LIMNOLOGY AND OCEANOGRAPHY
Letters

ASLO
Open Access

Limnology and Oceanography Letters 2021

doi: 10.1002/lol2.10203

DATA ARTICLE

LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S.

Kendra Spence Cheruvelil,[1,2]* *Patricia A. Soranno*,[1] *Ian M. McCullough*,[1] *Katherine E. Webster*,[1] *Lauren K. Rodriguez*,[1,3] *Nicole J. Smith*[1]

[1]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan; [2]Lyman Briggs College, Michigan State University, East Lansing, Michigan; [3]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, Maryland

Scientific Significance Statement

LAGOS-US LOCUS v1.0 is a research-ready open data platform that will facilitate studying lakes at broad scales. It includes 479,950 lakes and reservoirs larger than or equal to 1 ha across the entire conterminous U.S. The module includes spatially-explicit information on lakes and their watersheds, including location information, physical characteristics, and identifiers so that it can be linked with other important national-scale datasets.

Abstract

Macroscale studies of aquatic ecosystems are needed to address many contemporary broad-scale problems related to global change, particularly to inform national-scale environmental policies. In this data paper, we fill two important gaps in data availability for lakes at the scale of the conterminous U.S., the lack of: (1) high-resolution geographic representations of lakes and their watersheds and (2) an open data platform to connect disparate U.S. national-scale lake datasets. We describe the LAGOS-US LOCUS v1.0 data module that includes a detailed User Guide, which is part of the LAGOS-US extensible research-ready open data platform. This platform can be used to study the 479,950 lakes and reservoirs larger than or equal to 1 ha across the entire conterminous U.S. at multiple scales of space and time. The LOCUS module includes spatially-explicit information on lakes and their watersheds, including location information, physical characteristics, and identifiers of other important national-scale datasets.

*Correspondence: ksc@msu.edu

Associate editor: Jill Baron

Author Contribution Statement: P.A.S. and K.S.C. conceived of the data module, created and managed the research team, and provided the structure for the data and documentation. K.S.C. supervised the writing of the data paper. K.S.C. and I.M.M. drafted the background/motivation and data description sections, N.J.S. and K.E.W. drafted the methods section, and P.A.S. drafted the technical validation section. Figures were drafted by the following: 1 by L.R. and N.J.S.; 2 and 3 by K.S.C. and P.A.S.; 4 by I.M.M.; 5 by L.R. and K.E.W.; 6 by L.R.; 7–10 and 12 by K.E.W.; 11 and 13 by N.J.S. Table 1 was created by P.A.S. N.J.S. created the vision behind the database design, methods, and data model for LOCUS. She led data creation, documentation, and validation processes for LOCUS, with K.E.W. and L.R. providing essential assistance. In particular, K.E.W. created the naming conventions for the module, led and conducted the QAQC analyses, and created lake bounding geometry metrics and developed associated flags; L.R. provided key assistance for the lake_link table, QAQC process, and general data management tasks. All co-authors collectively framed and reviewed the paper.

Data Availability Statement: Data, metadata, and code are available on the Environmental Data Initiative (EDI) Data Portal at https://doi.org/10.6073/pasta/e5c2fb8d77467d3f03de4667ac2173ca.

There is an increasing need for conducting research on environmental ecosystems, such as lakes, streams, and forests, at broad scales of space and time. For example, macroscale (i.e., regional- to continental-scale) research is needed to understand global change and the role of ecosystems in global cycles, and to enhance management and conservation efforts critical to informing national environmental policy. Such studies require a wide range of information on different components of the environment. However, efforts have been hampered by data sources with different identifiers, lack of data standards, and a dearth of associated metadata and documentation. These limitations make it notoriously difficult to connect, integrate, and harmonize disparate data sources to answer macroscale questions.

Macroscale databases can come about in three main ways. First, there are increasing numbers of big-science initiatives such as the U.S. EPA National Aquatic Resource Surveys (www.epa.gov/national-aquatic-resource-surveys), the U.S. National Ecological Observatory Network (Thorpe et al. 2016), and the European Multi-Lake Survey (Mantzouki et al. 2018). Second, there are researchers using modeling and/or remotely sensed products to generate broad-scale estimates of key ecological processes such as estimating chlorophyll *a* (Chl *a*) at the river basin scale (Kuhn et al. 2019) or fire extent and severity using remote sensing (Eidenshink et al. 2007). Third, databases may be created from the harmonization of disparate and smaller datasets into integrated data products such as LAGOS-NE (Soranno et al. 2017) and the Canadian Lake Data Archive (Huot et al. 2019). This data paper describes the outcomes of applying this last approach to create the open-access platform, LAGOS-US, which uses common identifiers to connect disparate data sources and, where possible, uses existing or establishes new standards for data, metadata, and documentation. This extensible, research-ready platform can be used to study all 479,950 lakes and reservoirs (hereafter referred to as lakes) larger than or equal to 1 ha in the conterminous U.S. (i.e., 48 states and the District of Columbia) at multiple spatiotemporal scales (Fig. 1). We focus on the national scale to leverage many existing valuable datasets on lakes and other ecological characteristics that are available in the U.S. and to build an open data research platform that could be used as a model for national efforts in other countries.

We identified two key knowledge gaps hampering the study of lakes in the conterminous U.S. at broad scales of space or time. First, watersheds are a critical characteristic for lake research and management because they represent the contributing land area and the surface waters that drain into a lake, and thus, are highly influential on lake physical, chemical, and biological properties. However, there is currently no publicly available, high resolution, polygon watershed dataset for U.S. lakes. We filled this gap by delineating watersheds for all lakes in the conterminous U.S. Second, because individual efforts cannot produce a single database that has all relevant information necessary for macroscale lake research, there needs to be an easy way to connect existing valuable lake datasets. Therefore, we included unique identifiers from the major conterminous U.S. lake datasets to facilitate the combined use of multiple lake datasets.

This paper describes LOCUS, the first of three core database modules that make up the open data platform LAGOS-US (Fig. 2). LOCUS has locational, identifying, and physical information for all lakes ≥ 1 ha in surface area and their watersheds in the conterminous U.S. This first core module is the foundation for the LAGOS-US platform because it includes the basic information needed to connect to any other LAGOS-US module and serves as a foundation for future lake studies of the conterminous U.S. The second core module is GEO, which includes the geospatial and temporal ecological context variables (e.g., land use, climate, hydrology) for all lakes in LOCUS characterized at multiple spatial divisions (e.g., equidistant buffers around lakes, watersheds, ecoregions). The third core module, LIMNO, includes limnological physical, chemical, and biological measurements through time for a subset of lakes ≥ 1 ha.

LAGOS-US extension modules contain information for specific lake characteristics that often require different data sources or new analyses on either all or a subset of lakes in LOCUS. Upcoming extension modules that will be published soon include the following. RESERVOIR provides a predicted classification of all 137,465 lakes ≥ 4 ha as either a natural lake or a reservoir using a machine-learning algorithm and aerial imagery. LAKE DEPTH includes mean and/or maximum depth measurements of over 17,675 lakes ≥ 1 ha that were manually compiled from a wide range of online sources. NETWORKS uses graph theory to identify 898 lake networks that include 86,511 lakes ≥ 1 ha and provide quantitative surface water connectivity metrics for those networks and lakes (King et al. in press). LANDSAT provides predicted water quality measurements for Chl *a*, Secchi depth, and colored dissolved organic matter for the 137,465 lakes ≥ 4 ha using machine-learning models based on atmospherically corrected Landsat imagery and LIMNO data, in addition to lake-wide values of reflectance for each Landsat band and satellite overpass.

All LAGOS-US data modules are designed for reuse and extension by the broader research community. Therefore, each module will be made available in individual data repositories in the Environmental Data Initiative data repository, will include a detailed User Guide and metadata documentation, and will be accessible using the "LAGOSUS" R package (Stachelek 2021). The LOCUS module includes a detailed User Guide (Smith et al. 2021), a crosswalk table that facilitates connections to existing lake database resources, GIS data files of lakes and their watersheds, and extensive documentation of methods, code, and data.

Data description

The LAGOS-US LOCUS v 1.0 data module contains information on all 479,950 lakes and watersheds in the

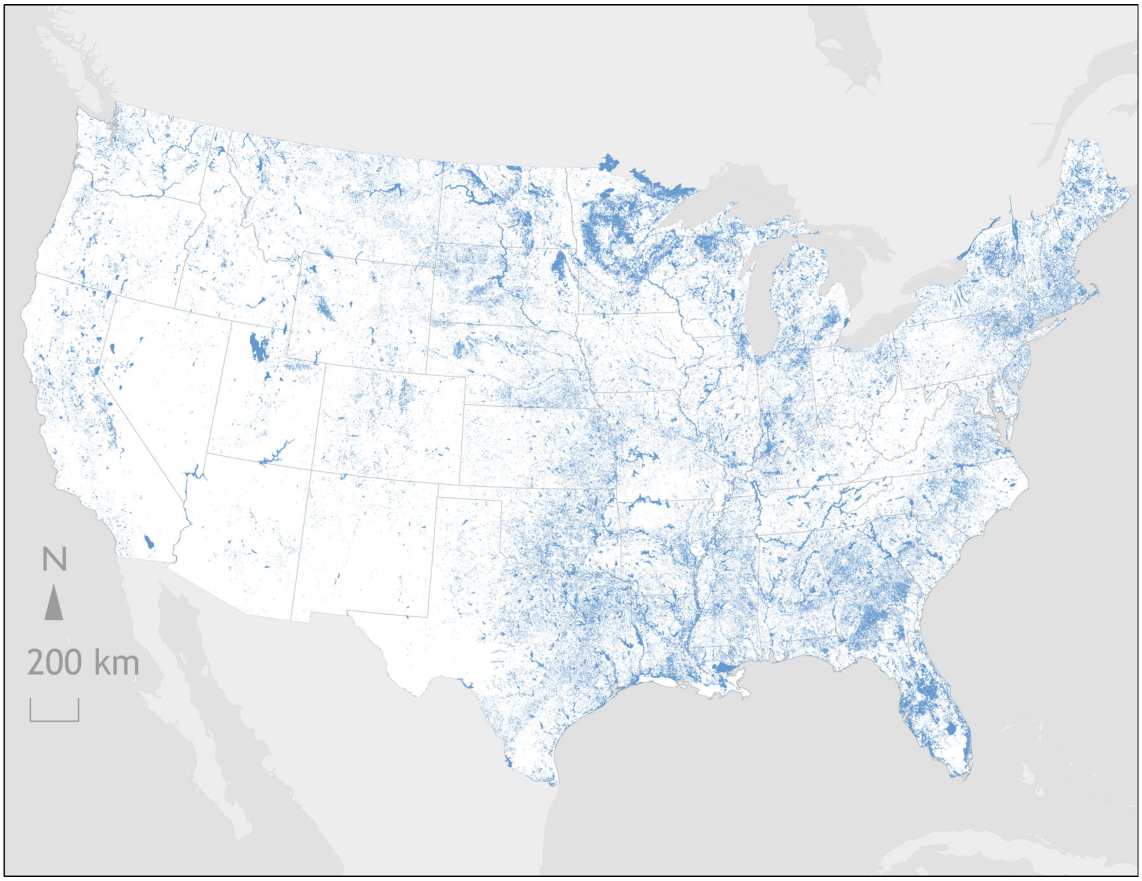

Fig. 1. Map of the 479,950 lakes ≥ 1 ha in the 48 states and the District of Columbia (i.e., conterminous U.S.) that are in LAGOS-US.

conterminous U.S. ≥ 1 ha in surface area, which we define as the “census lake population.” Lakes and watersheds are the two entities in the LOCUS module that are linked with other LAGOS modules using the common LOCUS lake identifier *lagoslakeid* and the common spatial division identifier *zoneid*.

A lake is defined as a perennial body of relatively still water ≥ 1 ha with a geographically defined polygon in the high-resolution National Hydrography Dataset (NHD HR). LOCUS includes lakes and reservoirs that are either completely natural, modified natural (i.e., a water control structure on a natural lake), or highly modified (i.e., a fully impounded stream or river); however, lakes are not coded using these categories due to the lack of a classification based on human modification. However, we explicitly exclude entirely artificial basins or those built for high-intensity human use based on our interpretation of labels assigned by the NHD HR—these include but are not limited to sewage treatment ponds, aquaculture ponds, and retention ponds.

