunately, limitations that are already in place for studying long-term changes due to small sample sizes are exacerbated by two additional realities. First, while a 20-yr record is often sufficient for identifying trends or generating robust predictions for some variables (Knowlton and Jones 2006; Lottig and Carpenter 2012; Henson et al. 2016), in other cases, 30 or more years of data with multiple samples per year and without gaps may be needed to detect even strong directional change (Henson et al. 2016; Gray et al. 2018). As of 2010, such a threshold was not reached in any lake for half of the eight variables. Second, the uneven geographic distribution of these sites provides a restricted view of long-term dynamics in U.S. lakes at decadal time scales in general and for variables such as TN or DOC in particular. Thus, current perceptions of how lakes vary and if they are experiencing directional changes at these time scales are being shaped by the dynamics of a small number of geographically clustered sites. This same pattern also occurs at the global scale, given high densities of long-term water quality datasets in Canada and northern European countries and sparse information elsewhere (e.g., *see* Hampton et al. 2017; Dugan et al. 2017*b*; McCrackin et al. 2017).

Despite the scarcity of lakes with long-term records, a remarkable amount of limnological data has been collected over the last three decades across the LAGOS-NE domain and beyond. However, the variety of sampling dates and typically small number of samples per site lead to questions regarding the reliability of this information for characterizing lakes at large spatial scales. Average errors in estimating a lake's long-term median based on the average number of observations per lake varied from 7.5% (Secchi) to 24% (Chl *a*). Translating percentages to absolute values for Chl *a*, a lake with a long-term median of 5.2 μg L^{-1} (the grand median Chl *a* for all LAGOS-NE observations) that was sampled three times could produce an estimate of 4.0 or 6.4 μg Chl *a* L^{-1}. This appears to be a moderate amount of error, although even this small difference could result in a lake being placed into different trophic categories in some lake classification schemes (e.g., Carlson 1977).

Results of this sample size analysis should be interpreted carefully given our underlying assumptions. We referred to a lake's long-term median for Secchi, Chl *a*, and TP as the "true" median, and assumed that 40 or more observations from 20 different years are sufficient for determining this "true" median. Built into this decision is the additional assumption that all lakes have some persistent average condition for these three variables. Clearly, this is unlikely to be the case. Our rationale for including all lakes regardless of their long-term dynamics was to mimic the use of data from different years in large-scale studies (e.g., Filstrup et al. 2014; Collins et al. 2017; DelSontro et al. 2018) and because of evidence that only a small fraction of lakes (< 11%) with observations distributed over several years in the LAGOS-NE region have changed significantly over time for Secchi (Lottig et al. 2014), TP, and Chl *a* (Oliver et al. 2017). Indeed, the modest percent errors calculated in our resampling exercise are consistent with observations of relatively static conditions in most of these systems (Oliver et al. 2017) and suggest that an average summertime state for these variables can be reasonably approximated from > 5 temporally distributed observations per lake. Nonetheless, these results should be used cautiously going forward, particularly in the context of ongoing environmental change.

Where

The preference for sampling large lakes—or avoiding small, hydrologically isolated lakes—has been described in prior studies within (Peterson et al. 1999; Wagner et al. 2008) and beyond (Hamill and Lew 2006) the LAGOS-NE domain. In the United States, this well-known bias inspired the use of probability-based surveys in which sampling efforts are stratified by the distribution of lake surface areas (Peterson et al. 1999; Peck et al. 2013). Preferential oversampling of larger lakes is driven by ease of sampling, greater human use of these water bodies, and the corresponding interest and need to monitor and manage widely used natural resources (Paulsen et al. 1998)—motives that are notably distinct from evaluating status and trends of a large population of lakes distributed across a broad geographic area.

