LIMNOLOGY
and
OCEANOGRAPHY

Limnol. Oceanogr. 68, 2023, 2759–2773

doi: 10.1002/lno.12457

A continuous classification of the 476,697 lakes of the conterminous US based on geographic archetypes

Jean-Francois Lapierre,[1,2*] Katherine E. Webster,[3,4] Ephraim M. Hanks,[5] Tyler Wagner,[5,6] Patricia A. Soranno,[3] Ian M. McCullough,[3] Kaitlin L. Reinl,[7] Marcella Domka,[3] Noah R. Lotting[8]

[1]Département de Sciences Biologiques, Faculté des Arts et Sciences, Université de Montréal, Montreal, Quebec, Canada
[2]Groupe de Recherche Interuniversitaire en Limnologie (GRIL), Montreal, Quebec, Canada
[3]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan, USA
[4]Center for Limnology Hasler Laboratory of Limnology, University of Wisconsin Madison, Madison, Wisconsin, USA
[5]Department of Statistics, The Pennsylvania State University, Pennsylvania, USA
[6]U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, Pennsylvania, USA
[7]University of Wisconsin-Madison Division of Extension Natural Resources Institute, Lake Superior National Estuarine Research Reserve, Madison, Wisconsin, USA
[8]Center for Limnology Trout Lake Station, University of Wisconsin-Madison, Madison, Wisconsin, USA

Abstract

A variety of classification approaches are used to facilitate understanding, prediction, monitoring, and the management of lakes. However, broad-scale applicability of current approaches is limited by either the need for in situ lake data, incompatibilities among approaches, or a lack of empirical testing of approaches based on ex situ data. We developed a new geographic classification approach for 476,697 lakes ≥ 1 ha in the conterminous U.S. based on lake archetypes representing end members along gradients of multiple geographic features. We identified seven lake archetypes with distinct combinations of climate, hydrologic, geologic, topographic, and morphometric properties. Individual lakes were assigned weights for each of the seven archetypes such that groups of lakes with similar combinations of archetype weights tended to cluster spatially (although not strictly contiguous) and to have similar limnological properties (e.g., concentrations of nutrients, chlorophyll *a* (Chl *a*), and dissolved organic carbon). Further, archetype lake classification improved commonly measured limnological relationships (e.g., between nutrients and Chl *a*) compared to a global model; a discrete archetype classification slightly outperformed an ecoregion classification; and considering lakes as continuous mixtures of archetypes in a more complex model further improved fit. Overall, archetype classification of US lakes as continuous mixtures of geographic features improved understanding and prediction of lake responses to limnological drivers and should help researchers and managers better characterize and forecast lake states and responses to environmental change.

There is a long history of classifying aquatic ecosystems into groups based on their physical, ecological, or geographic properties, sharing the common goals of improving understanding, prediction, or management (Hutchison et al. 1958;

*Correspondence: jean-francois.lapierre.1@umontreal.ca

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

Author Contribution Statement: J.F.L., P.A.S., K.E.W., E.H., I.M.M., T.W., and I.M.M. conceived the idea for the manuscript. J.F.L., K.L.R., T.W., K.E.W., M.D., E.H., and P.A.S. provided critical interpretation. J.F.L., P.A.S., and K.E.W. conducted the literature review with help from E.H. and T.W. N.L. and K.E.W. compiled and synthesized the data. E.H., K.E.W., T.W., and K.L.R. performed data analysis. K.E.W., I.M.M., T.W., and M.D. drafted figures and tables. J.F.L. wrote the first draft of the introduction, results and discussion. K.E.W. and E.H. wrote the first draft of the methods. M.D. and T.W. contributed significant amounts of text in the introduction, methods and results. P.A.S. and K.E.W. contributed significant amounts of text in the introduction and discussion. Authors are listed in decreasing order of contribution. All authors reviewed and commented on the text.

Heiskary et al. 1987; Emmons et al. 1999; Phillips et al. 2008; Dodds et al. 2019). Classification systems are diverse in their structure and aims, with some of them focusing on in situ observations of temperature, water chemistry, or biodiversity of the ecosystem studied, while others have focused on the physical or geographic properties of the aquatic ecosystem itself or its surrounding terrestrial landscape. Using these approaches, researchers have identified regions or types of lakes with (1) similar ecosystem properties such as biodiversity, nutrient concentrations, or alkalinity (Emmons et al. 1999; Phillips et al. 2008; Dodds et al. 2019; Lemm et al. 2021), or (2) similar driver-response relationships, as exemplified by stronger limnological relationships between concentrations of total Nitrogen (TN) and total Phosphorus (TP), or between algal biomass and nutrients within rather than among regions or types (Yuan et al. 2014; Poikane et al. 2022). Classification approaches have thus been successful in categorizing lakes into meaningful and useful groups, providing a predictive framework for understanding limnological properties and responses among broadly distributed ecosystems.

An implicit assumption of classification approaches, beyond grouping lakes with similar properties and limnological responses, is that it is more practical to use a limited number of intuitive classes than a large number of predictive variables. While all classification approaches essentially achieve this variable reduction, a lack of consensus on methods and approaches, perhaps explained by a combination of variability in spatial and temporal scales, project objectives, and data availability, has complicated efforts to make and compare inferences across different classification systems. For example, European countries have developed tens of lake classification systems, often at the national level, resulting in hundreds of differently defined lake types under the European Water Framework Directive (Poikane et al. 2022).

Furthermore, classification systems can be discrete and rely on a smaller number of in situ variables available for already sampled lakes such as trophic status (Carlson 1977; Wetzel 2001), stratification (Lewis Jr 1983), or alkalinity (Kelly et al. 2012; Solheim et al. 2019; Poikane et al. 2022), but cannot be readily applied to lakes for which the classifying variables are not available. Other approaches have used a larger number of ex situ variables for all lakes within a region (e.g., topographic or geological variables available from geospatial datasets; Hill et al. 2018; Solheim et al. 2019), but empirical demonstrations of the ability of these classifications to describe lake functioning remain rare. A common disadvantage of both traditional in situ and ex situ lake classification systems is that their distinct boundaries ignore potentially large within-class variability in ecologically relevant lake and watershed properties. In particular, spatially contiguous classifications overlook fine-scale spatial heterogeneity in important lake features such as area or depth that may be poorly correlated with coarser, regional variables (Lapierre et al. 2018). Thus, there is a need for an alternative classification approach that treats lakes as a mixture of lake-focused properties, their surrounding watersheds, as well as the regions in which they are located. Such an approach could better capture macroscale patterns in lakes and fine-scale heterogeneity within broader regions simultaneously, be reproducible, and be applicable to the complete population of lakes of a given area, including unsampled lakes.

Archetype modeling approaches provide a potentially valuable method to classify lakes that has not been explored to date. Traditional cluster analyses, like k-means clustering (Steinley 2006), sort lakes into a limited number of groups, with individual lakes classified within certain ranges in terms of in-situ measurements, or clustered together around a centroid value in a multivariate analysis (Fig. 1a). In contrast, archetype classification approaches represent individual entities (here, lakes) by a weighted average of multiple archetypes representing gradients of input characteristic values. Archetype classification approaches have been used in sustainability research (Sietz et al. 2019) to identify contrasting endpoint scenarios for development or natural resources use, and to evaluate the ecological outcomes for different combinations of archetypical scenarios (Eisenack et al. 2019; Harrison et al. 2019). In the context of lake classification, this approach

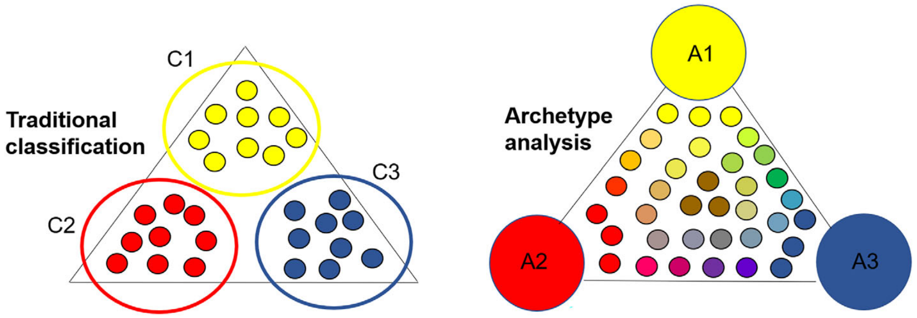

Fig. 1. Conceptual depiction of a traditional, centroid based clustering approach (left) vs. archetype analyses that classify each lake (small circles within the triangle) as a continuous mixture of lake and catchment properties along a gradient (right). Each lake within a class in the left panel is discretely assigned to the same cluster (C1, 2, or 3) of the same color, whereas each lake in the right panel has its own color, that is, a weighted combination of characteristics resembling the yellow, red, and blue archetypes (A1, 2, and 3, respectively).

allows one to identify the most contrasting combinations of lake and catchment properties (i.e., lake geographic archetypes) and to generate a specific linear combination of archetype weights that best characterizes each lake (Fig. 1b).