We define a lake watershed as an area of land that drains surficially into a lake either through potential land surface flow or via connected streams or upstream lakes. We use a limnological interpretation of the landscape that recognizes lakes as sinks, trapping nutrients and sediments and affecting the transport of such materials to downstream streams and lakes (e.g., Marcarelli and Wurtsbaugh 2007; Brett and Benjamin 2008; Zhang et al. 2012). Therefore, our lake watershed delineations factor in the effect of upstream water bodies. However, recognizing that small water bodies will have minimal impacts on downstream lakes based on simple accounting, we selected a lake surface area threshold of 10 ha to indicate whether an upstream drainage is sufficient to result in limnologically important downstream contributions to a focal lake (i.e., the downstream lake). Thus, when a lake has an upstream-connected lake that is ≥ 10 ha, we assume that the large upstream lake serves as a depositional basin of the upstream-draining land and we do not include the drainage areas of that upstream lake in the focal lake’s watershed. Lakes without an outflow ≥ 10 ha are similarly considered as depositional basins for this definition and are treated as sinks that drain local regions excluded from the focal lake’s watershed for the lake watershed delineation. Based on this conceptualization, every lake has a unique lake watershed. For each lake connected to one or more larger lakes (≥ 10 ha), we delineated the network watershed that consists of the unique lake

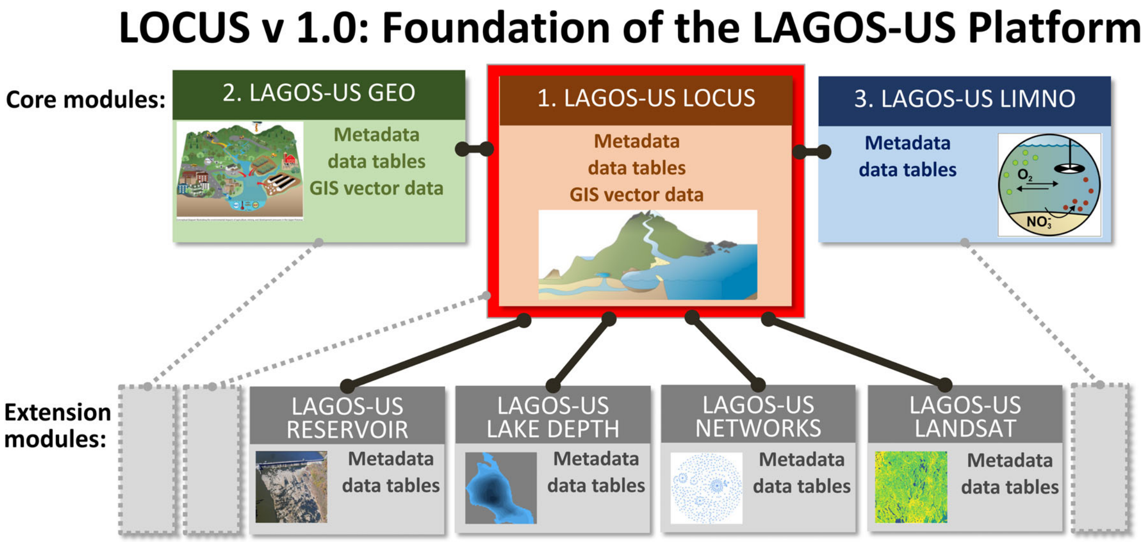

Fig. 2. The LAGOS-US platform: Core and extension modules. LAGOS-US includes three core modules, with the LOCUS v 1.0 data module playing a central role in the platform by connecting to all other LAGOS-US modules through unique identifiers. LOCUS has locational, identifying, and physical information for lakes and their watersheds. GEO includes the geospatial and temporal ecological context variables (e.g., land use, climate, hydrology) of all lakes in LOCUS characterized for multiple spatial divisions (e.g., equidistant buffers around lakes, watersheds, ecoregions). LIMNO includes surface-water limnological physical, chemical, and biological measurements for a subset of lakes in LOCUS through time. The LAGOS-US platform also includes four in-development extension modules: RESERVOIR, LAKE DEPTH, NETWORKS, and LANDSAT; *see* text for details). Gray boxes with dashed lines are placeholders for future extension modules that can be created by any developer or user and that can link to any of the three core modules. To facilitate data reuse, each module contains multiple data and metadata tables, with additional observation-level metadata embedded in the data tables; when applicable, modules also contain GIS vector data. An R package called "LAGOSUS" is available to facilitate access (Stachelek 2021). LANDSAT image courtesy of the U.S. Geological Survey and images for GEO, LOCUS, LIMNO, and RESERVOIR are from the Integration and Application Network, University of Maryland Center for Environmental Science (https://ian.umces.edu/imagelibrary/).

watershed plus the combined watersheds of all upstream lakes. Note that all representations and calculations of lake watersheds do not include the lake itself, even though we describe lakes as "nested" within watersheds. *See* "Lake watersheds" section below for a more detailed description.

LOCUS provides important information regarding the characteristics of lakes including: variables that can be obtained from GIS data such as location and geometry; variables that can be derived using GIS processing such as lake hydrologic connectivity, lake watersheds, and watershed geometry; and commonly used identifiers from other data products useful for linking with LAGOS-US. Some lake, lake watershed, and network watershed boundaries in LOCUS extend slightly beyond the conterminous U.S. crossing the borders of either Canada or Mexico. Because NHD international data harmonization efforts have provided source data extending beyond the political boundary, LOCUS variables including the connectivity variables are not limited by data availability beyond the U.S. border.

LOCUS consists of data as tables, GIS vector data, observation level flags, two metadata tables (Fig. 3), and a detailed User Guide (Smith et al. 2021). LOCUS metadata includes a data source table, a data dictionary table, and individual observation-level flags for some variables that are included within the data tables. The accompanying geospatial data are provided as a GIS vector dataset consisting of lakes as polygons and centroids, and watersheds as polygons. The four LOCUS data tables contain the observations of variables, grouped based on similarity of data source, methods used to create the variable, or general category of variable: lake information, lake characteristics, lake link, and lake watersheds (*see* below for more detail).

Lake information: This data table includes common identifiers such as the LAGOS lake name (collated from multiple sources), locational information (horizontal and vertical geographic coordinates, U.S. state; Fig. 4), and flags associated with these observations. Refer to Soranno et al. (2020) for a more detailed description of U.S. lake names, including regional patterns and commonly used, derogatory, or unusual names.

Lake link: This data table includes single or multiple identifiers for each lake from other commonly used national-scale data products including: Water Quality Portal (WQP), Geographic Names Information System (GNIS), NHD (medium-resolution) Plus v2.1, LAGOS-NE, and the EPA National Lakes Assessment surveys from 2007 and 2012. Linked identifiers are included in the table only when a lake is present in the LAGOS-US lake population.

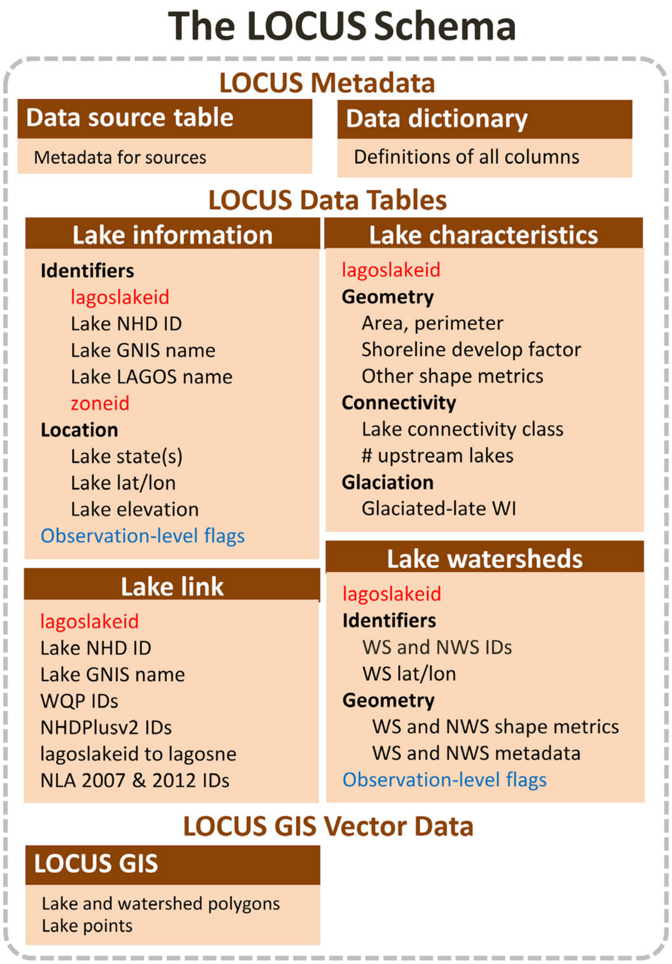

Fig. 3. The LAGOS-US LOCUS schema. Along with a detailed User Guide (Smith et al. 2021), LOCUS includes metadata in the form of a source table and a data dictionary, four data tables (lake_information.csv, lake_characteristics.csv, lake_watersheds.csv, and lake_link.csv), and a geopackage. The tables are connected to each other and other LAGOS-US modules via two unique, key identifiers depicted with red text: *lagoslakeid* and *zoneid*. Three of the four tables also include observation-level flags, depicted with blue text. The variables in the four data tables that are listed in black text are representative examples, rather than exhaustive. The census population of lakes is $N = 479{,}950$; however, n for the lake link and watershed tables is smaller. WS is lake watershed, NWS is network watershed, GNIS is Geographic Names Information System, WQP is Water Quality Portal, and NLA is the U.S. EPA's National Lakes Assessment program.

Lake characteristics: For each lake, this data table includes derived metrics for lake geometry, whether the lake is located within an area that was glaciated during the Late Wisconsin glaciation, and lake hydrologic connectivity. Geometry variables include lake area measured multiple ways (e.g., water only, water and island area combined), lake perimeter measured as lake only and with the perimeter of any islands within lakes (i.e., lake edge), and shape metrics such as shoreline development factor, which indicates shoreline complexity (Fig. 5).

Lakes are classified into one of six classes of hydrologic connectivity determined from the NHD network and distinguished by three features: the presence of inflows, presence of outflows, and connections to upstream lakes ≥ 10 ha in area. "Isolated" lakes lack both inflows and outflows; "Headwater" lakes have an outflow but no inflow; "Drainage" lakes have both inflow(s) and outflow(s) and lack upstream connected lake ≥ 10 ha in area; "DrainageLk" lakes have inflow(s) and outflow(s) and have upstream connected lake(s) ≥ 10 ha; "Terminal" lakes have inflows but no outflows and lack an upstream connected lake ≥ 10 ha in area; and "TerminalLk" lakes have inflow(s) but no outflow(s) and have an upstream connected lake(s) ≥ 10 ha (Fig. 6). Drainage and Isolated lakes are the most common connectivity classes while Terminal and TerminalLk classes are relatively rare (Fig. 7a–f). About a third of lakes have connectivity classes that fluctuate when only permanent stream connections are considered (Fig. 7g). Lakes were also classified as "glaciated" or "not glaciated" based on whether the lake is in an area that was glaciated during the Wisconsin glaciation that ended approximately 11,000 years ago (Ehlers et al. 2011); 30% of lakes are in formerly glaciated regions (Fig. 7h). For each lake, we calculated the area and number of upstream lakes connected by surface streams according to different size classes (≥ 1 ha, ≥ 4 ha, ≥ 10 ha) using the NHD network and considering both permanent and intermittent/ephemeral stream flow.

Lake watersheds: This data table includes identifiers, location, and geometry for the lake watersheds (WS) and network watersheds (NWS). Defining a lake watershed depends on the upstream surface water connections of a lake as well as the topography of the surrounding basin; the watershed boundaries do not include the focal lake. Each lake has a lake watershed (WS) that is one of three sub-types depending on the presence of lake inlets and upstream lakes ≥ 10 ha (Fig. 8). These sub-types are defined as:

a. *Local catchment (LC)*: The LC is the area of land that directly drains into a lake or a stream. Every lake and stream segment in LAGOS-US has a LC calculated for it, which is used to "accumulate" when creating the other two watershed types. For lakes that do not have any permanent upstream stream connections, the watershed accumulation process is complete and the LC represents the lake WS and is equivalent to the topographic watershed.
b. *Drainage-watershed (DWS)*: For lakes with upstream connections but no lakes ≥ 10 ha, we created a DWS by accumulating all upstream LCs. Thus, the DWS is the accumulation of the LCs for all upstream streams and small upstream lakes (< 10 ha).
c. *Inter-drainage-lake watershed (IDWS)*: For lakes with upstream connections with lakes ≥ 10 ha, we created an IDWS. These are similar to the DWS in that they include the area of land that directly drains into a lake and includes the area that drains into upstream-connected streams and