Given evidence of differences between small and large lakes for a variety of attributes (*see* Downing 2010) and the emphasis that has been placed on matching survey designs to lake surface area distributions to ensure the veracity of broad-scale assessments, we were surprised to find that correcting for the lake area bias produced only subtle differences in the statistical distributions of the eight focal variables. We can imagine three possible explanations for this result. First, differences in lake surface area have been found to affect some measures of lake water quality, but not others. Variation in DOC concentrations among lakes has been related to surface area differences in several studies (e.g., Xenopoulos et al. 2003; Hanson et al. 2007; Kankaala et al. 2013), consistent with the slight upward shift in median DOC and Color after we corrected for lake area bias in the LAGOS-NE data. However, similar evidence for other variables considered here is limited. For example, Hanson et al. (2007) found a surface area effect for TN as well as DOC but not for TP. Similarly, surface area was not a significant predictor of N and P among lakes distributed across New Zealand (Abell et al. 2011), of P among reference lakes in Europe (Cardoso et al. 2007), or of TP, TN, or their ratio in lakes of the LAGOS-NE study area (Collins et al. 2017). Second, surface area effects may be weak or difficult to detect among lakes > 1 ha. Significant relationships between surface area and limnological attributes often include systems < 1 ha (e.g., Xenopoulos et al. 2003; Hanson et al. 2007; Holgerson and Raymond 2016), and these relationships can be particularly steep at the smaller end of the size gradient (e.g., Kankaala et al. 2013), where the distinction between lakes and wetlands becomes blurred (Thornton et al. 2016). Finally, other factors not taken into consideration (e.g., geology and land use) may overwhelm a surface area effect at the scale of the entire LAGOS-NE region. In this case, the influence of surface area could be secondary or may in fact have little capacity to explain observed variance in water quality parameters of interest here.

Regardless of the presence or absence of a lake area effect, two points can be extracted from our comparison of observed and corrected datasets. First, the rationale for incorporating surface area in broad-scale surveys is that all else being equal, this attribute accounts for some amount of the observed variance in water quality measurements. Several lake and landscape features can be robust predictors of water quality (e.g., land use/land cover, lake depth, water residence time, and elevation), leading to the question of which of these (if any) should be built into sampling designs? Some drivers are easier to include than others, particularly those that can be obtained from remotely sensed data (Hollister et al. 2016). Unfortunately, other lake features that affect water quality such as depth (e.g., Jeppesen et al. 2003; Taranu and Gregory-Eaves 2008; Read et al. 2015) or water residence time (e.g., Xenopoulos et al. 2003; Sobek et al. 2007) are sparsely quantified and difficult to predict with certainty (Hollister et al. 2011; Oliver et al. 2016). However, rather than exclude these sorts of attributes, alternative strategies may be possible, such as using surrogate metrics (e.g., ratios of watershed area to lake area as an indicator of water residence time; Xenopoulos et al. 2003; Sobek et al. 2007) or categorical approaches (e.g., high- and low-latitude lakes or shallow and deep lakes; Søndergaard et al. 2005; Hamill and Lew 2006).

The second lesson that can be gleaned from these analyses is that while integrated datasets such as LAGOS-NE are not derived entirely from probabilistic surveys, they can be used to identify sampling tendencies, and in some cases, may allow us to correct for these biases via statistical approaches such as resampling. That is, resampling large integrated datasets can, in some cases, provide a way to generate a representative set of observations with which to make statistical and ecological inferences.

Conclusion

As the availability of environmental datasets increases, we expect the recent trend of synthesizing and analyzing information on lakes and watersheds derived from a variety of sources to accelerate. This reality inspired our questions about patterns of data availability, and because it integrates multiple independent datasets, LAGOS-NE provided a convenient resource for addressing these questions. Nonetheless, the specific outcomes of our analyses were inevitably shaped by the geographic context of LAGOS-NE. Sampling tendencies in other countries—or even in other regions of the United States—are likely to differ in some fashion. Heterogeneity in data availability over both space and time in the LAGOS-NE region can be connected to responses to specific management challenges (in this case, eutrophication and acid rain), the prominence of citizen science programs, and/or a small number of programs for specific variables and long-term datasets. We expect these same tendencies to exist elsewhere, although they may be driven by other or additional management priorities or circumstances. It is encouraging that some regional- and national-level programs have expanded their scope over time to include a greater diversity of limnological variables rather than focusing on issue-specific parameters (e.g., Peck et al. 2013; Fölster et al. 2014). However, overcoming the limited amount, type, and geography of long-term data may be particularly challenging for future investigations. Inconsistencies in the availability of limnological data will inevitably persist regardless of the type of dataset (integrated or probabilistic), and biased sampling can lead to biased or spurious conclusions about limnological pattern and process at macroscales. Thus, a critical step in the use of broad-scale datasets, particularly those compiled from multiple sources that often cross multiple political or organizational boundaries, is to determine the what, where, and when of data being used to discover or examine limnological trends and patterns at long temporal and broader spatial scales.