We developed a classification of 476,697 conterminous US lakes ≥ 1 ha (excluding the Great Lakes) as continuous mixtures of geographic archetypes based on a combination of climate, landscape, and hydrological features considered to be unrelated to human influence but known to be important for lake functioning, and specifically related to nutrient transport. By explicitly accounting for the multi-scale drivers of lake ecosystem properties, our approach has the potential to improve existing lake classification systems both by preserving nuances among individual ecosystems and by acting as a robust and flexible framework for prediction and management in a global change context. Our objectives were to: (1) classify all lakes into an optimal number of archetypes that captures variation in the geographic features known to influence lake functioning, (2) describe the geographic distribution of these archetypes across the contiguous US, and (3) determine to what extent the archetypes capture lakes with similar limnological states and responses to environmental drivers.

Methods

Data description

To meet our objective to define lake archetypes across wide gradients of lake-specific and landscape context features, we used the census population of 476,697 lakes ≥ 1 ha within the conterminous US defined in the LOCUS v1.0 module of the LAGOS-US platform (Cheruvelil et al. 2021). LAGOS-US is an extensible modular data platform with a core module, LOCUS that includes data tables with geospatial characteristics of lakes and their watersheds, including geometry and location. The linked modules on ecological context (GEO v1.0, Smith et al. 2022; https://portal.edirepository.org/nis/mapbrowse?packageid=edi.1136.3) and lake maximum depth (DEPTH v1.0, Stachelek et al. 2021; https://portal.edirepository.org/nis/mapbrowse?packageid=edi.1043.1), provided the predictor variables and the majority of the post hoc variables needed to meet the study objectives. Lake water quality data for limnological response modeling were obtained from the US Environmental Protection Agency's 2012 and 2017 National Lakes Assessment (NLA; USEPA 2016, 2022). Variables used as predictors in the archetype analysis, in post-hoc data exploration and in response modeling, are defined in Table 1; a summary of values for the study lake population is provided in Supplementary Table S1.

The 16 predictor variables (Table 1) used to generate the archetypes capture hydrology, soil properties, and terrain features that collectively influence hydrologic flowpaths to lakes, as well as other key ecological context variables known to influence lake nutrient status and productivity. These include long-term climatic regimes and "natural" land cover characteristics (forest and wetland). Variables were selected to encompass a balanced number of predictors among "climate," "hydrology," "geology," "terrain," and "land cover" groups while limiting redundant information among variables as determined using a principal component analysis. We did not include lake depth or area in the set of predictor variables because we wanted to create archetypes based on landscape setting rather than in-lake features with little to no spatial structure across the study extent (Lapierre et al. 2018). Following this data reduction exercise, we went from 37 candidate variables to 16 variables for modeling; these predictor variables exhibited wide ranges across the study lakes (Supplementary Table S1). Finally, we examined distributions of each potential predictor variable to identify the appropriate data transformation (Supplementary Table S1), which was applied prior to statistical analyses described below.

Modeling and characterizing Lake archetypes

We modeled among-lake variation in the characteristics described above using archetype analysis (Cutler and Breiman 1994; Oberlack et al. 2019). Archetypes represent extreme combinations of data values such that each entity (i.e., a lake) can be represented by weighted mixtures of the archetypes. The number of archetypes, and the mixture of weights for individual archetypes are computed to minimize root squared error of the observations within the multivariate space (Cutler and Breiman 1994). The approach thus preserves the individual characteristics of each lake as well as representing lakes as located across continuous gradients defined by each archetype. In other words, using a discrete classification, a group of lakes would be assigned the dominant or majority feature of the group (i.e., lakes with a watershed that is 55% agricultural and 45% forested would be labeled as agricultural), whereas archetype analysis represents each lake as a weighted mixture of the characteristics defining the different archetypes.

Specifically, let x_i be a vector of m characteristics for the lake i

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$$

The formulation of the archetypal analysis model we use considers the observed lake characteristics x_i as being normally distributed with mean expected x_i ($E(x_i)$)

$$E(x_i) = \sum\nolimits_{k=1}^{K} w_{ik} a_k,$$

where a_k is a linear combination of the lake characteristics corresponding to the k-th archetype across all lakes, and w_{ik} is the weight of the k-th archetype associated with the i-th lake. The weights $\{w_{ik}\}$ are estimated from the data, as are the archetypes, and the archetypes

Table 1. Description and sources of lake, landscape, hydrological, geological, land cover, and climatic variables used as predictors for developing the archetypes, post-hoc exploratory analysis, and limnological modeling. Features were assessed at the lake, watershed or the HU12 (e.g., 12-digit hydrologic unit of the USGS Watershed Boundary Dataset) scales. See Supplementary Table S1 for summary statistics and data source.

Analysis role	Variable name	Variable description	Units	Reference
Background	Latitude	Latitude of central point of the lake polygon in decimal degrees; NAD83 projection	Decimal degrees	USGS (2017*a*)
Background	Longitude	Longitude of central point of the lake polygon in decimal degrees; NAD83 projection	Decimal degrees	USGS (2017*a*)
Predictor	Elevation	The elevation of the lake polygon central point	Meters	USGS (2017*b*)
Predictor	Topographic wetness	An index of topographic control on hydrologic processes with high values reflecting flat terrain	None	USGS/EROS (2003)
Predictor	Topographic roughness	An index of terrain ruggedness equal to the absolute difference in meters between the elevation of the focal cell and its immediate neighbors; high values indicate more complex terrain	Meters	USGS (2017*c*)
Predictor	Baseflow index	Mean within the HU12 of the percentage of streamflow that can be attributed to groundwater discharge into streams	Percent	Wolock (2003)
Predictor	Annual runoff	Mean within the HU12 of annual runoff	Mm yr^{-1}	Gebert et al. (1987)
Predictor	Watershed : Lake area	Ratio between watershed area and lake water area	None	USGS (2019*a*)
Predictor	Stream density	Density of streams within the watershed	m ha^{-1}	USGS (2021)
Predictor	Annual precipitation	Mean value at the lake central point for the mean annual total precipitation from 1981 to 2010	Mm yr^{-1}	PRISM (2019)
Predictor	Annual temperature	Mean value at the lake central point for the mean annual temperature from 1981 to 2010	°C	PRISM (2019)
Predictor	Shrub land cover	Percent of the watershed classified as shrub and scrub in 2016	Percent	USGS (2019*b*)
Predictor	Forest land cover	Percent of the watershed classified in 2016 as forest; the sum of coniferous, deciduous and mixed forest	Percent	USGS (2019*b*)
Predictor	Wetland land cover	Percent of the watershed classified in 2016 as wetland; the sum of woody and emergent herbaceous wetlands	Percent	USGS (2019*b*)
Predictor	Depth to bedrock	Average absolute depth to bedrock within the watershed	m	Hengl et al. (2017)
Predictor	Soil erodibility	Average soil erodibility factor, not adjusted for the effect of rock fragments	None	Miller and White (1998)
Predictor	Sandy soils	Average percentage mass fraction of sand, 50 to 200 μm, in the 0 to 5 cm depth soil layer	Percent	Hengl et al. (2017)
Predictor	Silty soils	Average percentage mass fraction of silt, 2 to 50 μm, in the 0 to 5 cm depth soil layer	Percent	Hengl et al. (2017)
Post hoc exploratory	Maximum depth	Maximum depth of the lake	m	Stachelek et al. (2021)
Post hoc exploratory	Shoreline development factor (SDF)	a measure of lake shoreline complexity; calculated as the lake perimeter (m) divided by the product of 2 times the square root of pi times lake water area (ha)	None	USGS (2017*a*)
Post hoc exploratory	Lake area	Surface area of lake waterbody polygon from NHD comprised of open water; islands are excluded	Ha	USGS (2017*a*)

(Continues)