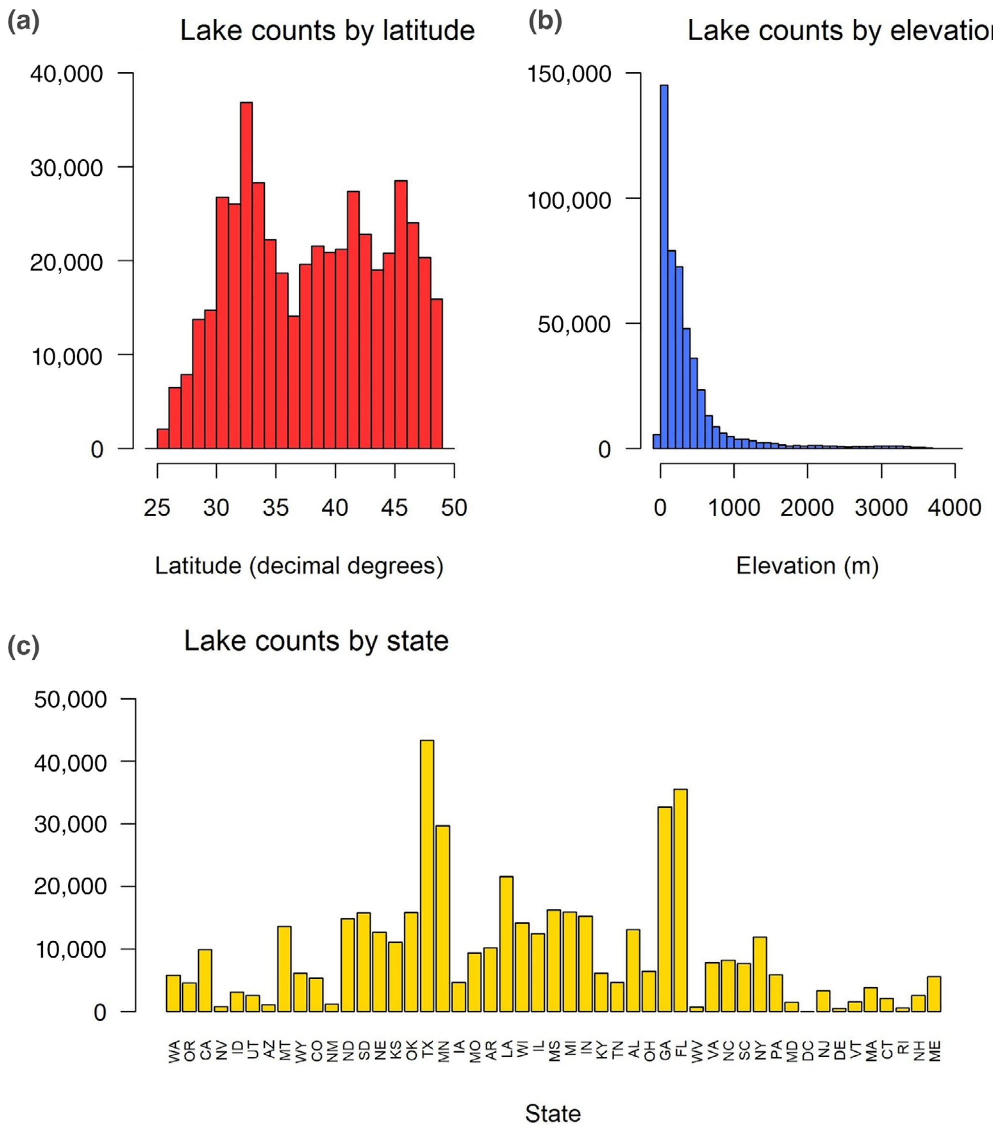

Fig. 4. Depiction of locational information found in the lake information table: (**a**) latitude, (**b**) elevation, and (**c**) U.S. state (lower 48 states and the District of Columbia) arranged generally from west to east and indicated with their standard two-letter code.

lakes < 10 ha. However, the IDWS accumulation "stops" at the outlet of any perennially connected upstream lakes ≥ 10 ha. Additionally, isolated and terminal lakes ≥ 10 ha are treated as sinks and are excluded from the focal lake IDWS.

In addition to the three sub-types of lake watersheds (WS) defined above, we also delineated a larger watershed for lakes with an IDWS, e.g., those with an upstream connected lake ≥ 10 ha. This watershed is called the network watershed (NWS) and is defined as the area of land that directly drains into a lake combined with the total area that drains into all upstream perennially-connected streams and lakes (including those ≥ 10 ha). Only DrainageLk and TerminalLk lakes have upstream lakes ≥ 10 ha and an NWS is quite large in some cases (Fig. 9a). To compile a set of total watersheds for all lakes, users can combine data for the NWS, derived for lakes with upstream lakes ≥ 10 ha, with that for the WS for those lakes that lack upstream lakes ≥ 10 ha. The flag *ws_equalsnws* provides a helpful variable to identify which lakes do or do not have an NWS.

For both WS and NWS, the number of observations reflects the dominance of small lakes and the relatively small proportion of lakes with network watersheds in LOCUS (Fig. 9a). While watershed area increases with lake area (Fig. 9b), an opposite decreasing pattern occurred for the ratio of watershed area to lake area; this pattern was flatter across lake area for WS compared to NWS watersheds (Fig. 9c).

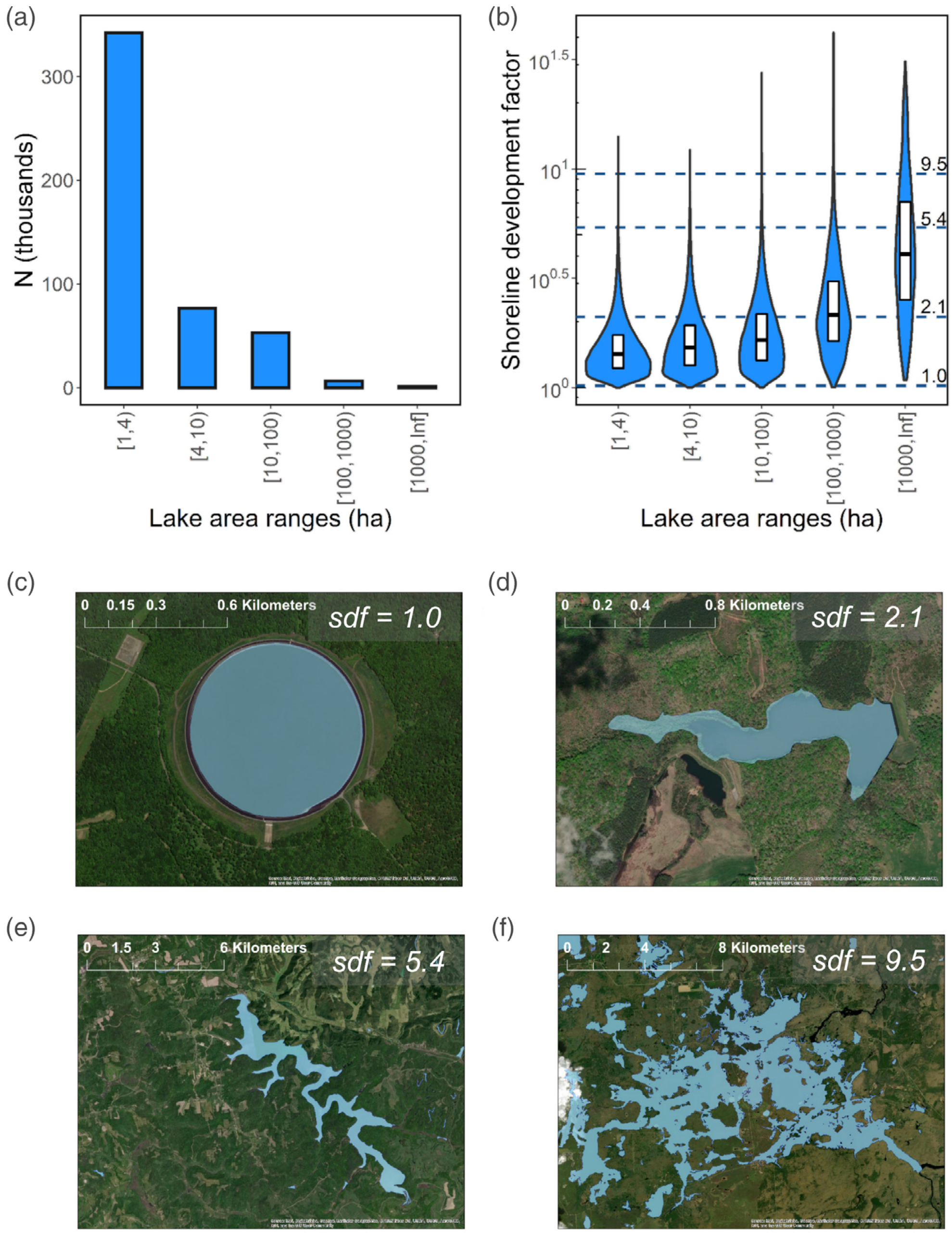

Fig. 5. Representative depictions of lake geometry. Number of lakes by lake area size bin (**a**) and violin plots of the shape metric "shoreline development factor" (SDF) plotted following $\log_{10}$ transformation (**b**). Violin plots show the kernel density distribution of shoreline development factor for each lake area class as well as the range of values. Embedded boxplots show the median value and the interquartile range (25th and 75th percentiles) of the $\log_{10}$ transformed data. Dashed reference lines show SDF values for panels **c**–**f**. The SDF is calculated as the ratio between the perimeter of a circle with area equal to the lake area and the measured perimeter. Lakes that are circular have an SDF approaching 1, while very reticulate lakes have a greater SDF. Shown below the plots are aerial images of lakes approximating the SDF values denoted in blue above: A very circular lake (c), an average lake (d), and two reticulated lakes (e and f). Image attribution: Esri, DigitalGlobe, GeoEye, i-cubed, USDA FSA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community.

Methods

Eleven main datasets were used to create LOCUS: NHD, High Resolution (USGS 2017*a*); NHDPlus High Resolution, Beta (USGS 2021*a*); NHDPlus v2.1, Medium Resolution (USGS 2019); National Elevation Dataset (USGS 2017*b*); Wisconsin Glaciation (Ehlers et al. 2011); TIGER/Line Boundaries (US Census Bureau 2017); Watershed Boundary Dataset, Hydrologic Unit (HU) boundaries (USGS 2016);

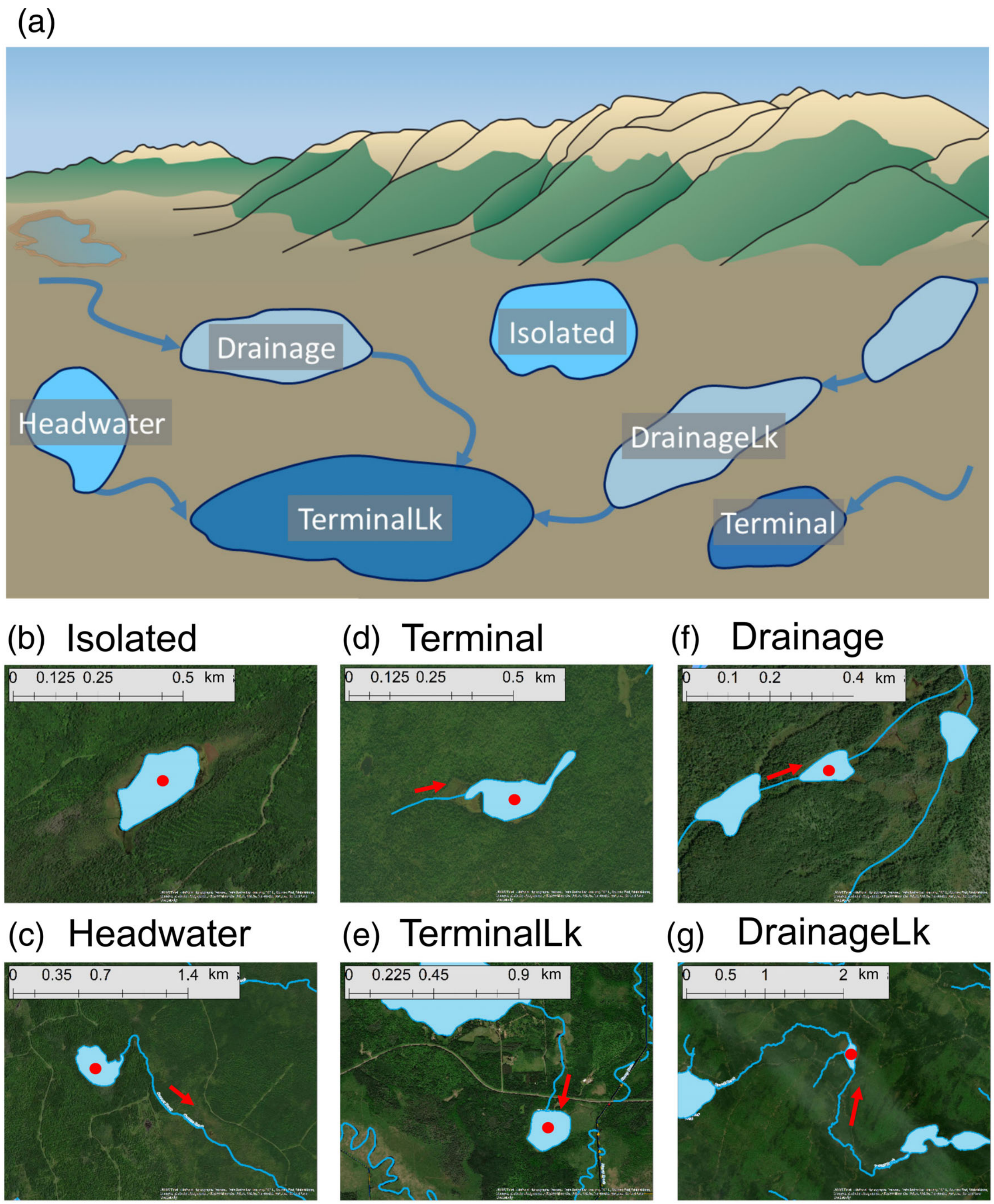

Fig. 6. Cartoon (top, (**a**)) and aerial images (**b–g**) depicting the six lake connectivity classes with focal lake marked with red dot and red arrows showing the direction of flow. Drainage and DrainageLk classes have inflow(s) and may or may not have outflow(s), with the DrainageLk also including one or more upstream lakes ≥ 10 ha; Headwater: outflow only; Terminal: inflow only, TerminalLk: inflow only also including one or more upstream lakes ≥ 10 ha; Isolated: no inflows or outflows. Image attribution: Esri, DigitalGlobe, GeoEye, i-cubed, USDA FSA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community. Symbols for diagrams courtesy of the Integration and Application Network (http://ian.umces.edu/symbols).