References

Abell, J. M., D. Özkundakci, P. Hamilton, and S. D. Miller. 2011. Relationships between land use and nitrogen and phosphorus in New Zealand lakes. Mar. Freshw. Res. **62**: 162–175. doi:10.1071/MF10180

Behmel, S., M. Damour, R. Ludgwig, and M. J. Rodriguez. 2017. Water quality monitoring strategies—a review and future perspectives. Sci. Tot. Environ. **571**: 1312–1329. doi: 10.1016/j.scitotenv.2016.06.235

Bergström, A. K., and M. Jansson. 2006. Atmospheric nitrogen deposition has caused nitrogen enrichment and eutrophication of lakes in the northern hemisphere. Glob. Change Biol. **12**: 635–643. doi:10.1111/j.1365-2486.2006.01129.x

Bertilsson, S., and others. 2013. The under-ice microbiome of seasonally frozen lakes. Limnol. Oceanogr. **58**: 1998–2012. doi:10.4319/lo.2013.58.6.1998

Bigham Stephens, D. L., R. E. Carlson, C. A. Horsburgh, M. V. Hoyer, R. W. Bachmann, and D. E. Canfield Jr. 2015. Regional distribution of Secchi disk transparency in waters of the United States. Lake Reserv. Manage. **31**: 55–63. doi: 10.1080/10402381.2014.1001539

Boucher, J., K. C. Weathers, H. Norouzi, and B. Steele. 2018. Assessing the effectiveness of Landsat 8 chlorophyll a retrieval algorithms for regional freshwater monitoring. Ecol. Appl. **28**: 1044–1054. doi:10.1002/eap.1708

Cardoso, A. C., A. Solimini, G. Premazzi, L. Carvalho, A. Lyche, and S. Rekolainen. 2007. Phosphorus reference concentrations in European lakes. Hydrobiologia **584**: 3–12. doi:10.1007/s10750-007-0584-y

Carlson, R. E. 1977. 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

Correll, D. L. 1998. The role of phosphorus in the eutrophication of receiving waters: A review. J. Environ. Qual. **27**: 261–266. doi:10.2134/jeq1998.00472425002700020004x

DelSontro, T., J. J. Beaulieu, and J. A. Downing. 2018. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. Limnol. Oceanogr. Lett. **3**: 64–75. doi:10.1002/lol2.10073

Denfeld, B. A., P. Kortelainen, M. Rantakari, S. Sobek, and G. A. Weyhenmeyer. 2016. Regional variability and drivers of below ice CO_2 in boreal and subarctic lakes. Ecosystems **19**: 461–476. doi:10.1007/s10021-015-9944-z

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

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

Downing, J. A. 2010. Emerging role of small lakes and ponds: Little things mean a lot. Limnetica **29**: 9–24. doi:10.23818/limn.29.02

Du, E., W. de Vries, J. N. Galloway, X. Hu, and J. Fang. 2014. Changes in wet nitrogen deposition in the United States between 1985 and 2012. Environ. Res. Lett. **9**: 095004. doi: 10.1088/1748-9326/9/9/095004

Dugan, H. A., and others. 2017*a*. Salting our freshwater lakes. Proc. Natl. Acad. Sci. USA **114**: 4453–4458. doi:10.1073/pnas.1620211114

Dugan, H. A., and others. 2017*b*. Long-term chloride concentrations in North American and European freshwater lakes. Sci. Data **4**: 170101. doi:10.1038/sdata.2017.101

Estes, L., P. R. Elsen, T. Treuer, L. Ahmed, K. Caylor, J. Chang, J. J. Choi, and E. C. Ellis. 2018. The spatial and temporal domains of modern ecology. Nat. Ecol. Evol. **2**: 819–826. doi:10.1038/s41559-018-0524-4

Filstrup, C. T., T. Wagner, P. A. Soranno, E. H. Stanley, C. A. Stow, K. E. Webster, and J. A. Downing. 2014. Regional variability among nonlinear chlorophyll-phosphorus relationships in lakes. Limnol. Oceanogr. **59**: 1691–1703. doi: 10.4319/lo.2014.59.5.1691

Fölster, J., R. K. Johnson, M. N. Futter, and A. Wilander. 2014. The Swedish monitoring of surface waters: 50 years of adaptive monitoring. Ambio **43**: 3–18. doi:10.1007/s13280-014-0558-z

Gobler, C. J., J. M. Burkholder, T. W. Davis, M. J. Harke, T. Johengen, C. A. Stow, and D. B. Van de Waal. 2016. The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. Harmful Algae **54**: 87–97. doi:10.1016/j.hal.2016.01.010