Table 1. Continued

Analysis role	Variable name	Variable description	Units	Reference
Post hoc exploratory	Cultivated crop land use	Percent of the watershed classified in 2016 as cultivated crops	Percent	USGS (2019*b*)
Post hoc exploratory	Grassland land cover	Percent of the watershed classified in 2016 as grassland or herbaceous	Percent	USGS (2019*b*)
Post hoc exploratory	Pasture land use	Percent of the watershed classified in 2016 as pasture and hay	Percent	USGS (2019*b*)
Post hoc exploratory	Agricultural land use	Percent of the watershed classified in 2016 as agriculture; the sum of pasture and cultivated crop land use classes	Percent	USGS (2019*b*)
Post hoc exploratory	Development land use	Percent of the watershed classified in 2016 as developed; the sum of open, low, medium and high development land use classes	Percent	USGS (2019*b*)
Post hoc exploratory	Total deposition N	Mean annual total deposition to the watershed of nitrogen during 2010	Kg ha^{-1}	NADP (2022)
Post hoc exploratory	Total deposition S	Mean annual total deposition of sulfur to the watershed during 2010	Kg ha^{-1}	NADP (2022)
Post hoc exploratory	Clay soils	Average percentage mass fraction of clay, 0 to 2 μm, in the 0 to 5 cm depth soil layer	Percent	Hengl et al. (2017)
Post hoc exploratory	Coarse soils	Average percentage by volume of coarse fragments in the 0 to 5 cm soil depth	Percent	Hengl et al. (2017)
Post hoc exploratory	Soil organic C	Average organic carbon content, fine earth fraction, in the 0 to 5 cm soil layer	g kg^{-1}	Hengl et al. (2017)
Limological response	Chl*a*	Chlorophyll *a* (Chl *a*) concentration	μg L^{-1}	USEPA (2016, 2022)
Limological response	DOC	Dissolved organic carbon concentration	mg L^{-1}	USEPA (2016, 2022)
Limological response	TN	Total nitrogen concentration	μg L^{-1}	USEPA (2016, 2022)
Limological response	TP	Total phosphorus concentration	μg L^{-1}	USEPA (2016, 2022)
Limological response	Secchi	Secchi disk transparency	m	USEPA (2016, 2022)

$$\{a_k, k = 1, \ldots, K\}$$

are constrained to be linear combinations of the n lake characteristics, with each

$$a_k = \sum\nolimits_{i=1}^{n} c_{ki} x_i$$

for some set of $\{c_{ki}\}$. The weights for each lake sum to one

$$\sum\nolimits_{k=1}^{K} w_{ik} = 1.$$

Thus, archetype analysis finds the optimal set of *K* archetypes and optimal representation of each lake as a weighted average of these *K* archetypes. Each estimated archetype is a weighted average of observed lake and ecological context variables and can be interpreted as a representative combination of lake and ecological context variables.

The estimated archetype weights for each lake $\{w_{ik}\}$ represent the model fit for each lake in the dataset, with each lake represented by a weight for all estimated archetypes. As the weights sum up to 1 for each lake, a lake that is very similar to one archetype will have a very high weight (w_{ik} close to 1) in that archetype and low weights (close to zero) in all other archetypes. We fit the archetypal analysis model using the "archetypes" package (Eugster and Leisch 2009) in the R statistical computing environment (R Core Team 2021). We fit archetype models across a range of 2–12 archetypes. To minimize the impact of outliers, we used absolute error (rather than squared error) as our criteria for model fit. Using a screeplot, we chose seven archetypes as the best fit, as the improvement in model fit (as represented by absolute error) was steep from 2 to 7 and leveled off from 7 to 12.

We then examined patterns in lake-specific characteristics, watershed land use, and other features not used in the classification. These include lake morphometry variables such as water area and SDF from LAGOS-US LOCUS (Cheruvelil et al. 2021; Smith et al. 2021); lake maximum depth for a small subset of lakes from the LAGOS-US DEPTH module (Stachelek et al. 2021); human influenced land use variables (agriculture and development), TN, and sulfur (S) atmospheric deposition in 2010 and soil-related features from LAGOS-US GEO (Smith et al. 2022); and in situ measurements of lake water quality from the NLA data collected in 2012 and 2017 (USEPA 2016, 2022).

Prior to analysis, predictor and response variables were transformed (Supplementary Table S1) and, in the case of archetype analysis, standardized to have a mean of zero and standard deviation of one. After merging, cases with missing rows were deleted, providing a total of 476,697 lakes in the final analysis. Post-hoc variables were also transformed prior to analysis.

Predictive model fitting

To examine the usefulness of archetype weights in predictive models, we included them in models of the relationships between pairs of the limnological variables TN, TP, Chl *a*, and Secchi depth. In each case, we fit linear models between pairs of these variables, with four alternative models explaining variation in the relationships. First, we considered a global model ("Global," Table 2), in which the mean of one variable (e.g., Chl*a*) was assumed to be a linear function of another variable (e.g., TN). Second, we considered a hierarchical model ("Ecoregion," Table 2) where the linear relationship could change based on ecoregion membership in the National Aquatic Resource Survey 9-Level Ecosystems (Herlihy et al. 2008), with each ecoregion having its own distinct linear relationship. Third, we considered a similar model, but instead of using ecoregion, we used a categorical variable denoting the archetype most resembling each lake ("Max Archetype," Table 2). Finally, we considered a model in which each lake's linear relationship is a weighted average of the seven archetypes (Weighted Archetypes, Table 2). This fourth model is equivalent to a model with interactions between the seven estimated lake archetype weights and the predictor variable (e.g., TN). All models were fit in R using the "lm" function and model fits were compared using R-squared and AIC.

Results

The seven lake archetypes of the conterminous US

The following are descriptions of the predictors and post hoc variables investigated for the seven archetypes we identified for lakes in the conterminous US. We emphasize that these general descriptions do not represent distinct lake classes nor geographically contiguous lake ecoregions, but rather represent characteristics of lakes that received high weights (i.e., $> = 0.75$) from a particular archetype. Of all 476,697 lakes in the analysis, 10.4% received high weights for a given archetype. Based on the highly weighted lakes, we provide a synthetic interpretation of the archetypes below.

Archetype 1

Lakes with high weight in Archetype 1 ($n = 18{,}007$) are found in the sandiest soils and among the hottest, most humid, relatively flat, and low elevation areas. Their watersheds have average baseflow, higher runoff, and typically low stream density, depth to bedrock, silty soils, soil erodibility, and topographic roughness. They have an average watershed to lake area ratio and average coverage of forests and wetlands in their local watershed (Fig. 2). They are most common in the southeastern US, being particularly dominant in Florida (Fig. 3). They are not distinguishable from other archetypes in terms of the post-hoc variables lake area, SDF or depth, with lower TP but average Secchi depth, concentrations of TN, Chl*a*, DOC, soil organic C, and coarse soil material. They tend to have average agricultural land use (low cultivated but somewhat higher pasture). Lakes in this archetype tend toward higher proportions of developed areas and receive among the highest atmospheric deposition of N and S (Fig. 4).

Archetype 2

Lakes with high weight in Archetype 2 ($n = 2302$) are found in moderate temperature and high precipitation areas. These lakes are distinguished from other archetypes based on the high percentage of silty soils and soil erodibility of their watersheds. Watersheds are also characterized by low baseflow and stream density, high runoff, and the lowest watershed to lake area ratios of all archetypes and are typically found in low-elevation areas with high topographic roughness indicating complex terrain. Watersheds have high forest but low wetland land cover (Fig. 2). The lakes are mostly found along the central part of the conterminous US, mainly along the valleys of the Mississippi and Ohio Rivers (Fig. 3). The lakes tend to be smaller and shallower with low TP, DOC, and TN but high Chl*a* and low clarity. Their watersheds have average agricultural land use but low cropland and receive high atmospheric deposition of N and the highest S deposition. Developed land in these watersheds was overall average and intermediate between the higher values of archetypes 1 and 5 and lower values of the remaining archetypes (Fig. 4).