WQP (USGS et al. 2018); GNIS (USGS and US Board on Geographic Names 2018); US National Lakes Assessment 2007 and 2012 (USEPA 2010, 2016); and LAGOS-NE (LIMNO) v1.087.3 (Soranno et al. 2017; Soranno et al. 2019). Details of these sources and how they were used are included in the LOCUS User Guide (Smith et al. 2021) and the LAGOS-US GIS Toolbox (LAGOS_GIS_Toolbox v2.0 Beta).

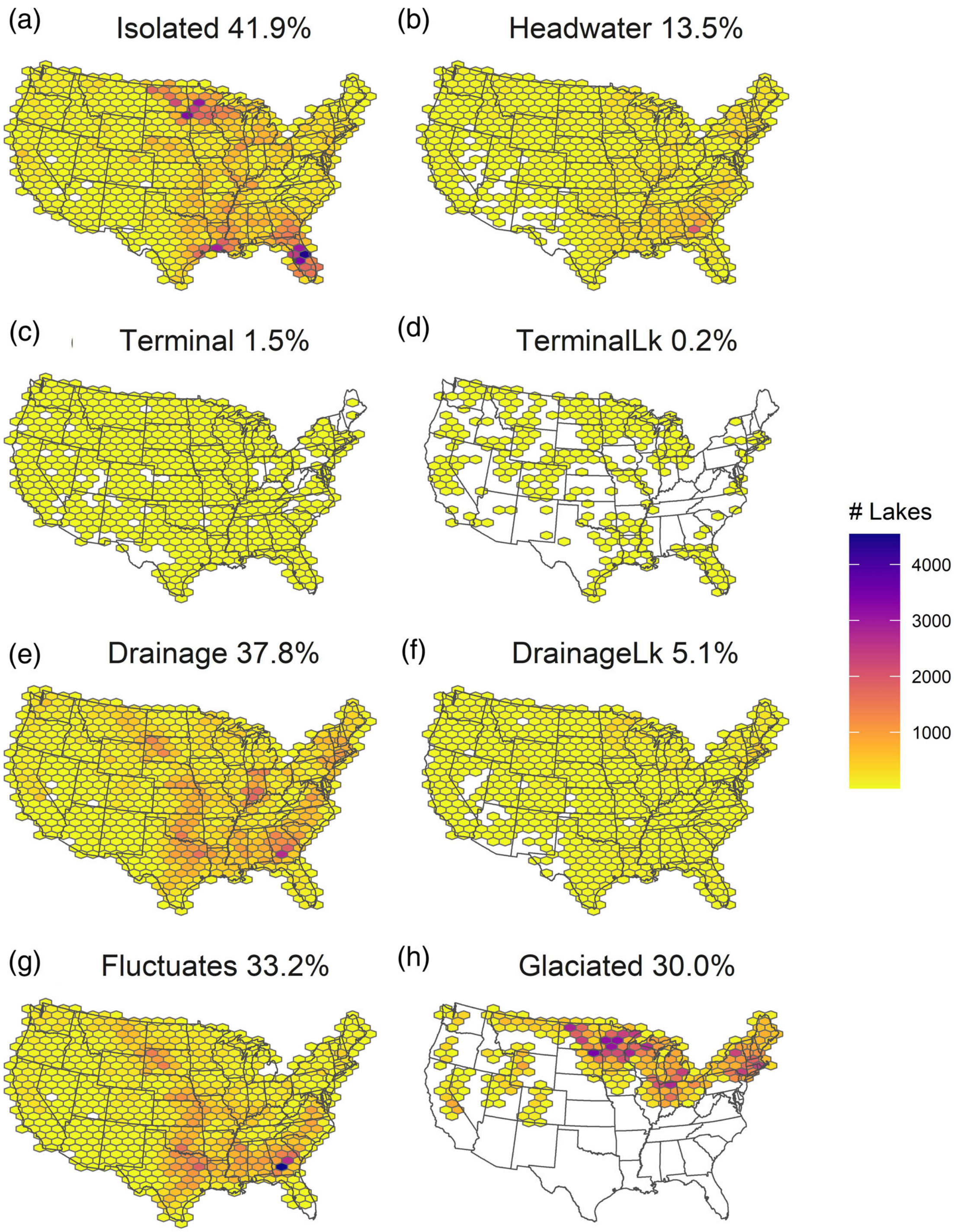

Fig. 7. Geographic distributions and lake densities mapped as hexagonal bin plots for the six lake connectivity classes (**a**–**f**); lakes with connectivity class fluctuations if only permanent stream connections are considered (**g**); and lakes located in regions that were glaciated during the Late Wisconsin glaciation (**h**). The percentages in the panel labels are the percent representation in the LOCUS lake population. *See* text and Fig. 6 for details about connectivity class types.

Metadata tables

LOCUS has a data source table (source_table_locus) and a data dictionary table (data_dictionary_locus). The source table includes official names, descriptions, citations, and other relevant metadata related to each of the source datasets in which variables from LOCUS were obtained. The data dictionary

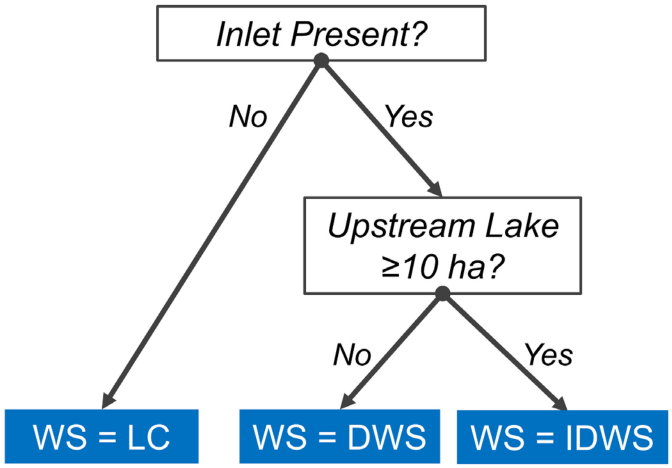

Fig. 8. A taxonomy of lake watershed sub-types. The decision tree indicates three scenarios for lake watershed (WS) delineation. Watershed subtypes are local catchment (LC), drainage-watershed (DWS), and inter-drainage-lake watershed (IDWS).

table includes official names, taxonomic information, units, and other relevant metadata related to each variable in LOCUS.

GIS vector data

GIS vector data consist of lakes as polygons and points, and watersheds as polygons. These spatial data are for future researchers to visualize the lakes and watersheds and to conduct further analyses that require spatial representation of lakes and watersheds. A complete polygon dataset is provided for all of the lake WS. The polygon dataset for the NWS includes a polygon for each lake that has the IDWS lake watershed sub-type (i.e., a subset of all lakes). A LC polygon dataset for lakes only (not streams/flowlines) is provided that can be combined with the publicly available NHDPlus HR Beta "catchment" GIS dataset. Methods for these vector data are detailed in the LAGOS-US GIS Toolbox and associated scripts (LAGOS_GIS_Toolbox v2.0 Beta).

Data tables

The four LOCUS data tables, described in detail below, contain the observations of variables that are grouped into tables based on similarity of data source, methods used to create the variable, or general category of variable: lake information, lake characteristics, lake link, and lake watersheds. These data tables also include individual observation-level flags for some variables, when applicable.

Data table: Lake information

The unique lake identifier (*lagoslakeid*) was generated as a sequential integer for each of the 479,950 lakes maintained in LAGOS-US: LOCUS. This *lagoslakeid* matches that in LAGOS-NE (an earlier version of the LAGOS research platform for 17 U.S. states; Soranno et al. 2017) when the NHD HR Permanent_Identifier used in the NE snapshot matched the Permanent_Identifier from the most recent NHD HR US snapshot. In a few rare cases, LAGOS-NE identifiers were completely removed from the lake population and were not recycled due to NHD identifier changes through time.

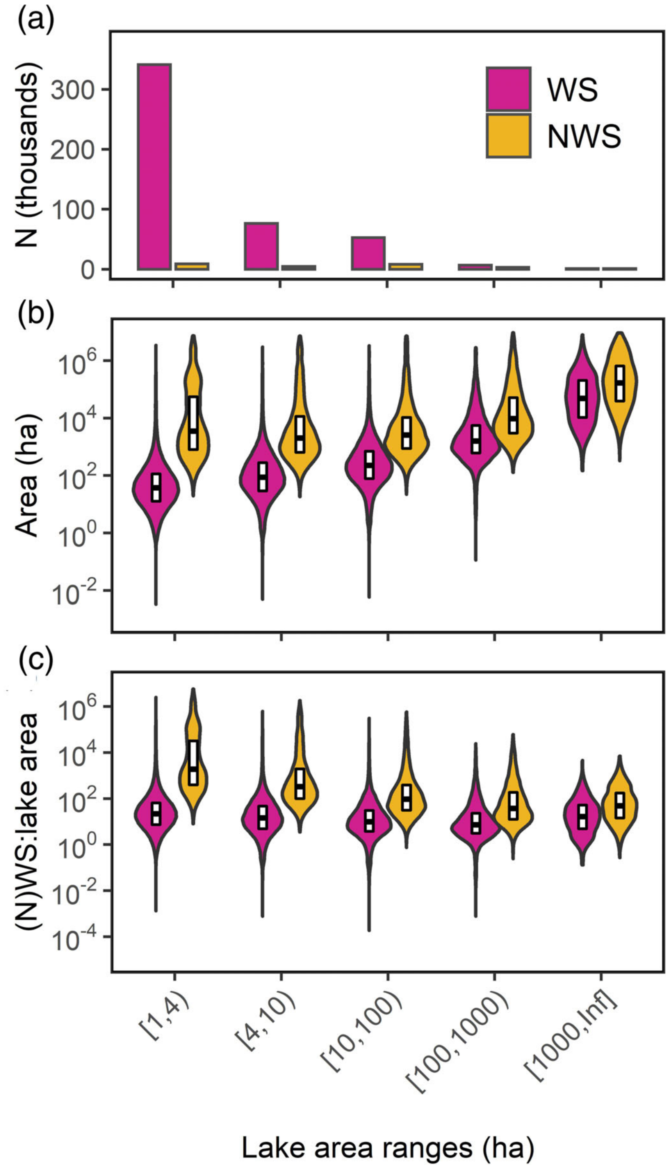

Fig. 9. Number in thousands (**a**), watershed area (**b**), and the watershed to lake water area ratio (**c**) for lake watersheds (WS) and lake network watersheds (NWS) across lake area ranges. Violin plots show the kernel density distribution and the range of values for each combination of watershed type and lake area interval. Embedded boxplots show the median value and the interquartile range (25th and 75th percentiles) of the data. Observations are binned by lake water area; for example, the [1,4) bin comprises lakes with water areas ≥ 1 and < 4 ha.

Official lake names were gathered from multiple data sources (e.g., NHD, GNIS) with significant manual review. All names were de-duplicated and concatenated using a semicolon delimiter to create *lake_namelagos*, so that multiple names could be provided for each lake, if available. Additional names were always linked with a lake if the (point) geographic location associated with the lake name from another data source was located within the LAGOS-US lake polygon. Additional names were sometimes linked if the location was within 100 m of the lake, contingent upon manual review to determine whether the point and lake polygon referenced the same lake. More names were identified manually when lake sampling programs provided unique names with the sampling observations that are part of a different LAGOS-US core module. Some names in LAGOS-US were manually identified using internet map services when reviewing lake locations. Names were also reused from LAGOS-NE in the following way: when lakes were larger than 50 ha, the LAGOS-NE name was validated before reuse; for smaller lakes, source text for names were used without further validation and names appear in upper case.

Several geographic location variables in lake_information were based on the point representation of lakes. The point representations are always within the lake polygon (a central "labeling" location) and are not based on the true centroid of the lake. Latitude and longitude for each lake were calculated with the Calculate Geometry tool in ArcGIS using the NAD83 geographic coordinate system (*lake_lat_decdeg*, *lake_lon_decdeg*) and lake surface elevation, which was based on the elevation value from the NED source dataset and was assigned to lakes using the Extract Values to Points tool in ArcGIS (*lake_elevation_m*).

The LOCUS lake_information table includes spatial division zone identifiers (e.g., zoneids) that connect lakes to the spatial division that contains their centers. There are many additional features of the zones that are provided as part of a different LAGOS-US core module. These 15 spatial divisions either: surround the lake shore (at buffer distances of 100 and 500 m); contribute water to the lake (both WS and NWS watershed types as described below); or reflect the location of the lake centroid within USGS nested hydrologic units (HU12 to HU4), political entities (county and state), or each of six widely used ecoregion or regionalization frameworks. *See* Smith et al. (2021) for a complete list of regionalization frameworks included and their citations. The unique identifier for the zone intersecting each lake point was recorded for all divisions (*hu4_zoneid*, **_zoneid*); if no zone intersected the point (as sometimes occurred along the U.S. border), the closest zone was used. All states that intersect the lake polygon were identified using the Spatial Join tool in ArcGIS and were recorded as the two-letter postal abbreviation(s) for each intersected state strung together into a single value separated by semicolons (*lake_states*).