Gray, D. K., S. E. Hampton, C. M. O'Reilly, S. Sharma, and R. S. Cohen. 2018. How do data collection and processing methods impact the accuracy of long-term trend estimation in lake surface-water temperatures? Limnol. Oceanogr. Methods **16**: 504–515. doi:10.1002/lom3.10262

Häkanson, L. 1992. Considerations on representative water quality data. Int. Revue ges. Hydrobiol. **77**: 497–505. doi: 10.1002/iroh.19920770312

Hamill, K., and D. Lew. 2006. Snapshot of lake water quality in New Zealand. New Zealand Ministry for the Environment, Available from http://www.mfe.govt.nz/publications/fresh-water-environmental-reporting/snapshot-lake-water-quality-new-zealand. Last accessed on 30 Jan 2019.

Hampton, S. E., and others. 2017. Ecology under lake ice. Ecol. Lett. **20**: 98–111. doi:10.1111/ele.12699

Hanson, P. C., S. R. Carpenter, J. A. Cardille, M. T. Coe, and L. A. Winslow. 2007. Small lakes dominate a random sample of regional lake characteristics. Freshwat. Biol. **52**: 814–822. doi:10.1111/j.1365-2427.2007.01730.x

Hartmann, J., R. Lauerwald, and N. Moosdorf. 2014. A brief overview of the GLObal River CHemistry database, GLORICH. Proc. Earth Planet. Sci. **10**: 23–27. doi:10.1016/j.proeps.2014.08.005

Heffernan, J. B., and others. 2014. Macrosystems ecology: Understanding ecological patterns and processes at continental scales. Front. Ecol. Environ. **12**: 5–14. doi:10.1890/130017

Henson, S. A., C. Beaulieu, and R. Lampitt. 2016. Observing climate change trends in ocean biogeochemistry: When and where. Glob. Change Biol. **22**: 1561–1571. doi:10.1111/gcb.13152

Holgerson, M. A., and P. A. Raymond. 2016. Large contribution to inland water CO_2 and CH_4 emissions from very small ponds. Nat. Geosci. **9**: 222–226. doi:10.1038/NGEO2654

Hollister, J. W., W. B. Milstead, and M. A. Urrutia. 2011. Predicting maximum lake depth from surrounding topography. PLoS One **6**: e25764. doi:10.1371/journal.pone.0025764

Hollister, J. W., W. B. Milstead, and B. J. Kreakie. 2016. Modeling lake trophic state: A random forest approach. Ecosphere **7**: e01321. doi:10.1002/ecs2.1321

Hughes, B. B., and others. 2017. Long-term studies contribute disproportionately to ecology and policy. Bioscience **67**: 271–281. doi:10.1093/biosci/biw185

Hughes, R. M., and D. V. Peck. 2008. Acquiring data for large aquatic resource surveys: The art of compromise among science, logistics, and reality. J. N. Am. Benthol. Soc. **27**: 837–859. doi:10.1899/08-053.1

Huser, B. J., M. N. Futter, R. Wang, and J. Fölster. 2018. Persistent and widespread long-term phosphorus declines in boreal lakes in Sweden. Sci. Total Environ. **613–614**: 240–249. doi:10.1016/j.scitotenv.2017.09.067

Jeppesen, E., and others. 2003. The impact of nutrient state and lake depth on top-down control in the pelagic zone of lakes: A study of 466 lakes from the temperate zone to the Arctic. Ecosystems **6**: 313–325. doi:10.1007/s10021-002-0145-1

Kankaala, P., J. Huotari, T. Tulonen, and A. Ojala. 2013. Lake-size dependent physical forcing drives carbon dioxide and methane effluxes from lakes in a boreal landscape. Limnol. Oceanogr. **58**: 1915–1930. doi:10.4319/lo.2013.58.6.1915

Knowlton, M. F., and J. R. Jones. 2006. Temporal variation and assessment of trophic state indicators in Missouri reservoirs: Implications for lake monitoring and management. Lake Reserv. Manage. **22**: 261–271. doi:10.1080/07438140609353904

Lapierre, J. F., D. A. Seekell, C. T. Filstrup, S. M. Collins, C. E. Fergus, P. A. Soranno, and K. S. Cheruvelil. 2017. Continental-scale variation in controls of summer CO_2 in United States lakes. J. Geophys. Res. Biogeosci. **122**: 875–885. doi: 10.1002/2016JG003525

Lindenmayer, D. B., G. E. Likens, C. J. Krebs, and R. J. Hobbs. 2010. Improved probability of detection of ecological "surprises". Proc. Natl. Acad. Sci. USA **107**: 21957–21962. doi: 10.1073/pnas.1015696107