Archetype 3

Lakes with high weight in Archetype 3 ($n = 2956$) are distinguished by their very high elevation and cold temperatures, high precipitation, very high baseflow index, annual runoff, and stream density, and have the most rugged terrain as indicated by extremely high topographic roughness of their watersheds; in other words, they are typically cold-climate mountain lakes in hydrologically active and highly forested areas. Soils tend toward higher percent silt and average sand, with average depth to bedrock and erodibility. Land cover consists of very high percent forest with moderate shrub but very low wetland (Fig. 2). Lakes are mostly found in high elevation areas along the eastern and western mountain chains of the conterminous US, as well as in wet and cold northern areas of the Midwestern US (Fig. 3). These are among the deepest lakes in the dataset, have the lowest concentrations of TN, and Chl*a*, low TP and DOC, and are the most transparent of any archetype. Human-influenced land use of either agriculture or development is very low, with average grassland, like the high elevation lakes resembling archetype 7. Their watershed soils have the highest concentrations of soil organic carbon and proportions of coarse soils, and they receive

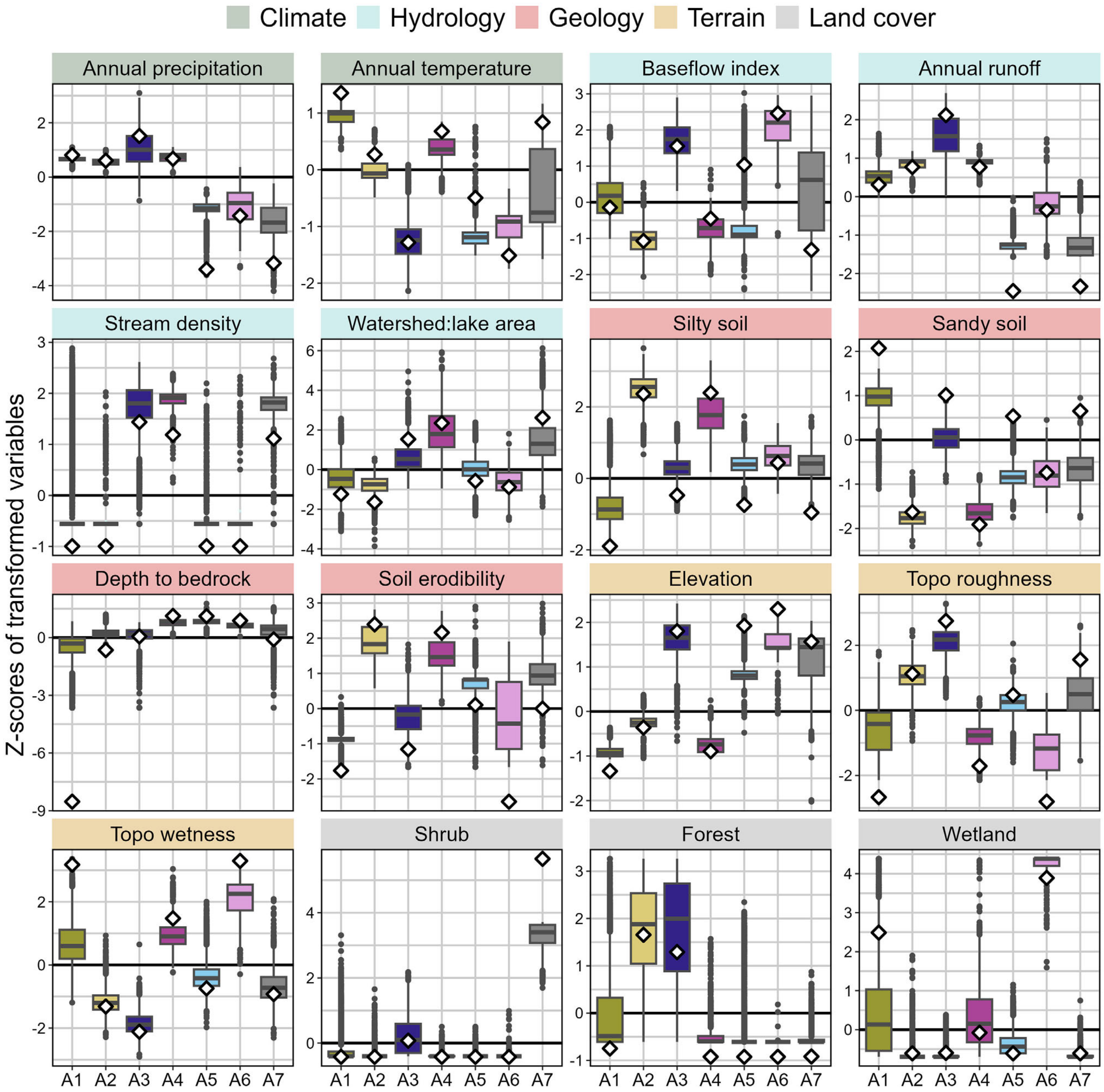

Fig. 2. Distribution of the predictor variables grouped by theme included in the archetype analysis for lakes with weights ≥ 0.75 in each archetype (A1–A7). Values were transformed prior to calculating z-scores. Box height represents the interquartile range from the 25th to 75th percentiles while the solid bar indicates the median value for these highly weighted lakes. The white diamond denotes the median for the five lakes most resembling each archetype. Barring a few exceptions, these represent more extreme values distinguishing lakes from this archetype from the median for all sites.

among the lowest levels of atmospheric deposition of N and S of any archetype (Fig. 4).

Archetype 4

Lakes with high weight in Archetype 4 ($n = 1320$) have high precipitation, temperature, and runoff, low baseflow, and very high stream density. Watershed to lake area ratio is also very high. They also tend to have deep soils that are very silty with very high erodibility. These lakes are located at low elevation in relatively flat terrain (e.g., low topographic roughness) with high topographic wetness and moderate wetland cover; forest cover is low (Fig. 2). Lakes resembling this archetype are typically found in the central region of the conterminous US, between the high elevation lakes resembling archetypes 3, 5, and 7 and the lakes in the valley of the Mississippi and Ohio river resembling archetype 2 (Fig. 3). These lakes tend to have

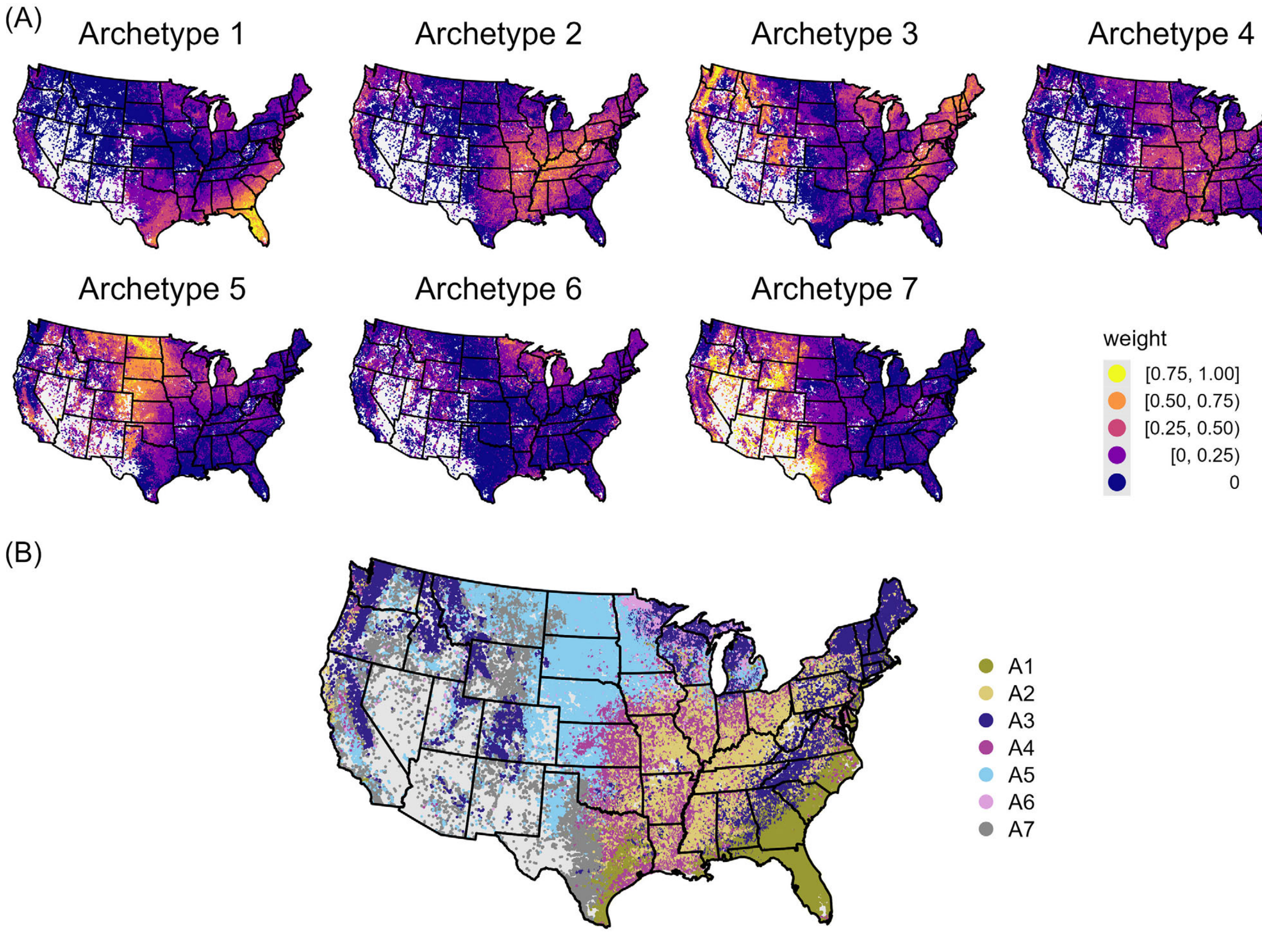

Fig. 3. (**a**) The upper seven panels show maps of the continuous representation of the weights (1 being the highest value and indicating a lake best representing the archetype) assigned to each of the ~ 476,697 study lakes for each of the seven archetypes. (**b**) Study lake location color-coded by the maximum archetype weight associated with the lake. The background color of light gray reflects areas of the country where lake density is very low.