The following variables were assigned a flag value of "Y" based on criteria described in depth in the data documentation (Smith et al. 2021): lake polygons intersecting the U.S. land border at Mexico or Canada (data source: TIGER) (*lake_onlandborder*), multipart lakes based on the shape's IsMultipart attribute calculated in ArcGIS (*lake_ismultipart*), and lakes without a *lagoslakeid* match in the lake_watersheds table (*lake_missingws*).

Data table: Lake link

Below, we provide a summary of the methods used to create the crosswalk table that links LAGOS-US with other broad-scale lake databases. Full documentation for the creation of the LAGOS-US lake_link table is available as an R notebook in the code repository. Additionally, detailed descriptions of the methods used are included in the LOCUS User Guide (Smith et al. 2021).

Six main datasets were used to create the lake_link table: LAGOS-US, GNIS, WQP, NHDPlusV2 (medium-resolution), NLA 2007 and 2012, and LAGOS-NE. There were several obstacles to reliably identifying common lakes among these datasets, including inadequate location accuracy, disagreement in the definitions affecting classification of waterbodies as lakes vs. another type of waterbody (e.g., swamps, streams, etc.), disagreement concerning the delineation of lake spatial extents or divisions between adjoining lakes, conflicting names, unaccounted changes in identifiers within a dataset, and asynchronous, unique update schedules for each dataset. Our methods were intended to overcome as many of these obstacles as possible, while keeping the number of false links to a minimum. This table was created in R, relying on the dplyr and sf libraries for data manipulation. For each dataset, the workflow consisted of seven steps as follows:

1. **Import**: modify fields and prepare the data frame for later work.
2. **Filter**: prune categories of entities that are less likely to represent either natural lakes or reservoirs, even if those categories sometimes connect to a LAGOS-US lake. The lake_link table only includes relationships if they ultimately connect back to a LAGOS-US lake.
3. **Convert**: move between spatial and non-spatial data formats, as needed. Spatial formats were projected to the USGS Albers Conic Equal Area projection (EPSG code 5070) for consistency.
4. **Select**: keep only necessary columns.
5. **Join**: individually connect LAGOS-US with each other dataset. This step included spatial joins, joins based on common identifiers, or the incorporation of links established or confirmed with a manual linking review process (for LAGOS-NE lakes and NLA lakes only).
6. **Select again**: keep only necessary columns.
7. **Final join**: merge the results of the multiple joins in step 5 into a single crosswalk table with only the necessary fields remaining.

Data table: Lake characteristics

The lake_characteristics table includes variables that characterize lake geometry, surface water connectivity, and glaciation. Below, we summarize the methods used for these variables and associated flags; full documentation of the methods used are included in the LAGOS-US LOCUS User Guide (Smith et al. 2021) and the LAGOS-US GIS Toolbox.

Lake geometry variables

The NHD HR lake polygons were the basis for calculating these metrics. Several variables describe lake geometry (i.e., area, perimeter, shape), and they were created with the Calculate Geometry tools in ArcGIS, using the Albers USGS Conical Area Projection.

Metrics include calculated shoreline development factor (SDF; *lake_shorelinedevfactor*; Fig. 5); water-only lake area (*lake_waterarea_ha*); total lake area including islands, if present (*lake_totalarea_ha*); area of islands, if present (*lake_islandarea_ha*); lake perimeter (*lake_perimeter_m*); and perimeter of islands, if present (*lake_islandperimeter_m*; Fig. 10a). Water-only lake area (*lake_waterarea_ha*) equals the area of the original polygon, whereas the remaining metrics were calculated by removing holes (islands) from the original lake polygons using ArcGIS Eliminate Polygon Part. Island area and perimeter metrics were calculated as differences between the area and perimeters of the original polygon and the hole-free polygons.

Other lake shape metrics reflect two types of minimum bounding (mbg) polygons based on (1) convex hulls and (2) rectangles (Fig. 10b). These polygons were generated in ArcGIS using the Minimum Bounding Geometry tool in the Data Management Toolbox/Features toolset and selecting the respective CONVEX_HULL or RECTANGLE_BY_AREA option. First, we used the minimum bounding geometry convex hull (defined as the smallest convex polygon enclosing the lake polygon; one way to envision this is as an elastic stretched around a polygon) to calculate three metrics: the maximum distance between any two vertices of the convex hull (*lake_mgbconhull_length_m*), the maximum width of the convex hull defined along an axis perpendicular to the axis defined by the convex hull length (*lake_mbgconhull_width_m*), and a lake orientation metric in degrees from 0 (or North) to 180° of the line defining the convex hull length (*lake_mbgconhull_orientation_deg*). Second, we used the minimum bounding geometry rectangle, which is defined as the smallest rectangle by area enclosing the lake polygon (Fig. 10b), to generate two metrics: rectangle length (*lake_mgbrect_length_m*) and rectangle width (*lake_mgbrect_width_m*).

To generate the *lake_shapeflag* that flags lake polygons with very elongate or angular shapes, we calculated the ratio between the water-only lake area and the area of the minimum bounding geometry rectangle (*lake_mbgrect_arearatio*). Elongate lakes are very thin relative to their length and might be considered closer to riverine systems (Fig. 10c). Angular lakes have shapes that closely match either a rectangle or triangle suggesting that they were human-made (Fig. 10c). Lakes that meet neither of these criteria were given a value of "noflag."

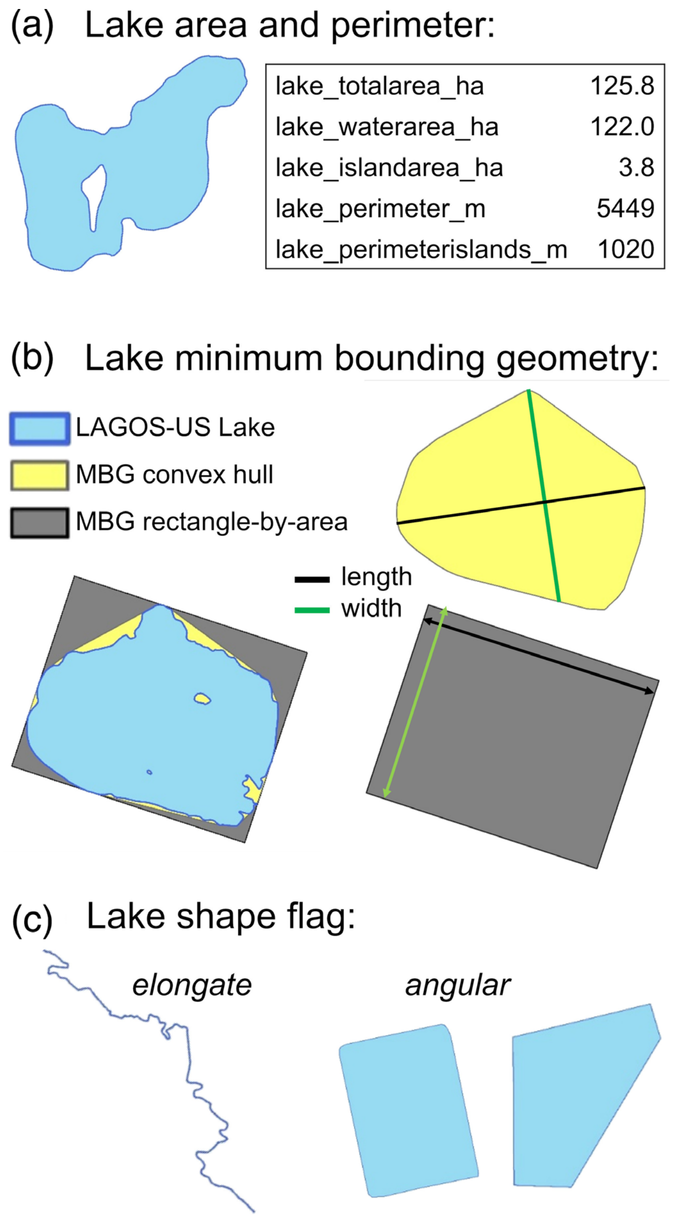

Fig. 10. Illustrations depicting lake geometry metrics and the lake shape flag. Specifically, (**a**) geometry values for a lake polygon with an island; (**b**) minimum bounding geometry (MBG) features with the original lake polygon (blue) depicted overlaid on the convex hull polygon (yellow) and both overlaid on the rectangle-by-area polygon (gray); (**c**) lakes with lake_shapeflag values equal to "elongate" and "angular." The angular lake examples shown illustrate criteria of low vertex counts (both) and shape approximating a rectangle (angular lake to the right).

Connectivity variables

We characterized lake surface water connectivity using a variety of metrics that share a unified methodology: lake connectivity class, upstream lake counts, and accumulation of LCs into the lake watersheds and network lake watersheds. We used network tracing implemented in a custom

Python tool (NHDNetwork, *see* LAGOS_GIS_Toolbox v2.0 Beta) and summarized the contents of the NHDFlow table in the NHD source data to make connectivity determinations. A trace represents the path(s) that starts at the origin point specified upon initiating the trace and follows the flow in the specified direction (upstream or downstream) until either defined barrier locations or the end of the network is reached. All metrics were calculated separately for each subregion (HU4-level river watershed) due to the way the NHD flow network data were packaged. The network connectivity metrics were calculated using the NHDPlus HR Beta 2021 snapshot, in correspondence with the watershed delineations.

Lake connectivity class: Lake connectivity classes were created using all lake network traces and were calculated in both flow directions for all lake and reservoir waterbodies in the NHDWaterbody layer. Maximum lake connectivity (*lake_connectivity_class*) was classified based on the following classes that represent six connection possibilities between lakes, streams, and upstream lakes ≥ 10 ha (Fig. 6): (1) Isolated—traces in both directions were empty (no network connectivity); (2) Headwater—only the downstream trace contained network connectivity; (3) DrainageLk—traces in both directions had network connectivity, and the upstream trace contains the identifier of one or more lakes ≥ 10 ha; (4) Drainage—all lakes that do not meet one of the prior three criteria; traces in both directions had network connectivity and the upstream trace did not contain any identifiers for lakes ≥ 10 ha; (5) Terminal—only the upstream trace contained network connectivity; and (6) TerminalLk—only the upstream trace contained network connectivity and the upstream trace contains the identifier of one or more lakes ≥ 10 ha.

These lake connectivity classes were based on flow traces of all upstream and downstream connectors. However, some of these connectors are considered ephemeral or intermittent. Therefore, permanent connectivity (*lake_connectivity_permanent*), defined as the connectivity class that persists throughout and between years, was assessed using the same rules as above after dropping stream segments labeled as either "intermittent" or "ephemeral" flow.

After calculating both maximum and permanent connectivity classes, if the two classes were unequal, the *lake_connectivity_fluctuates* flag was assigned the "Y" value to indicate the lake may have variable connectivity within or between years. For example, lakes that are isolated most of the year, but have ephemeral inflow after rainstorms are common in the arid southwest.

Upstream lake counts: The upstream network from each focal lake in the LAGOS-US lake population was traced and broken into three hierarchically nested size classes: all lakes ≥ 1 ha, all lakes ≥ 4 ha, and all lakes ≥ 10 ha. The count and area of connected lakes in each size class are summarized in six variables.

Glaciation variables

The ArcGIS Tabulate Intersection tool was used to calculate the percentage of each watershed or lake that was glaciated during the Wisconsin glaciation, terminating approximately 11,000 years ago (Ehlers et al. 2011). For lakes, if any percentage of the lake was glaciated, the "Glaciated" class was assigned; otherwise, the lake was designated as "Not Glaciated."

Table: Lake watersheds

The lake_watersheds table includes identifying information and variables associated with lake watersheds (Fig. 11). The delineation of lake watersheds was completed in two phases: (1) delineation of LCs and (2) accumulation of LCs into both watersheds and, where applicable, network watersheds. Watersheds were delineated by modifying NHDPlus HR files to add defined catchment outlets for all the lakes in the LAGOS-US lake population. This process allowed us to meet the goal of delineating a LC for each lake in the LAGOS-US lake population. Below, we summarize the methods used for delineating watersheds, calculating watershed geometry metrics, and assigning watershed flags; full documentation of the methods used are included in the LAGOS-US LOCUS User Guide (Smith et al. 2021), as well as the LAGOS-US GIS Toolbox and associated scripts (LAGOS_GIS_Toolbox v2.0 Beta).

Delineating watersheds

The watershed delineation process included two major steps—creating LCs for all waterbodies and then, accumulating the LCs to create watersheds for each lake. We describe these two major steps at a high level here (*see* User Guide for details; Smith et al. 2021). LCs were delineated by modifying the published NHDPlus HR intermediate watershed delineation data products to emphasize the LAGOS-US lakes as catchment outlets while retaining the value added by the NHDPlus HR product standards.

We reused or modified the following datasets published in the NHDPlus HR geodatabase and raster files: NHDPlus NHDPlusIDGridCode, NHDPlusSink, catseed.tif, hydrodem.tif, filldepth.tif, and fdr.tif. The modifications implemented in LAGOS-US were intended to ensure each LAGOS-US lake would have a catchment delineated, regardless of surface water connectivity. We also mitigated an acknowledged error with excessive small sinks in the NHDPlus HR Beta version.