Lottig, N. R., and S. R. Carpenter. 2012. Interpolating and forecasting lake characteristics using long-term monitoring data. Limnol. Oceanogr. **57**: 1113–1125. doi:10.4319/lo.2012.57.4.1113

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

McCrackin, M. L., H. P. Jones, P. C. Jones, and D. Moreno-Mateos. 2017. Recovery of lakes and coastal marine ecosystems from eutrophication: A global meta-analysis. Limnol. Oceanogr. **62**: 507–518. doi:10.1002/lno.10441

Mendonça, R., R. A. Müller, D. Clow, C. Verpoorter, P. Raymond, L. J. Tranvik, and S. Sobek. 2017. Organic carbon burial in global lakes and reservoirs. Nat. Commun. **8**: 1694. doi:10.1038/s41467-017-01789-6

National Research Council. 2004. Confronting the nation's water problems: The role of research. National Academies Press.

Obrecht, D. V., M. Milanick, and B. D. Perkins. 1998. Evaluation of data generated from lake samples collected by volunteers. Lake Reserv. Manag. **14**: 21–27. doi:10.1080/07438149809354106

Oliver, S. K., and others. 2016. Prediction of lake depth across a 17-state region in the U.S. Inland Waters **6**: 314–324. doi: 10.5268/IW-6.3.957

Oliver, S. K., S. M. Collins, P. A. Soranno, T. Wagner, E. H. Stanley, J. R. Jones, C. A. Stow, and N. R. Lottig. 2017. Unexpected stasis in a changing world: Lake nutrient and chlorophyll trends since 1990. Glob. Change Biol. **23**: 5455–5467. doi:10.1111/gcb.13810

O'Reilly, C. M., and others. 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys. Res. Lett. **42**: 10773–10781. doi:10.1002/2015GL066235

Paulsen, S. G., R. M. Hughes, and D. P. Larsen. 1998. Critical elements in describing and understanding our nation's aquatic resources. J. Am. Water Resour. Assoc. **34**: 995–1005. doi:10.1111/j.1752-1688.1998.tb04148.x

Peck, D. V., A. R. Olsen, M. H. Weber, C. Peterson, and S. M. Holdsworth. 2013. Survey design and extent estimates for the National Lakes Assessment. Freshwat. Sci. **32**: 1231–1245. doi:10.1899/11-075.1

Peterson, S. A., N. S. Urquhart, and E. B. Welch. 1999. Sample representativeness: A must for reliable regional lake estimates. Environ. Sci. Technol. **33**: 1559–1565. doi: 10.1021/es980711l

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Available from https://www.R-project.org/. Last accessed on 30 Jan 2019.

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

Read, E. K., and others. 2017. Water quality data for national-scale aquatic research: The water quality portal. Water Resour. Res. **53**: 1735–1745. doi:10.1002/2016WR019993

Schindler, D. W. 1974. Eutrophication and recovery in experimental lakes: Implications for lake management. Science **184**: 897–899. doi:10.1126/science.195.4275.260

Seekell, D. A., J. F. Lapierre, and K. S. Cheruvelil. 2018. A geography of lake carbon cycling. Limnol. Oceanogr. Lett. **3**: 49–56. doi:10.1002/lol2.10078

Sharma, S., and others. 2015. Globally distributed lake surface water temperatures collected in situ and by satellites; 1985-2009. Sci. Data **2**: 150008. doi:10.1038/sdata.2015.8

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

Sobek, S., L. J. Tranvik, Y. T. Prairie, P. Kortelainen, and J. J. Cole. 2007. Patterns and regulation of dissolved organic carbon: An analysis of 7,500 widely distributed lakes. Limnol. Oceanogr. **52**: 1208–1219. doi:10.4319/lo.2007.52.3.1208

Søndergaard, M., E. Jeppesen, J. P. Jensen, and S. L. Amsinck. 2005. Water framework directive: Ecological classification of Danish lakes. J. Appl. Ecol. **42**: 616–629. doi:10.1111/j.1365-2664.2005.01040.x

Søndergaard, M., T. L. Lauridsen, L. S. Johansson, and E. Jeppesen. 2017. Nitrogen or phosphorus limitation in lakes and its impact on phytoplankton and submerged macrophyte cover. Hydrobiologia **795**: 35–48. doi:10.1007/s10750-017-3110-x

Soranno P. A., and K. 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. doi:10.6073/pasta/b1b93ccf3354a7471b93eccca484d506