the most convoluted shorelines (high SDF) and the shallowest depths of any archetype. They can be considered the most eutrophic with high nutrient concentrations, low water clarity, and very high Chl*a*. In addition, they are surrounded by watersheds with the highest percentages of total agricultural and crop land among all archetypes. They receive high levels of atmospheric deposition of N and moderate S deposition (Fig. 4).

Archetype 5

Lakes with high weight in Archetype 5 ($n = 10{,}813$) tend to be found at high elevation but low relief areas with deep soils. They are located in very cold and low precipitation areas with watersheds having among the lowest values for runoff, baseflow index and stream density. Terrain is moderately rugged with low topographic wetness; shrub, forest, and wetland land covers are low. They have average values for watershed : lake area ratio, silty and sandy soils, and soil erodibility (Fig. 2) These lakes are mostly found along the central region of the contiguous US, with the highest concentration of lakes toward the northern part (Fig. 3). Lakes resembling Archetype 5 have moderately high agricultural (mostly cultivated crops) and very high grassland land use, are relatively shallow, and have high concentrations of TN, TP, Chl*a*, and DOC, with the lowest transparency of any archetype. The lakes tend to be more circular (low SDF) and receive moderate levels of atmospheric deposition of N and S (Fig. 4).

Archetype 6

The small number of lakes with high weights in Archetype 6 ($n = 351$) are distinguished by the highest watershed wetland land cover and topographic wetness of any archetype and are located in areas with very high baseflow index values and low stream density. Their watersheds receive low precipitation and temperatures are very low, as are stream density, watershed : lake area ratio, and soil erodibility; terrain in the watersheds of these lakes is relatively flat as indicated by the very low topographic roughness (Fig. 2). As a group, the elevation of these lakes is very high. Archetype 6 lakes are mostly found in the northern parts of the Midwestern states,

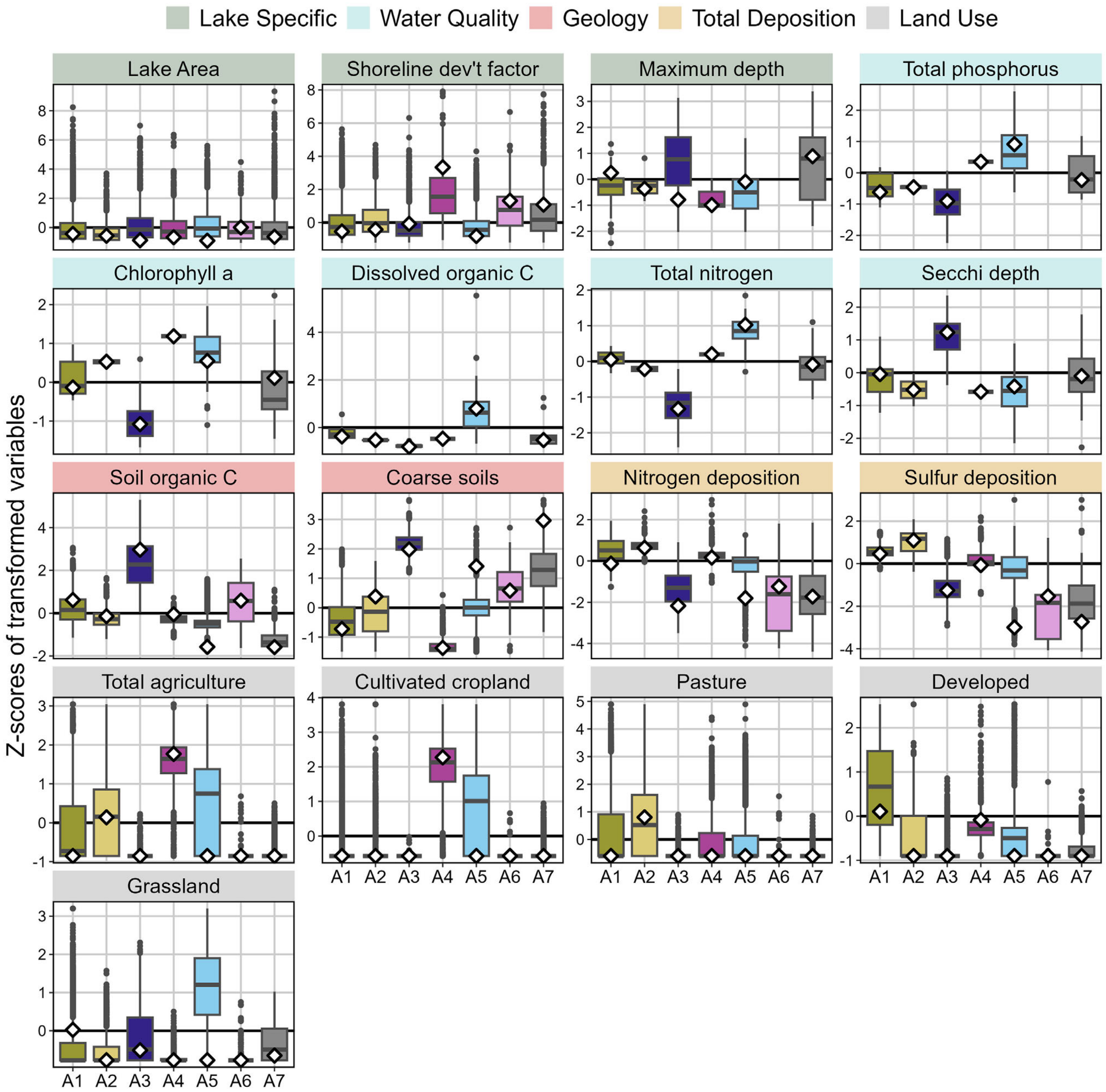

Fig. 4. Post-hoc distributions of geographic and limnological variables for lakes closely resembling (weight ≥ 0.75) each archetype (A1–A7). Box height represents the interquartile range from the 25th to 75th percentiles while the solid bar indicates the median value for these highly weighted lakes. White diamond denotes the median for the five lakes most resembling each archetype. Barring a few exceptions, these represent more extreme values distinguishing lakes from this archetype from the median for all sites.

as well as in other relatively cold and wetland-rich areas (Fig. 3), suggesting that they are primarily found in landscapes with high abundance of peatlands or water-saturated soils (See representative lakes for Archetype 6 in Fig. 5). The watersheds of these lakes have among the lowest percentages for non-wetland land cover categories, as well as the lowest levels of atmospheric deposition of N and S (Fig. 4). This small group of highly weighted lakes in Archetype 6 were not represented in the datasets we used to assess depth and limnological data.

Archetype 7

Lakes with high weights in Archetype 7 ($n = 2906$) are comparable to high elevation lakes from Archetype 3 but have much lower precipitation, baseflow, and runoff, are not as cold, and watersheds are dominated by shrub land cover as opposed to forest. Lakes have very high watershed : lake area ratios and their watersheds have very high stream density. Soils are relatively deep with more silt than sand and erodibility is high while topographic wetness is low (Fig. 2). They are mostly found in west-central part of the conterminous US around the dry areas of the Rocky Mountains, as well as in steeper regions of Texas (Fig. 3). Like lakes resembling Archetype 3, lakes in Archetype 7 are moderate in depth, nutrients, and transparency with low DOC; they differ in having more moderate Chl*a*. Watershed soils are very coarse with low carbon concentration, have low human land use, and receive low levels of atmospheric deposition of N and S (Fig. 4).