Processing conducted by the source data providers: The NHDPlus HR watershed delineation process is documented in detail by Moore et al. (2019), and implements these major steps:

1. Mosaic the National Elevation Dataset 3DEP DEMs to cover the extent of the HU4 (subregion) and convert elevation units to centimeters.

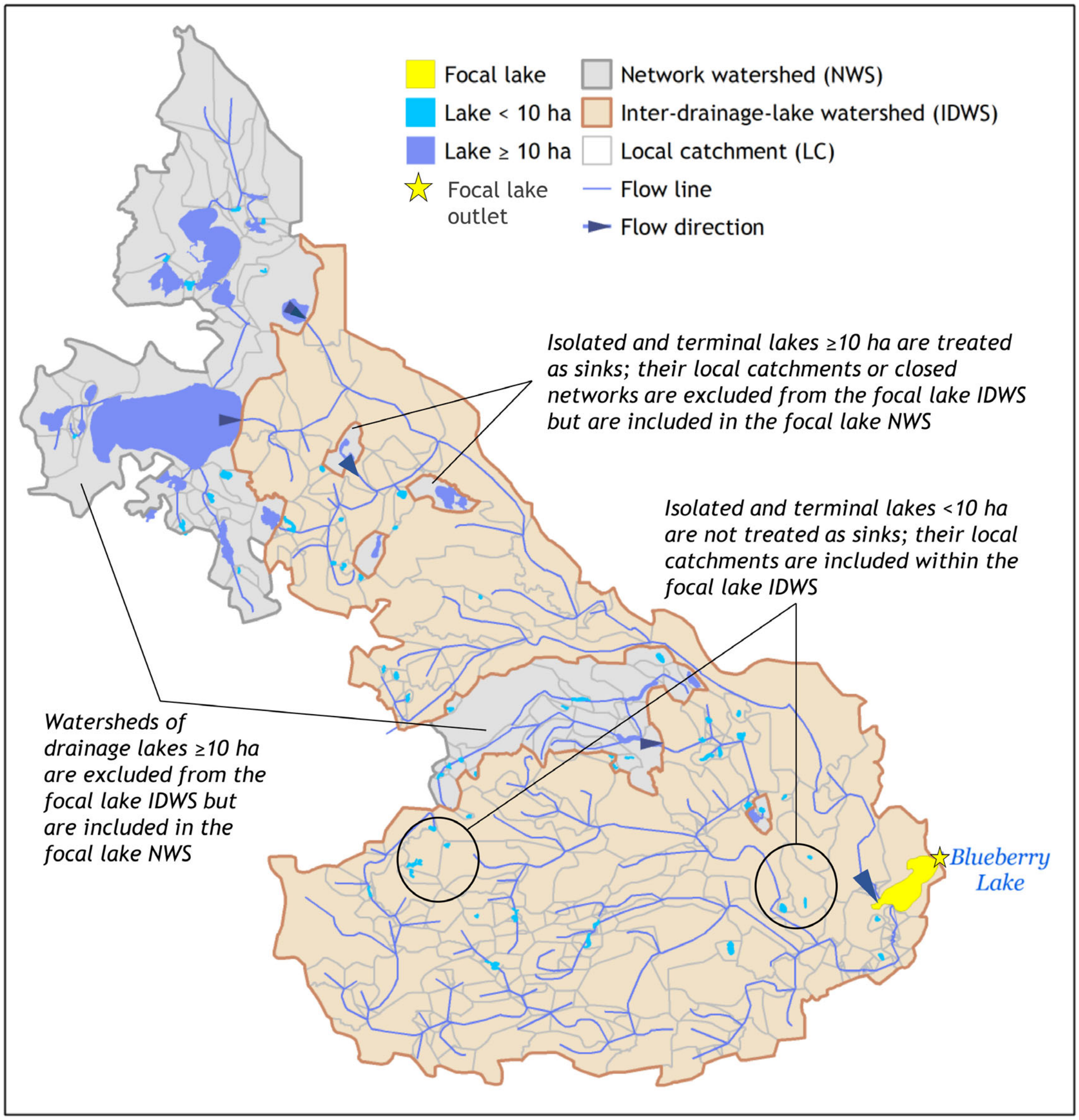

Fig. 11. Description of watershed definitions that are based on different stream and lake connections in LAGOS-US. *See* text for details and Fig. 8 for a decision tree explaining the watershed taxonomy. Note, that our operational minimum lake size is 1 ha (*see* Technical validation).

2. Pre-process and clean up the NHD vector features in preparation for burning features into the DEM raster. Examples of pre-processing are excluding pipelines and ditches with uninitialized flow from the network, trimming headwater flowlines that conflict with WBD, and identifying valid sinks to burn. (NHDPlusSink).
3. Use the results of step 2 to burn the network flowlines and waterbodies into the DEM with a beveled drop following the AGREE methodology (Hellweger and Maidment 1997). This process forces alignment between the DEM and the vector network.
4. Add walls at HU12 boundaries with cuts for flowlines/streams.
5. Fill sinks in the landscape so the associated regions can drain into the flow network. Those sinks defined as valid in prior steps are protected as not fillable. Save the fill depth as a separate raster. (hydrodem.tif, filldepth.tif).
6. Calculate eight-direction flow using the hydro-conditioned DEM and clip the flow direction raster to the HU4 boundary. (fdr.tif).
7. Define catchment pour points from the burned flowline features using the GridCode to translate identifiers to raster pixel values. (catseed.tif, NHDPlusNHDPlusIDGridCode).

Delineating LCs: To create a LC for every lake, we reused and modified intermediate data products in the NHDPlus HR to build on their processing while also enforcing lakes being handled as sinks in the network. We then repeated the processes of filling sinks, calculating flow direction, and delineating catchments in the same fashion as was used to create the NHDPlus HR to maximize geographic alignment with that data product aside from the influence of our focus on the lakes. We used these steps to enforce the modifications:

1. Reconstitute the unfilled hydro-enforced DEM by subtracting the cell values of filldepth.tif from those in hydrodem.tif.
2. Ensure burning all LAGOS-US lakes into the hydro-enforced DEM.
3. Protect isolated and terminal LAGOS-US lakes as valid, unfillable sinks in order to delineate LCs for these waterbodies.
4. Fill excessive small sinks ("NHDWaterbody closed lake" not included in LAGOS-US population) in the hydro-conditioned DEM that were indicated as erroneously permitted in the beta version of the NHDPlus HR.
5. Modify the pour points raster catseed.tif to use all LAGOS-US lakes as primary catchment pour points, prevent artificial paths associated with lakes from being used as pour points, and remove any pour points associated with the excessive small sinks.
6. Calculate eight-direction flow from the modified hydrodem.tif and clip the output to the same HU4 boundary as the published fdr.tif.
7. Delineate catchments using the modified catseed.tif and new fdr.tif.

Delineating lake WS and lake NWS by accumulating LCs: For each lake in LOCUS, the NWS was delineated as an accumulation of all qualifying upstream LCs by tracing the entire upstream network without defining any flow barrier points to control the tracing. The traces were calculated using the same tool (NHDNetwork) that was used to calculate lake connectivity metrics. If the focal lake had no upstream network (i.e., isolated and headwater lakes), the NWS delineated is the same as the focal lake's LC. If the focal lake had upstream network connectivity, the focal lake's LC plus all the upstream catchments were aggregated, dissolved into a single polygon, and the single polygon had all holes removed. The focal lake polygon itself (not its catchment) was erased from its own watershed.

To accumulate the lake WS delineations, we used a workflow with three major steps: upstream tracing with barriers (i.e., lakes ≥ 10 ha), dissolving watershed polygons to remove sinks aside from those belonging to lakes ≥ 10 ha, and erasing regions excluded by definition (those with flow sunk into lakes ≥ 10 ha). The second two steps of this workflow were necessary to accommodate the geometric operations needed to sink flow into any non-focal lakes with and without outlets ≥ 10 ha, while also incorporating flow from smaller sinks into the focal lake's watershed. More details about these steps, as well as the erasing step are available in the User Guide (Smith et al. 2021).

Watershed geometry variables

Minimum bounding watershed geometry-based variables were calculated using ArcGIS 10.5 in the same manner as the corresponding lake geometry variables described above, using the convex hull and the rectangle-by-area. The minimum bounding geometry rectangle is defined as the smallest rectangle by area enclosing the WS or NWS polygon. The following WS and NWS variables were generated from the convex hull polygons: the maximum possible distance between any two vertices of the convex hull (**_mbgconhull_length_m*), the maximum width of the convex hull defined along an axis perpendicular to the axis defined by the convex hull length (**_mbgconhull_width_m*), and degrees from 0 (or North) of the line defining the convex hull length (**_orientation_deg*). The following variables were calculated from the minimum bounding geometry rectangle: the length (**_mbgrect_length_m*) and the width of the rectangle (**_mbgrect_width_m*). Watershed and network watershed/lake area ratios (**_lake_arearatio*) were derived with Calculate Field in ArcGIS (using **_focallakewaterarea_ha* and **_area_ha*). Watershed and network mean width (**_meanwidth_m*) were derived with Calculate Field in ArcGIS using **_area_ha* x 10,000 /**_mbgconhull_length_m*).

Watershed flags

Several watershed and watershed geometry variables are associated with flags, some of which are informational whereas others are cautionary. For example, if the difference between the area of the WS and the NWS is effectively zero (< 10 square meters), then *ws_equalsnws* was set to "Y." If the focal lake's upstream trace did not include any connectivity, *ws_subtype* was set to "LC" (local catchment). If the focal lake had upstream connectivity and the *ws_equalsnetwork* flag was "N," then *ws_subtype* was set to "IDWS" (inter-drainage-lake watershed). Otherwise, the flag was set to "DWS" (drainage-watershed).

We also provide flags for watersheds that cross boundaries or that have multiple parts. For example, watersheds and network watersheds with their polygon intersecting the U.S. land border at Mexico or Canada (data source: TIGER) or the oceanic or Great Lakes coasts were flagged "Y" on the respective **_onlandborder*, **_oncoast* flags. Multipart watersheds and network watersheds (**_ismultipart*) were identified in ArcGIS using the IsMultiPart attribute and flagged with a "Y."

Occasionally, very skinny "sliver" WS were produced that appeared to be artifacts from geoprocessing in very flat terrain or other difficult settings (Fig. 12). To identify slivers (*ws_sliverflag*), we created an index of how much a watershed polygon deviated from a rectangular shape as the ratio between the WS or NWS area and the area of the

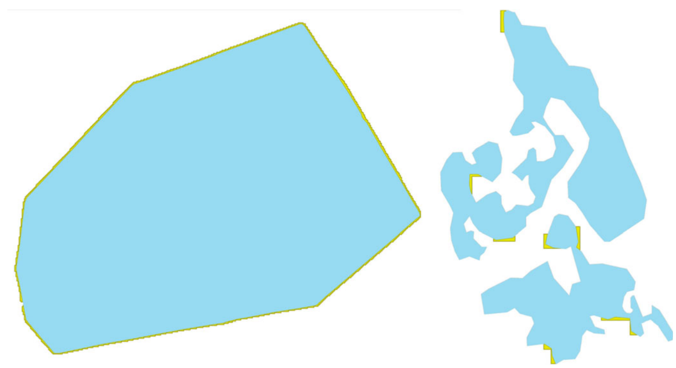

Fig. 12. Example of watersheds flagged with *ws_sliverflag* = "Y." Lake polygon in blue; watershed polygon in yellow.

corresponding minimum bounding geometry rectangle ([*nws* or *ws*] *mbgrect_arearatio*). The criteria for assigning a *ws_sliverflag* were arrived at by trial and error and based on multiple criteria (*see* User Guide for details).

Technical validation

We used existing data sources to create the many variables that are in LAGOS-US LOCUS. We either copied those data directly or we created new derived variables by either processing single variables or combining more than one variable using different functions. Because all variables are based on underlying public datasets with their own published technical validation, we do not repeat it here. Instead, we describe the many steps that we took throughout our workflow to ensure data quality for the variables in LOCUS.

Note that as is often the case for such integrated, large, and complex datasets, it was not our goal to definitively differentiate between "good" or "bad" data during technical validation. Instead, we validated our methods or data to: (1) ensure that our processing was doing what was intended, (2) ensure that there were no errors in accounting or data transfer, and (3) identify where there may be potential data concerns that may limit some future data users (but not all). Our validation approach used automated procedures because the large volume of data in LOCUS could not be processed and evaluated manually. We created observation-level flags that indicate some possible concerns or special information for specific data values that future users may want to consider when using those data values. The most critical aspect of our workflow was to use reproducible and transparent methods wherever possible. We achieved this goal by minimizing manual processing (i.e., point and click approaches) by using scripts, incorporating data checks at multiple points, conducting a final data check on all variables including visualization and mapping, making our scripts available for other users, and detailed methodological documentation in extensive documentation that is published with our data.

Next, we describe the technical processing used to determine the minimum lake size for our lake population. Then, we provide a high-level description of our data validation that was a critical part of our workflow. We refer the reader to the methods section of this article, as well as the detailed technical documentation that is provided with the dataset for additional details (Smith et al. 2021).