Soranno P. A., and K. Cheruvelil. 2017*b*. LAGOS-NE-LOCUS v1.01: 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. doi:10.6073/pasta/940b25d022c695b440e1bdbc49fbb77b

Soranno P. A., and K. Cheruvelil. 2017*c*. 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. doi:10.6073/pasta/b88943d10c6c5c480d5230c8890b74a8

Soranno, P. A., K. S. Cheruvelil, K. E. Webster, M. T. Bremigan, T. Wagner, and C. A. Stow. 2010. Using landscape limnology to classify freshwater ecosystems for multi-ecosystem management and conservation. Bioscience **60**: 440–454. doi:10.1525/bio.2010.60.6.8

Soranno, P. A., and others. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science through data reuse. GigaScience **4**: 28. doi:10.1186/s13742-015-0067-4

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 U.S. lakes. GigaScience **6**: 1–22. doi:10.1093/gigascience/gix101

Sprague, L. A., G. P. Oelsner, and D. M. Argue. 2017. Challenges with secondary use of multi-source water-quality data in the United States. Water Res. **110**: 252–261. doi:10.1016/j.watres.2016.12.024

Stachelek J., S.K. Oliver, and F. Masrour. 2017. LAGOS: R interface to the LAke multi-scaled GeOSpatial & temporal database. R package version 1.0.0. Available from https://CRAN.R-project.org/package=LAGOSNE. Last accessed on 30 Jan 2019.

Stanley, E. H., S. M. Collins, N. R. Lottig, S. K. Oliver, K. E. Webster, K. S. Cheruvelil, and P. A. Soranno. 2019. LAGOS: Lake nutrient, carbon and chlorophyll data to evaluate biases in lake water quality sampling practices in a 17-state region of the US. Environmental Data Initiative. doi:10.6073/pasta/ad2516f5a98df32f1a8cbd7e658c088f

Stich, H. B., and A. Brinker. 2005. Less is better: Uncorrected versus phaeopigment-corrected photometric chlorophyll-a estimation. Arch. Hydrobiol. **162**: 111–120. doi:10.1127/0003-9136/2005/0162-0111

Taranu, Z. E., and I. Gregory-Eaves. 2008. Quantifying relationships among phosphorus, agriculture, and lake depth at an inter-regional scale. Ecosystems **11**: 715–725. doi:10.1007/s10021-008-9153-0

Thornton, B. F., M. Wik, and P. M. Crill. 2016. Double-counting challenges the accuracy of high-latitude methane inventories. Geophys. Res. Lett. **43**: 12569–12577. doi:10.1002/2016GL071772

US EPA. 2016. National Lakes Assessment 2012. EPA 841-R-16-113. U.S. Environmental Protection Agency, Available from https://www.epa.gov/sites/production/files/2016-12/documents/nla_report_dec_2016.pdf. Last accessed on 30 Jan 2019.

Vallentyne, J. R. 1969. Definition of a limnologist. Limnol. Oceanogr. **14**: 815. doi:10.4319/lo.1969.14.5.0815

Wagner, T., P. A. Soranno, K. S. Cheruvelil, W. H. Renwick, K. E. Webster, P. Vaux, and R. J. F. Abbitt. 2008. Quantifying sample bias of inland lake sampling programs in relation to lake surface area and land cover. Environ. Monit. Assess. **141**: 131–147. doi:10.1007/s10661-007-9883-z

Xenopoulos, M. A., D. M. Lodge, J. Frentress, T. A. Kreps, S. D. Bridgham, E. Grossman, and C. J. Jackson. 2003. Regional comparisons of watershed determinants of dissolved organic carbon in temperate lakes from the upper Great Lakes region and selected regions globally. Limnol. Oceanogr. **48**: 2321–2334. doi:10.4319/lo.2003.48.6.2321

Acknowledgments

We are grateful to all individuals and groups who provided data and input to the LAGOS initiative. Special thanks to Melanie Cofrin at the New Hampshire Department of Environmental Services for assistance and provision of updated data. We appreciate the thoughtful and constructive input provided by two anonymous reviewers. Funding was provided by the National Science Foundation Macrosystems Biology Program through a collaborative grant to M.S.U. (EF-1638679) and U.W. (EF-1685534); S.M.C. also received support from DBI-1401954.

Conflict of Interest

None declared.

Submitted 04 November 2018
Revised 10 January 2019
Accepted 14 January 2019

Associate editor: Paul Frost