Lakes as continuous archetype mixtures with similar states and functions

Limnological and geographic properties tend to be more similar among lakes closely resembling an archetype than among lakes resembling different archetypes (Fig. 4), presumably because they lie in catchments with contrasting geographic properties that are important drivers of lake functioning. These geographic properties are apparent in Fig. 5, which depicts the lakes most closely resembling each archetype for three size classes: the surrounding of lakes of the three size classes resembling Archetypes 4 and 5 (and to a lesser degree, Archetype 2) illustrate how disturbed their lake catchments are, a pattern that is the opposite of archetypes 3, 6, and 7 (Fig. 5).

We tested how accounting for archetype mixtures influenced the fit of widely reported limnological relationships compared to (1) a global model with no classification, (2) a model using the ecoregion classification that forms the basis of the NLA sampling scheme, and (3) a discrete classification model that used the maximum archetype of each lake (Table 2). The global model, which assumes a linear relationship between each pair of limnological variables, yielded R^2 ranging from 0.21 to 0.52 and had a wide range of AIC values (Table 2). All types of classification, discrete or continuous, led to an improvement in prediction (both R^2 and AIC) over the global model, and both models using archetypes had a better fit than the ecoregion model (Table 2). The continuous classification that weighted lakes as continuous mixtures of archetypes performed the best in all cases (Table 2), improving R^2 by 0.06 to 0.18 and reducing AIC by 104 to 216 compared to a global model, depending on the relationships.

Discussion

We found that the 476,697 lakes of the contiguous US are best classified as continuous mixtures of seven geographic

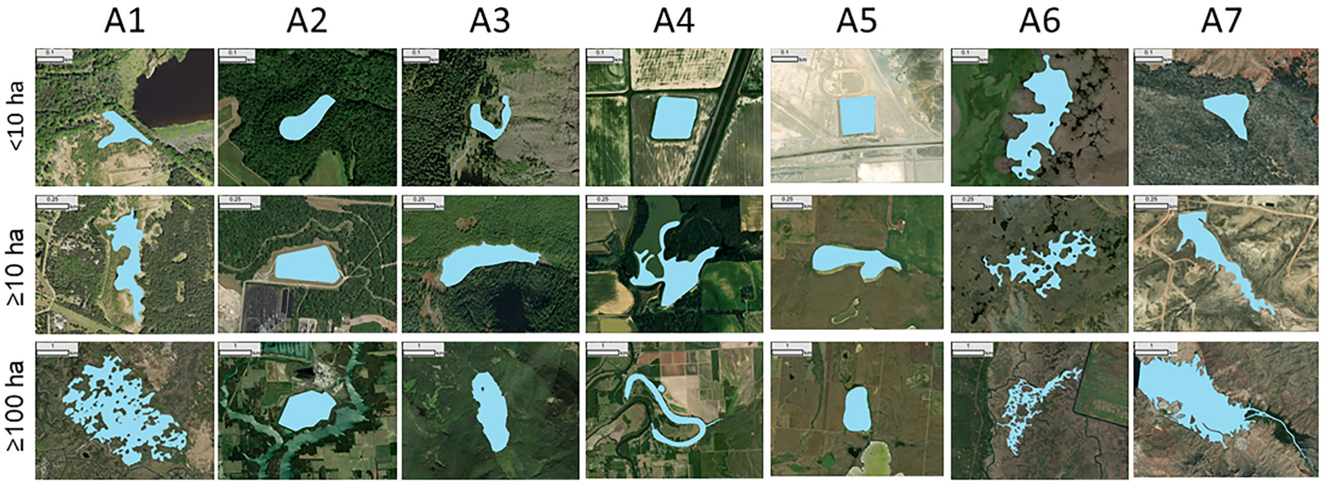

Fig. 5. Polygons for the lakes most closely resembling each archetype (A1–A7), for lakes < 10 ha, 10–100 ha, and > 100 ha, respectively, the top, middle, and bottom row. Polygons are from LAGOS-US LOCUS as derived from the National Hydrography Dataset, High Resolution (U.S. Geological Survey (USGS) 2017*a*).

Table 2. Comparison of a global (all lakes included) vs. discrete (ecoregion, maximum archetype score) and continuous (weighted archetype) classification model for commonly reported limnological relationships.

	R2				AIC			
Limnological relationships	Weighted archetype	Max. Archetype	Ecoregion	Global	Weighted archetype	Max. Archetype	Ecoregion	Global
Chl*a* a vs. TP	0.56	0.51	0.48	0.45	2861	3002	2980	3061
Chl*a* a vs. TN	0.58	0.56	0.53	0.52	2822	2854	2914	2926
Secchi vs. Chl*a*	0.39	0.33	0.23	0.21	3173	3259	3377	3389
Chl*a* a vs. TN : TP	0.62	0.59	0.55	0.52	2711	2799	2863	2917

archetypes defined by climate, hydrology, geology, terrain, and land cover of their watersheds. At the scale of the US, lakes closely resembling a specific archetype had a geographic distribution that was generally aligned with previously used classifications (e.g., ecoregions, Omernik 1987). However, accounting for lake-focused features and considering lakes as continuous mixtures of multiple archetypes allowed for a non-contiguous spatial distribution that better reflects the hierarchical, multi-scale phenomena that influence lake ecosystems. For example, lakes closely resembling a given archetype tended to have similar levels of atmospheric deposition, land use, and water quality, characteristics not included as predictor variables in the analysis; hence lakes in watersheds with similar geographic contexts tend to be more similar to each other in key limnological properties. Additionally, modeling well-established limnological relationships was improved by considering the archetype most closely resembling each lake as a categorical variable, and even more so by weighting the contribution of each archetype to each lake, and both models outperformed a global model with no classification and a widely used ecoregion classification. Therefore, lakes in watersheds with different geographic contexts tend to respond differently to a given environmental driver, and our classification approach provides an intuitive, quantitative, and reproducible framework to account for this geographic variability in lake functioning.

US lakes as continuous mixtures of seven archetypes for improved prediction and understanding of lakes

The archetype modeling provides a weight for each lake that describes its similarity to each of the seven archetypes, and these weights, or a discrete variable denoting the archetype most resembling each lake, can directly be used in limnological analyses as in other classification approaches. In other words, a simple model based on a single categorical variable with seven discrete states (i.e., the "max" archetype for each lake) or a more complex model with seven continuous variables representing the weight of each archetype for a given lake can be directly applied from our classification approach depending on research, management, or other needs. The continuous model using weighted archetypes allows more nuance that accounts for the fact that most lakes often resemble multiple archetypes. Although this model is more complex, the AIC metric indicates that the improvement in prediction may be worth the increased complexity (Table 2).

Our seven archetypes are comparable in number to the 11 Level I Ecoregions (Omernik 1987) in the US, which have successfully been used to group lakes with similar states or responses in a number of broad-scale studies (Taranu et al. 2017; Sprague et al. 2019; Garner et al. 2022). The geographic distribution of the maximum archetype weights for each lake (Fig. 3) is also roughly comparable to the geographic distribution of Level I Ecoregions, and other landscape classifications such as Holdridge Life Zones (Lugo et al. 1999). A key difference with the archetype approach, is that by focusing on a combination of lake-focused, watershed, and regional features, archetypes do not have a contiguity constraint and they can include watershed-specific predictors at finer scales than ecoregion approaches allow. Like the ecoregion approach, and unlike several classification approaches widely used in European countries (e.g., Solheim et al. 2019; Poikane et al. 2022), archetype modeling has the added benefit of not relying on in-lake data and can thus be applied to unsampled lakes.

The archetype predictors often have high levels of spatial auto-correlation at continental extents (see Lapierre et al. 2018); hence the weights of a given archetype or the maximum archetype for each lake still follow broadly structured spatial patterns. However, nearby lakes can be different archetypes, particularly toward the edges of areas dominated by a specific archetype (Fig. 3). Contiguity of regions can be advantageous from a management point of view, as it can be more practical to define and apply policies uniformly within an administrative unit (Cheruvelil et al. 2021 and references therein), and our results support that using a contiguous classification (e.g., ecoregions) is always better than no classification (i.e., global models). Therefore, considering some type of ecologically relevant classification appears to always improve lake predictions, particularly for water quality measures.