Determining the minimum lake size threshold

The census population of lakes, by definition, represents all lakes in the conterminous U.S. Therefore, we took steps to ensure that our lake data source (NHD HR), was not biased according to lake size. Due to resolution issues, small lakes could be missed in the digitization process when creating the NHD HR. Additionally, the NHD is edited on a regional basis, which could result in differences such that some states might digitally represent small lakes, whereas others might not. Therefore, we selected a minimum lake size that was adequately represented across the conterminous U.S. in the NHD HR to define our lake census population. Next, we describe the steps that we took to make this determination and to make the final decision of a 1 ha minimum lake size for LAGOS-US.

We created a lake size error validation dataset to examine the effect of lake size on the probability of being represented in the NHD HR. We selected nine states representing different physiographic regions of the U.S. (FL, ND, CA, AL, AZ, TX, WA, WI, IA) to quantify whether lakes were more likely to be represented in the NHD HR if they were larger. We compared lakes in the NHD HR with the highest resolution aerial imagery available (generally from the USDA's National Agriculture Imagery Program; NAIP). We selected five sampling regions per state, except for CA where we selected six regions due to its large size, to manually count lakes in the NHD HR and aerial imagery. Lakes that were present in imagery but were not in the NHD HR were roughly digitized. Finally, we calculated error rates and each lake was assigned an error status. The 43 sampling regions contained 1074 waterbodies identified as lakes in the NHD and 137 lakes identified in the imagery but not included in the NHD, for a total of 1211 lake candidates included in the assessment. Of these, 996 were identified as lakes by both the NHD and our assessment.

We found that the proportion of "correct" lakes in the NHD HR varied by lake surface area, but that there was no definitive threshold size class (Fig. 13a). Lakes ≥ 9 ha were over 91% correct and lakes ≥ 10 ha were 95% correct; however, because there are hundreds of thousands of lakes smaller than 9–10 ha (Fig. 13b), we decided to accept a higher misclassification rate. Lakes that are 1–2 ha in size have a 77% correct rate; however, if we combine the entire lake size error validation dataset, then the overall correct rate is 82%. In comparison, using a 2, 3, or 4 ha minimum lake size results in overall correct rates of 86%, 87%, and 87%, respectively. In fact, there was little evidence for a clear threshold (Fig. 13a).

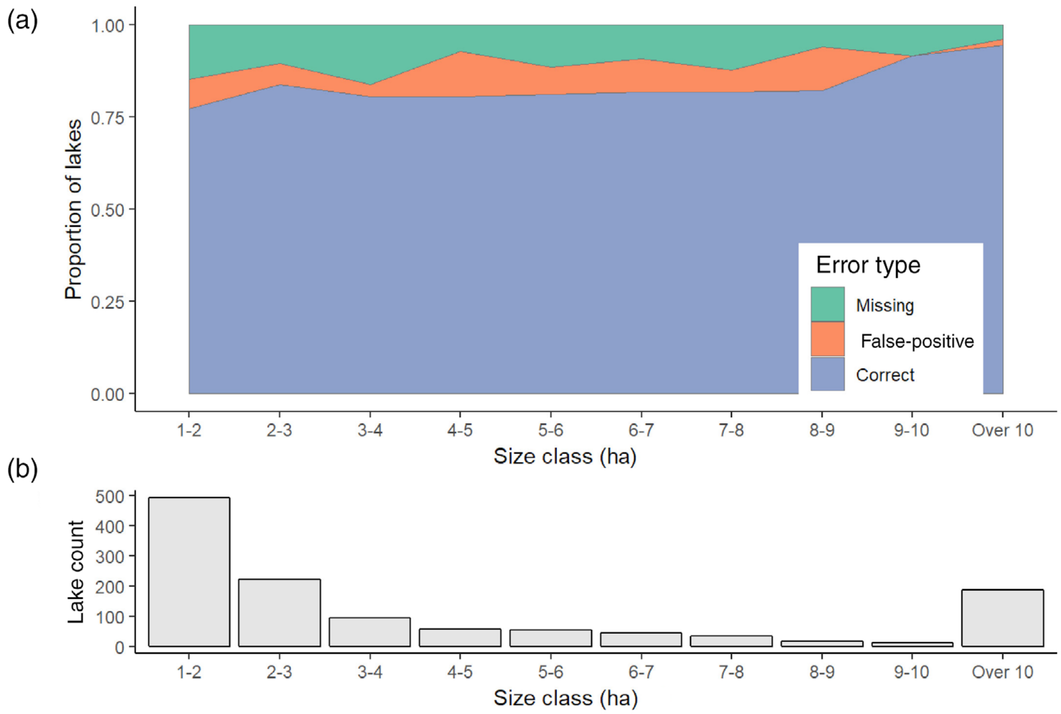

Fig. 13. Results from the error analysis using the 1074 lakes in the "lake size error validation dataset." The proportion correct, the proportion of false-positives, and the proportion of missing lakes by lake surface area class (size class, ha) (**a**). The number of sampled lakes by size class in the "lake size error validation dataset" (**b**).

We determined that a minimum size of 1 ha would include the largest lake population with a reasonable error rate. Therefore, we include lakes ≥ 1 ha in LAGOS-US LOCUS census lake population.

Data validation

The data check step was intended to ensure that the procedures used to create the feature values for LOCUS resulted in the intended outcomes. We queried each of the feature values generated by geoprocessing to identify potential data or geoprocessing issues and verify that data values are sensible (e.g., are within expected ranges and expected completeness of data). These checks were performed using semi-automated R scripts that imported each feature individually to run it through a companion R markdown script that queried the data, ensured comparability with the source GIS layer and data dictionary, summarized values and features (e.g., numbers of missing data, data that were negative, zeroes), produced maps of variables, and automatically generated scores for five main evaluation criteria (i.e., match with GIS data, match with metadata, percent composition sums check, spatial completion check, and missing values check). This entire process was documented in a full technical validation summary report in html format.

Data limitations

Due to the large number of lakes in LOCUS, we automated processing of lake and watershed polygons and the attributes provided by the NHD source data. Prior to data analysis, users should always consult the observation-level flags we created for the issues described above. In addition to these flags, we are aware of three other potential limitations. First, we recognize that the decision to only represent lakes ≥ 1 ha in LOCUS may exclude locally important lakes smaller than this threshold. There are likely regional differences in the presence of small lakes, as well as the perceived importance of them. For example, using lake names in the GNIS dataset, more lakes < 1 ha are named in the Southwest U.S. compared to the Upper Midwest and Northeast U.S., suggesting their regional importance. Second, our watershed delineations do not account for altered hydrological inputs (e.g., irrigation) that may be important influences on surface flowpaths in some cases. Third, because the foundation of LOCUS is the NHD HR, any limitations of the NHD HR apply to LOCUS. For example, the NHD HR is known to have delineation errors for some waterbodies, particularly those that fluctuate in depth (and thus, shape and perimeter) through time. Because LOCUS is based on a snapshot in time, it does not represent these fluctuations, nor does it include any updates made to the NHD HR following the download of our snapshot. However, we were able to fix one known issue with the NHDPlus

Table 1. Open data and software available for studying lakes at the scale of the conterminous U.S. as of 2020, arranged according to the type of lake data. All products are for the conterminous U.S. except for LAKEMORPHO, which is for the entire U.S. Res = reservoirs, chars = characteristics, wshed = watershed, obs = observations, repo = repository, cont = continuously, pub = publication, NA = not available. Variable for NHD products indicates that the number of lakes and minimum lake size vary depending on the definition of lake used.

SHORT NAME, full name and/or description	Water body type	Type of lake data	Total number of lakes	Min lake size (area, ha)	Product type	Lake depth info?	Year(s) created	Unique identifier in LOCUS lake_link?	Dataset spatial resolution	Cite***
LAGOS-US LOCUS, LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds	Lakes, res	Lake chars, wshed chars, and geospatial	479,950	1	Data package	No	2017–2021	Yes	1:24,000	This study*
LAGOS-US RESERVOIR, LAGOS-US RSVR v1.0: Data module classifying lakes as natural lakes or reservoirs	Lakes, res	Lake chars	137,465	4	Data package	No	2019–2021	Yes (same as in LAGOS-US)	1:24,000	In development*
RMD, reservoir morphology database for reservoirs greater than 101 ha	Res	Lake chars	3828	101	Data package	Yes	2016	No (but in LAGOS-US RSVR)	1:24,000	Rodgers 2017†
NHD high res, National Hydrography Dataset	Rivers, streams, lakes, res	Lake chars	Variable	Variable	Data package	No	Cont updated	Yes	1:24,000	§§§
NHDPlusV2, geospatial hydrologic framework dataset	Rivers, streams, lakes, res	Lake chars	Variable	Variable	Data package	No	2012	Yes	1:100,000	¶¶¶
NHDPlus HR, geospatial hydrologic framework dataset	Rivers, streams, lakes, res	Lake chars	Variable	Variable	Data package	No	2012–present	Yes	1:24,000	¶¶¶
LAGOS-US NETWORKS, LAGOS-US NETS v1.0: Data module of surface water networks characterizing connections among lakes, streams, and rivers	Lakes res	Lake connectivity	86,511	1	Data package	No	2020	Yes (same as in LAGOS-US)	1:24,000 and 1:100,000	King et al. in press
HYDROLINKS, characterizing hydrologic networks: Developing a tool to enable research of macroscale aquatic networks	Lakes res	Lake connectivity	6,500,264	NA	R package	No	2018	NA	NA	Winslow et al. 2018§

(Continues)

Table 1. Continued

SHORT NAME, full name and/or description	Water body type	Type of lake data	Total number of lakes	Min lake size (area, ha)	Product type	Lake depth info?	Year(s) created	Unique identifier in LOCUS lake_link?	Dataset spatial resolution	Cite***
LAGOS-US LAKE DEPTH, LAGOS-US DEPTH v1.0: Data module of observed mean and maximum lake depths for a subset of lakes	Lakes res	Lake depth and morphometry	17,700 with max depth, 9560 with mean depth	1	Data package	Yes	2019	Yes (same as in LAGOS-US)	NA	In development*
LAKEMORPHO, lakemorpho: Calculating lake morphometry metrics in R	Lakes res	Lake depth and morphometry	NA	NA	R package	No	~2016	NA	NA	Hollister and Stachelek 2017¶
LAGOS-US GEO, LAGOS-US GEO v1.0: Data module of geospatial ecological context at multiple spatial and temporal scales for lakes and their watersheds	Lakes res	Lake wsheds and beyond, geospatial	479,950	1	Data package	No	2017–2021	Yes (same as in LAGOS-US)	1:24,000	In development*
LakeCat, Lake and catchment dataset characterizing landscape features for lake basins	Lakes res	Lake wsheds and beyond	378,000	0.1	Data package	Yes‡‡‡	~2017	Yes	1:100,000	Hill et al. 2018**
LAGOS-US LIMNO, LAGOS-US LIMNO v1.0: Data module of in situ limnological physical, chemical, and biological measurements through time for lakes	Lakes res	In situ obs	Unknown	1	Data package	No	2017–2021	Yes (same as in LAGOS-US)	1:24,000	In development*
NLA-2007, National Lake Survey-2007 of the National Aquatic Resource Surveys	Lakes res	In situ obs	1157	4	Data package	Yes	2007–2012	Yes	NA	Pollard et al. 2018††
NLA-2012, National Lake Survey-2012 of the National Aquatic Resource Surveys	Lakes res	In situ obs	1287	1	Data package	Yes	2012–2016	Yes	NA	Pollard et al. 2018‡‡
NES, The National Eutrophication Survey: Lake characteristics and historical nutrient concentrations	Lakes res	In situ obs	775	6	Data package	Yes	1972–1975	No	NA	Stachelek et al. 2018§§

(Continues)

Table 1. Continued

***SHORT NAME*, full name and/or description**	**Water body type**	**Type of lake data**	**Total number of lakes**	**Min lake size (area, ha)**	**Product type**	**Lake depth info?**	**Year(s) created**	**Unique identifier in LOCUS lake_link?**	**Dataset spatial resolution**	**Cite*****
WQP/WQX, Water Quality Portal: National Water Quality Monitoring Council	All	In situ obs	Unknown†††	NA	Data repo	Some	NA	Yes (lake and res sites)	NA	Read et al. 2017¶¶
NWIS, National Water Information System: USGS	All	In situ obs	Unknown†††	NA	Data repo	Some	NA	No	NA	****
AQUASAT, matchup database of remote sensing reflectance values and in situ observations	Lakes res, streams, estuaries	In situ obs	Unknown†††	NA	Data package	No	2019	Yes	NA	Ross et al. 2019
LAGOS-US LANDSAT, LAGOS-US LANDSAT v1.0: Data module of remote sensing reflectance values and modeled in situ data	Lakes res	In situ obs	137,465‡‡‡	4	Data package	No	2020–2021	Yes (same as in LAGOS-US)	NA	In development*
LIMNOSAT-US, database of remote sensing reflectance values and modeled in situ data	Lakes res	In situ obs	56,792‡‡‡	10	Data package	No	2020	No	NA	Topp et al. 2021

*https://lagoslakes.org/
†https://pubs.er.usgs.gov/publication/ds1062
‡https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution
§https://cran.r-project.org/web/packages/hydrolinks/index.html
¶https://cran.r-project.org/web/packages/lakemorpho/lakemorpho.pdf
**https://www.journals.uchicago.edu/doi/abs/10.1086/697966
††https://www.epa.gov/national-aquatic-resource-surveys/national-lakes-assessment-2007-results
‡‡https://www.epa.gov/national-aquatic-resource-surveys/national-lakes-assessment-2012-results
§§https://www.earth-syst-sci-data.net/10/81/2018/
¶¶https://www.waterqualitydata.us/
***Footnotes point readers to a weblink for more information, when available.
†††Organized by site, not by water body.
‡‡‡Modeled, not measured.
§§§https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution
¶¶¶http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php
****https://waterdata.usgs.gov/nwis

HR at the time of our download—that of extra sinks not connected to the hydrological network. We removed these sinks prior to creating lake watersheds and incorporated several rules in our watershed accumulation algorithms that checked for these and other sinks.