Lake archetypes adequately capture among-lake variation as there was little overlap in the distributions of in-lake measurements (TP, TN, Chl *a*, dissolved organic carbon, Secchi depth) among lakes closely resembling a given archetype. This was the case even though limnological variables presented in Fig. 4 were collected by a sampling program (the National Lake Assessment, USEPA 2016) that bases its sampling design on terrestrial ecoregions, which led to imbalanced sample size among archetypes (note the absence of limnological data for Archetype 6, Fig. 4). In line with previous studies (Sadro et al. 2012; Rose et al. 2014), concentrations of all water quality variables were the lowest in the pristine, deep, and high elevation lakes from Archetypes 3 and 7. At the other end of the spectrum lie lakes resembling Archetypes 4 and 5. These lakes were mainly found on and downstream of the high elevation, flat plateau of the central US (Figs. 2, 3) and co-occurred with the densest concentrations of cultivated crops and the highest levels of atmospheric deposition (Fig. 4). These human drivers are likely sufficient in themselves to explain the higher concentrations of nutrients and lower water transparency measured in situ (Carpenter et al. 1998), but additional factors related to the geographic features of their watershed are likely contributing factors, such as morphometry or lake origin, which may influence in-lake nutrient processing (Read et al. 2015; Casas-Ruiz et al. 2021). There were no systematic differences in lake area among archetypes, but lakes resembling Archetypes 4 and 5 are typically shallow (Fig. 4). These features presumably explain the contrasting driver-response relationships among archetypes (Table 2), suggesting that lakes found in watersheds with contrasting geographic properties may respond differently to a given disturbance.

Implications for predicting lake response to broad scale change

Globally, lakes are subject to stressors such as climate and land use change that have complex interactions and diverse response trajectories (Hayes et al. 2015; Zia et al. 2016; Hansen et al. 2022). The design of monitoring programs to collect in situ ecosystem data is notoriously challenging because of the need to capture the full range of ecologically relevant heterogeneity that exists at both local and regional scales (Janousek et al. 2019; Soranno et al. 2020). Our results show that a continuous approach based on lake archetype classification can facilitate the identification of lakes with similar limnological states and responses, two key management targets that can sometimes form the foundation for such monitoring programs (Soranno et al. 2010; Yuan et al. 2014; Poikane et al. 2022). For example, because all 476,697 conterminous US lakes ≥ 1 ha are assigned a weight for each archetype, it is possible to use the archetype model output to estimate potential ranges of water quality related in-lake measurements (nutrients, carbon, clarity), even for lakes that have not been sampled. Such a predictive approach would provide water quality estimates to management agencies on all lakes in a region or jurisdiction that could aid in management decisions. These results are important because even if lakes have similar TP concentrations, the response of algal biomass to changes in nutrients is not universal and likely differs depending on the lake archetype mixture (Table 2). It has been previously shown that algal-nutrient relationships vary spatially (Fergus et al. 2016; Liang et al. 2020; Zhou et al. 2022), and the results presented here suggest that archetype classification can be used to group lakes with presumed similar responses, including lakes that have not been sampled.

Application of the seven archetypes described here to the assessment of other lake pressures of local to national interest has yet to be tested. Availability of archetype assessments at the scale of the conterminous US may complement such discrete assessments by uncovering potentially sensitive lake populations not previously identified due to lack of in-situ sampling information. This further points toward the need to develop better driver-response relationships for under-sampled lake archetypes (e.g., lack of lakes closely resembling archetype 6 in Fig. 4) to better understand their current functioning and future responses to direct and indirect human pressures. Finally, we suggest that finer, more local approaches may be needed in regions where most lakes appear to resemble a single archetype. Indeed for pressures such as acid rain where lake sensitivities have been well studied, a targeted discrete assessment approach may be more optimal, for example in acid-sensitive Rocky Mountain lake sub-populations (Nanus et al. 2009). However, we argue that archetype classification may be the most useful at broad spatial extents or in areas where several distinct archetypes are found within close proximity, where the population of lakes is very diverse, and knowledge about the effects of critical landscape features is limited. Under such circumstances, the flexible, continuous nature of archetype classification may be particularly useful for exploring emerging environmental pressures that are not as documented as long-standing problems such as eutrophication and acid rain.

Data availability statement

The data discussed in this article are available from Zenodo (DOI: 10.5281/zenodo.10008350). Two data tables with observations by lake are provided: one with the transformed predictors and archetype weights and maximum archetype and the second with the raw data for predictors, post-hoc response variables, and limnological modeling variables. The two tables can be linked with a unique lake identifier, lagoslakeid. A data dictionary defining columns in each data table is also provided.

References

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

Carpenter, S. R., N. F. Caraco, D. L. Correll, R. W. Howarth, A. N. Sharpley, and V. H. Smith. 1998. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. **8**: 559–568.

Casas-Ruiz, J. P., J. Jakobsson, and P. A. Del Giorgio. 2021. The role of lake morphometry in modulating surface water carbon concentrations in boreal lakes. Environ. Res. Lett. **16**. doi:10.1088/1748-9326/ac0be3

Cheruvelil, K. S., P. A. Soranno, I. M. McCullough, K. E. Webster, L. K. Rodriguez, and N. J. Smith. 2021. LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. Limnol. Oceanogr. Lett. **6**: 270–292. doi:10.1002/lol2.10203

Cutler, A., and L. Breiman. 1994. Archetypal analysis. Dent. Tech. **36**: 338–347. doi:10.1080/00401706.1994.10485840

Dodds, W. K., and others. 2019. The freshwater biome gradient framework: Predicting macroscale properties based on latitude, altitude, and precipitation. Ecosphere **10**. doi:10.1002/ecs2.2786

Eisenack, K., and others. 2019. Design and quality criteria for archetype analysis. Ecol. Soc. **24**. doi:10.5751/ES-10855-240306

Emmons, E. E., M. J. Jennings, and C. Edwards. 1999. An alternative classification method for northern Wisconsin lakes. Can. J. Fish. Aquat. Sci. **56**: 661–669. doi:10.1139/f99-008

Eugster, M. J. A., and F. Leisch. 2009. From spider-man to hero–archetypal analysis in R. J. Stat. Softw. **30**: 1–23. doi:10.18637/jss.v030.i08

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

Garner, R., S. Kraemer, V. Onana, Y. Huot, I. Gregory-Eaves, and D. Walsh. 2022. Protist diversity and metabolic strategy in Freshwater Lakes are shaped by trophic state and watershed land use on a continental scale. mSystems **7**: 7. doi:10.1128/msystems.00316-22

Gebert, W. A., Graczyk, D. J., and Krug, W. R. 1987. Average annual runoff in the United States, 1951–1980 (No. 710). US Geological Survey.

Hansen, G. J. A., K. Wehrly, K. Vitense, J. Walsh, and P. Jacobson. 2022. Quantifying the resilience of Coldwater lake habitat to climate and land use change to prioritize watershed conservation. Ecosphere **13**: e4172. doi:10.1002/ecs2.4172

Harrison, P. A., and others. 2019. Synthesizing plausible futures for biodiversity and ecosystem services in europe and central asia using scenario archetypes. Ecol. Soc. **24**. doi:10.5751/ES-10818-240227

Hayes, N. M., M. J. Vanni, M. J. Horgan, and W. H. Renwick. 2015. Climate and land use interactively affect lake phytoplankton nutrient limitation status. Ecology **96**: 392–402. doi:10.1890/13-1840.1

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., … and Kempen, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS one. **12**: e0169748.

Herlihy, A. T., S. G. Paulsen, J. Van Sickle, J. L. Stoddard, C. P. Hawkins, and L. L. Yuan. 2008. Striving for consistency in a national assessment: The challenges of applying a reference-condition approach at a continental scale. J. North Am. Benthol. Soc. **27**: 860–877. doi:10.1899/08-081.1

Hill, R. A., M. H. Weber, R. M. Debbout, S. G. Leibowitz, and A. R. Olsen. 2018. The Lake-catchment (LakeCat) dataset: Characterizing landscape features for lake basins within the conterminous USA. Freshw. Sci. **37**: 208–221. doi:10.1086/697966

Janousek, W. M., B. A. Hahn, and V. J. Dreitz. 2019. Disentangling monitoring programs: Design, analysis, and application considerations. Ecol. Appl. **29**: e01922. doi:10.1002/eap.1922

Kelly, F. L., A. J. Harrison, M. Allen, L. Connor, and R. Rosell. 2012. Development and application of an ecological classification tool for fish in lakes in Ireland. Ecol. Indic. **18**: 608–619. doi:10.1016/j.ecolind.2012.01.028

Lapierre, J. F., and others. 2018. Similarity in spatial structure constrains ecosystem relationships: Building a macroscale understanding of lakes. Glob. Ecol. Biogeogr. **27**: 1251–1263. doi:10.1111/geb.12781

Lemm, J. U., and others. 2021. Multiple stressors determine river ecological status at the European scale: Towards an integrated understanding of river status deterioration. Glob. Chang. Biol. **27**: 1962–1975. doi:10.1111/gcb.15504

Lewis, W. M., Jr. 1983. A revised classification of lakes based on mixing. Can. J. Fish. Aquat. Sci. **40**: 1779–1787. doi:10.1139/f83-207

Liang, Z., P. A. Soranno, and T. Wagner. 2020. The role of phosphorus and nitrogen on chlorophyll a: Evidence from hundreds of lakes. Water Res. **185**: 116236. doi:10.1016/j.watres.2020.116236

Lugo, A. E., S. B. Brown, R. Dodson, T. S. Smith, and H. H. Shugart. 1999. The Holdridge life zones of the conterminous United States in relation to ecosystem mapping. J. Biogeogr. **26**: 1025–1038. doi:10.1046/j.1365-2699.1999.00329.x

Miller, D. A., and White, R. A. 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth interactions. **2**: 1–26.