Data use and recommendations for reuse

We are making these data available as soon as they have been completed; therefore, just two articles have been published using a small part of this data module to date—lake names in the conterminous U.S. (Soranno et al. 2020) and lake point and polygon GIS vector files (McCullough et al. 2019). Because this data module includes all lakes and their watersheds in the conterminous U.S., characteristics of both lakes and watersheds, and GIS files, it is likely that this data module will be used as a base dataset for many future macroscale or continental-scale analyses of lakes. For example, LOCUS can be combined with the other two LAGOS-US core modules (LIMNO and GEO), the four in-progress extension modules (DEPTH, NETWORKS, RESERVOIR, LANDSAT), as well as with other data sources (*see* below). Finally, we have designed LOCUS, and all of LAGOS-US, so that users can develop additional extension modules (gray boxes and dashed lines in Fig. 2).

Comparison with existing datasets

This data module fills gaps in pre-existing open datasets. We provide watersheds for all lakes, in addition to a GIS data layer for future users to conduct their own analyses on all lakes and watersheds in the conterminous U.S. We have summarized all open data national-scale lake databases for the U.S. (Table 1) that will be helpful for future researchers conducting research in the conterminous U.S. or compiling data by country for a global analysis. No single database is likely to have everything a researcher conducting macroscale research on lakes might want. An examination of the gaps in Table 1 inspired us to create the LOCUS lake_link table that connects these different datasets through a common unique identifier (*lagoslakeid*), which will facilitate future macroscale lake studies. We also include key information that influences a researcher's ability to harmonize these different datasets, such as minimum lake size, the inclusion of lakes and reservoirs, and the total number of lakes in the dataset. A similar table could be constructed for global lake datasets; however, there are far fewer such datasets that include the range of variables in LOCUS or the datasets in Table 1. Furthermore, there are few global datasets that include all lakes, particularly those as small as 1 ha. An important future direction is to contribute to global studies by building the capacity to create these types of macroscale databases for all countries.

Code URL with permanent identifier

The code for the GIS processing to create many variables in the dataset is referred to as the *LAGOS GIS Toolbox* and is available in the data and metadata repository above.

Measurement(s)

Geographic location, boundaries, area, and other geometries.

Technology type(s)

NA

Temporal range

2010–2019.

Frequency or sampling interval

Snapshot.

Spatial scale

Continental.

References

Ehlers, J., P. L. Gibbard, and P. D. Hughes. 2011, *Quaternary glaciations-extent and chronology: A closer look*, v. **15**. Elsevier; [accessed 2019 February 21]. Available from https://booksite.elsevier.com/9780444534477/digital_maps.php.

Eidenshink, J., B. Schwind, K. Brewer, Z.-L. Zhu, B. Quayle, and S. Howard. 2007. A project for monitoring trends in burn severity. Fire Ecol. **3**: 3–21. doi:10.4996/fireecology.0301003

Hellweger, F., and D. Maidment. 1997, *AGREE-DEM surface reconditioning system*. Austin: Univ. of Texas, [accessed 2019 April 18]. Available from https://www.caee.utexas.edu/prof/maidment/gishydro/ferdi/research/agree/agree.html.

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

Hollister, J., and J. Stachelek. 2017. Lakemorpho: Calculating lake morphometry metrics in R. F1000 Research **6**: 1718. doi:10.12688/f1000research.12512.1

Huot, Y., C. A. Brown, G. Potvin, D. Antoniades, H. M. Baulch, B. E. Beisner, S. Bélanger, S. Brazeau, H. Cabana, J. A. Cardille, P. A. del Giorgio, I. Gregory-Eaves, M.-J. Fortink, A. S. Lang, I. Laurion, R. Maranger, Y. T. Prairie, J. A. Rusak, P. A. Segura, R. Sironq, J. P. Smol, R. D. Vinebrooke, and D. A. Walsh. 2019. The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate. Sci. Total Environ. **695**: 133668. doi:10.1016/j.scitotenv.2019.133668

King, K., Q. Wang, L. K. Rodriguez, and K. S. Cheruvelil. in press. Lake connectivity networks for the conterminous U.S. (LAGOS-US NETWORKS v1). Limnol. Oceanogr. Lett.

Kuhn, C., A. de Matos Valerio, N. Ward, L. Loken, H. Oliveira Sawakuchi, M. Kampel, J. Richey, P. Stadler, J. Crawford, R. Striegl, E. Vermote, N. Pahlevan, and D. Butmanan. 2019. Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-*a* and turbidity. Remote Sens. Environ. **224**: 104–118. doi:10.1016/j.rse.2019.01.023

Mantzouki, E., and others. 2018. A European multi lake survey dataset of environmental variables, phytoplankton pigments and cyanotoxins. Sci. Data 5: 180226. doi:10.1038/sdata.2018.226

Marcarelli, A.M. and Wayne A.W. 2007. Effects of upstream lakes and nutrient limitation on periphytic biomass and nitrogen fixation in oligotrophic, subalpine streams. Freshwater Biology **52**: 2211–2225. doi:10.1111/j.1365-2427.2007.01851.x

McCullough, I., K. S. Cheruvelil, J.-F. Lapierre, N. Lottig, M. Moritz, J. Stachelek, and P. A. Soranno. 2019. Do lakes feel the burn? Ecological consequences of increasing exposure of lakes to fire in the continental United States. Glob. Chang. Biol. **25**: 2841–2854. doi:10.1111/gcb.14732

Michael, T.B. and Mark, M.B. 2008. A review and reassessment of lake phosphorus retention and the nutrient loading concept. Freshwater Biology **53**: 194–211. doi:10.1111/j.1365-2427.2007.01862.x

Moore, R. B., L. D. McKay, A. H. Rea, T. R. Bondelid, C. V. Price, T. G. Dewald, and C. M. Johnston. 2019. User's guide for the National Hydrography dataset Plus (NHDPlus) high resolution: U.S. Geological Survey Open-File Report 2019-1096. doi:10.3133/ofr20191096

Pollard, A. I., S. E. Hampton, and D. M. Leech. 2018. The promise and potential of continental-scale limnology using the U.S. Environmental Protection Agency's National Lakes Assessment. Limnol. Oceanogr. Bull. **27**: 36–41. doi:10.1002/lob.10238

Read, E. K., L. Carr, L. DeCicco, H. A. Dugan, G. Hanson, J. A. Hart, J. Kreft, J. S. Read, and L. A. Winslow. 2017. Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resour. Res. **53**: 1735–1745. doi:10.1002/2016WR019993

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rodgers, K. D. 2017. *A reservoir morphology database for the conterminous United States*. U.S.G.S. doi:10.5066/F7GF0RQZ. Available from https://www.sciencebase.gov/catalog/item/587fb41ae4b085de6c11f38b.

Ross, M. R., S. N. Topp, A. Appling, X. Yang, C. Kuhn, D. Butman, M. Simard, and T. M. Pavelsky. 2019. AquaSat: A dataset to enable remote sensing of water quality for inland waters. Water Resour. Res. **55**: 10012–10025. doi:10.1029/2019WR024883

Smith, N. J., K. E. Webster, L. K. Rodriguez, K. S. Cheruvelil, and P. A. Soranno. 2021. LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. Ver 1. Environmental Data Initiative. doi:10.6073/pasta/e5c2fb8d77467d3f03de4667ac2173ca

Soranno, P.A., N.R. Lottig, A.D. Delany, and K.S. Cheruvelil. 2019. LAGOS-NE-LIMNO v1.087.3: 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 ver 3. Environmental Data Initiative. doi:10.6073/pasta/08c6f9311929f4874b01bcc64eb3b2d7

Soranno, P.A., and others. 2017. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. GigaScience **6**: 1–22. doi:10.1093/gigascience/gix101

Soranno, P.A., K. E. Webster, N. J. Smith, J. D. Vázquez, and K. S. Cheruvelil. 2020. What is in a "lake" name? That which we call a lake by any other name. Limnol. Oceanogr. Bull. **4**: 183–192. doi:10.1002/lob.10355

Stachelek, J. 2021. LAGOSUS: Interface to the lake multi-scaled geospatial and temporal database. R package version 0.0.1.

Stachelek, J., C. Ford, K. King, H. Miller, and R. Nagelkirk. 2018. The National Eutrophication Survey: Lake characteristics and historical nutrient concentrations. Earth Syst. Sci. Data **10**: 81–86. doi:10.5194/essd-10-81-2018

Thorpe, A. S., D. T. Barnett, S. C. Elmendorf, E.-L. S. Hinckley, D. Hoekman, K. D. Jones, K. E. Levan, C. L. Meier, L. F. Stanish, and K. M. Thibault. 2016. Introduction to the sampling designs of the National Ecological Observatory Network Terrestrial Observation System. Ecosphere **7**: e01627. doi:10.1002/ecs2.1627

Topp, S. N., T. M. Pavelsky, H. A. Dugan, X. Yang, J. Gardner, and M. R. V. Ross. 2021. Shifting patterns of summer lake color phenology in over 26,000 US lakes. Water Resour. Res. **57**: e2020WR029123. doi:10.1029/2020WR029123

U.S. Census Bureau. 2017. TIGER/line shapefiles (machine-readable data files), [accessed 2017 May 2]. Available from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-geodatabase-file.html.

U.S. Environmental Protection Agency. 2010. National Aquatic Resource Surveys: National Lake Assessment 2007, [accessed 2020 May 10]. Available from https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys.

U.S. Environmental Protection Agency. 2016. National Aquatic Resource Surveys: National Lake Assessment 2012, [accessed 2020 May 10]. Available from https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys.

U.S. Geological Survey. 2017*a*. National Hydrography Dataset (ver. USGS National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU) 4-2001), [accessed 2017 January 5]. Available from https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products.

U.S. Geological Survey. 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)), [accessed 2017 April 24]. Available from https://ned.usgs.gov/.

U.S. Geological Survey. 2019. National Hydrography Dataset (NHDPlus version 2.1 for hydrologic unit (HU) 4-2001 (published 20191002)), [accessed 2019 October 23]. Available from https://www.epa.gov/waterdata/nhdplus-national-data.

U.S. Geological Survey. 2021. National Hydrography Dataset (ver. USGS NHDPlus High Resolution (HR) Beta for Hydrologic Unit (HU) 4-2001), [accessed 2021 January 21]. Available from https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products.

U.S. Geological Survey. 2021. National Water Information System (USGS Water Data for the Nation), [accessed July 19, 2021]. Available from http://waterdata.usgs.gov/nwis.

U.S. Geological Survey & U.S. Board on Geographic Names. 2018. Geographic Names Information System (GNIS), [accessed 2018 March 19]. Available from https://geonames.usgs.gov/docs/stategaz/NationalFile_20180201.zip.

U.S. Geological Survey, U.S. Department of Agriculture-NRCS, & U.S. Environmental Protection Agency. 2016. National Watershed Boundary Dataset (WBD), [accessed 2016 December 5]. Available from ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/Hydrography/WBD/National/GDB.

U.S. Geological Survey, U.S. Environmental Protection Agency, & U.S. National Water Quality Monitoring Council. 2018. Water Quality Portal (WQP), [accessed 2018 February 14]. Available from https://waterservices.usgs.gov/rest/Site-Test-Tool.html.

Winslow, L. A., T. H. Hahn, T. H. Leach, and K. C. Rose. 2018. Characterizing hydrologic networks: Developing a tool to enable research of macroscale aquatic networks. Environ. Monit. Softw. **104**: 94–101. doi:10.1016/j.envsoft.2018.03.012

Zhang, T., Soranno, P.A., Cheruvelil, D.B. Kramer, M.T. Bremigan, A. Ligmann-Zielinska. 2012. Evaluating the effects of upstream lakes and wetlands on lake phosphorus concentrations using a spatially-explicit model. Landscape Ecology **27**: 1015–1030. doi:10.1007/s10980-012-9762-z

Acknowledgments

We thank Austin Delany, Corinna Gries, Arnab Shuvo, and Noah Lottig for their help with the data module. We also thank Allie Shoffner for her review of earlier drafts and the entire Continental Limnology Project team for contributions and discussions throughout the development of this data module. Thanks to the following individuals who are the lead authors of the following LAGOS-US extension module data products Jem Stachelek (LAKE DEPTH), Katelyn King (NETWORKS), Patrick Hanly (LANDSAT), and Sam Polus (RESERVOIR). Grant sponsor information: DEB-1638679 (MSU); DEB-1638539 (PSU); DEB-1638554 (UW); DEB-1638550 (Missouri).

Submitted 28 November 2020
Revised 26 June 2021
Accepted 28 June 2021