NADP (2022). Data obtained from a website: https://catalog.data.gov/dataset/nadp-total-deposition-data

Nanus, L., M. W. Williams, D. H. Campbell, K. A. Tonnessen, T. Blett, and D. W. Clow. 2009. Assessment of lake sensitivity to acidic deposition in national parks of the Rocky Mountains. Ecol. Appl. **19**: 961–973. doi:10.1890/07-1091.1

Oberlack, C., Sietz, D., Bonanomi, E. B., De Bremond, A., Dell'Angelo, J., Eisenack, K., … and Villamayor-Tomas, S. 2019. Archetype analysis in sustainability research. Ecology and Society. **24**: 19.

Omernik, J. M. 1987. Ecoregions of the conterminous United States. Ann. Assoc. Am. Geogr. **77**: 118–125. doi:10.1111/j.1467-8306.1987.tb00149.x

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

Poikane, S., and others. 2022. Estimating nutrient thresholds for eutrophication management: Novel insights from understudied lake types. Sci. Total Environ. **827**: 154242. doi:10.1016/j.scitotenv.2022.154242

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, http//www. R-project.org.

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

Rose, K. C., and others. 2014. Light attenuation characteristics of glacially-fed lakes. J. Geophys. Res. Biogeo. **119**: 1446–1457. doi:10.1002/2014JG002674

Sadro, S., C. E. Nelson, and J. M. Melack. 2012. The influence of landscape position and catchment characteristics on aquatic biogeochemistry in high-elevation Lake-chains. Ecosystems **15**: 363–386. doi:10.1007/s10021-011-9515-x

Sietz, D., U. Frey, M. Roggero, Y. Gong, N. Magliocca, R. Tan, P. Janssen, and T. Václavík. 2019. Archetype analysis in sustainability research: Methodological portfolio and analytical frontiers. Ecol. Soc. **24**. doi:10.5751/ES-11103-240334

Smith, N. J., K. E. Webster, L. K. Rodriguez, K. S. Cheruvelil, and P. A. Soranno. 2021. LAGOS-US LOCUS: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. Limnol. Oceanogr. Lett. **6**: 270–292.

Smith, N. J., K. E. Webster, L. K. Rodriguez, K. S. Cheruvelil, and P. A. Soranno. 2022. LAGOS-US GEO v1.0: Data module of lake geospatial ecological context at multiple spatial and temporal scales in the conterminous U.S. ver 3. Limnol. Oceanogr. Lett. **6**. doi:10.6073/pasta/0e443bd43d7e24c2b6abc7af54ca424a

Solheim, A. L., and others. 2019. A new broad typology for rivers and lakes in Europe: Development and application for large-scale environmental assessments. Sci. Total Environ. **697**: 134043. doi:10.1016/j.scitotenv.2019.134043

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. 2020. Ecological prediction at macroscales using big data: Does sampling design matter? Ecol. Appl. **30**: e02123. doi:10.1002/eap.2123

Sprague, L. A., R. M. Mitchell, A. I. Pollard, and J. A. Falcone. 2019. Assessing water-quality changes in US rivers at multiple geographic scales using results from probabilistic and targeted monitoring. Environ. Monit. Assess. **191**: 348. doi:10.1007/s10661-019-7481-5

Stachelek, J., and others. 2021. LAGOS-US DEPTH v1.0: Data module of observed maximum and mean lake depths for a subset of lakes in the conterminous U.S. ver 1. Environ. Data Initiative.

Steinley, D. 2006. K-means clustering: A half-century synthesis. Br. J. Math. Stat. Psychol. **59**: 1–34. doi:10.1348/000711005X48266

Taranu, Z. E., I. Gregory-Eaves, R. J. Steele, M. Beaulieu, and P. Legendre. 2017. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. Glob. Ecol. Biogeogr. **26**: 625–637. doi:10.1111/geb.12569

U.S. Environmental Protection Agency (USEPA). 2016. National Aquatic Resource Surveys. National Lakes Assessment 2012 (data and meta data files). U.S. Environmental Protection Agency (USEPA), https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Environmental Protection Agency (USEPA). 2022. National Aquatic Resource Surveys. National Lakes Assessment 2017 (data and meta data files). U.S. Environmental Protection Agency (USEPA), https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Geological Survey (USGS). 2017*a*. National hydrography dataset ver. USGS National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU). U.S. Geological Survey (USGS), https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products

U.S. Geological Survey (USGS). 2017*b*. National elevation dataset (ver. USGS 30 Meter Resolution, One-Sixtieth Degree National Elevation Dataset (NED) for CONUS, Alaska, Hawaii, Puerto Rico, and the U.S. Virgin Islands (published 20170424)). U.S. Geological Survey (USGS), https://ned.usgs.gov/

U.S. Geological Survey (USGS). 2017*c*. 3D elevation program 1-meter resolution digital elevation model. U.S. Geological Survey (USGS), https://www.sciencebase.gov/catalog/item/543e6b86e4b0fd76af69cf4c

U.S. Geological Survey (USGS). 2019*a*. National Hydrography Dataset (ver. USGS NHDPlus High Resolution (HR) Beta for Hydrologic Unit (HU) 4 (published 20190507)). U.S. Geological Survey (USGS), https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products

U.S. Geological Survey (USGS). 2019*b*. NLCD 2016 Land Cover Conterminous United States. U.S. Geological Survey (USGS), https://doi.org/10.5066/P937PN4Z

U.S. Geological Survey (USGS). 2021. National Hydrography Dataset (ver. USGS NHDPlus High Resolution (HR) Beta for Hydrologic Unit (HU) 4–2001). U.S. Geological Survey (USGS), https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products

U.S. Geological Survey Earth Resources Observation and Science (USGS/EROS). 2003. Elevation Derivatives for National Applications (EDNA) Compound Topographic Index (CTI). Earth Resources Observation and Science (EROS) Center. doi:10.5066/F7TD9VTQ

Wetzel, R. G. 2001. Limnology, Lake and River Ecosystems. Elsevier, p. 850.

Wolock, D. M. 2003. Base-flow index grid for the conterminous United States. U.S. Geological Survey. doi:10.3133/ofr03263

Yuan, L. L., A. I. Pollard, S. Pather, J. L. Oliver, and L. D'Anglada. 2014. Managing microcystin: Identifying national-scale thresholds for total nitrogen and chlorophyll *a*. Freshw. Biol. **59**: 1970–1981. doi:10.1111/fwb.12400

Zhou, J., P. R. Leavitt, Y. Zhang, and B. Qin. 2022. Anthropogenic eutrophication of shallow lakes: Is it occasional? Water Res. **221**: 118728. doi:10.1016/j.watres.2022.118728

Zia, A., and others. 2016. Coupled impacts of climate and land use change across a river-lake continuum: Insights from an integrated assessment model of Lake Champlain's Missisquoi Basin, 2000-2040. Environ. Res. Lett. **11**. doi:10.1088/1748-9326/11/11/114026

Acknowledgments

We thank the Continental-Limnology team (People–LAGOS (lagoslakes.org)) for feedback and discussion throughout the development of this study. This research was funded by the National Science Foundation (DEB-1638679; DEB-1638554; DEB-1638539; and DEB-1638550). P.A.S. also acknowledges support from the National Institute of Food and Agriculture, Hatch Project 176820. J.F.L. is supported by a NSERC-Discovery grant. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This study has considerably improved following the inputs of the Associate Editor (John Downing) and two anonymous reviewers.

Conflict of Interest

None declared.

Submitted 30 September 2022
Revised 12 October 2023
Accepted 04 November 2023

Associate editor: John Downing