LIMNOLOGY AND OCEANOGRAPHY
Letters

ASLO
Open Access

Limnology and Oceanography Letters 2023
© 2023 The Authors. *Limnology and Oceanography Letters* published by Wiley Periodicals LLC
on behalf of Association for the Sciences of Limnology and Oceanography.
doi: 10.1002/lol2.10299

DATA ARTICLE

LAGOS-US RESERVOIR: A database classifying conterminous U.S. lakes 4 ha and larger as natural lakes or reservoir lakes

Lauren K. Rodriguez[1,2] *Sam M. Polus*[1] *Danielle I. Matuszak*[1] *Marcella R. Domka*[1] *Patrick J. Hanly*[1]* *Qi Wang*[3]
Patricia A. Soranno[1] *Kendra S. Cheruvelil*[1,4]

[1]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan; [2]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, Maryland; [3]Computer Science and Engineering, Michigan State University, East Lansing; [4]Lyman Briggs College, Michigan State University, East Lansing, Michigan

Scientific Significance Statement

Naturally formed lakes differ from human-made lakes (i.e., reservoirs) in many ways. Although well distributed across the globe, it has not been possible to classify lakes into these two broad categories across broad geographic scales due to the lack of detailed information on lake origins. As a result, there has been no regional- to continental-scale data source that differentiates between natural lakes (NLs) and reservoirs, except for the very largest of lakes globally. LAGOS-US RESERVOIR v1, a research-ready data module, fills this gap with a machine learning model-based classification for all 137,465 U.S. conterminous lakes ≥ 4 ha. These data facilitate the macroscale study of both reservoirs and NLs, which is needed to better quantify and understand the role of surface water in global cycles and to test conventional wisdom about how NLs and reservoirs differ from each other across the broad scales that represent their full diversity.

Abstract

The LAGOS-US RESERVOIR data module classifies all 137,465 lakes ≥ 4 ha in the conterminous U.S. into three categories using a machine learning predictive model based on visual interpretation of lake outlines and a lake shape classification rule. Natural Lakes (NLs) are defined as naturally formed, lacking large, flow-altering structures; Reservoir Class A's (RSVR_A) are defined as lakes likely human-made or human-altered by a large water

*Correspondence: hanlypat@msu.edu

Associate editor: Bridget R. Deemer

Lauren K. Rodriguez, Sam M. Polus, Danielle I. Matuszak, and Marcella R. Domka indicate joint first authors listed in reverse alphabetical order.

Author Contribution Statement: SMP led and conducted the manual classification effort that created the datasets for testing and building the model. QW designed and tested the machine learning model, and PAS designed the overall approach. All authors framed the paper, wrote the Data Description section, and provided critical reviews of the entire manuscript. DIM, MRD, and KSC wrote the Background and Motivation; SMP, QW, LKR, and PJH wrote the Methods; PAS and PJH wrote the Technical Validation; and DIM, MRD, KSC, and PAS wrote the Data Use and Recommendations for Reuse section. Figures were created as follows: Figs. 2, 3, 8, 11, 12 by LKR; Fig. 1 by KSC; Figs. 4, 5 by DIM with contributions by PJH; Fig. 6 by MRD with contributions by PJH; Figs. 7, 9 by SMP; and Figs. 10, 13 by PJH. Tables and Box were created as follows: Tables 1–3 by LKR with contributions by PJH; Table 3 and Box 1 by DIM with contributions by KSC and PAS. PJH revised all figures, data, and metadata during the revision process. KSC and PAS provided intellectual, supervisory, and collaborative writing leadership for the data module.

Data Availability Statement: Data and metadata are available on EDI: https://portal-s.edirepository.org/nis/metadataviewer?packageid=edi.804.1. Code for the predictive models, raw image files, and figure code are viewable at https://doi.org/10.5281/zenodo.5584528.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

control structure; and Reservoir Class B's (RSVR_Bs) are lakes likely human-made but are not connected to streams and have a shape rare in NLs. We trained machine learning models on 12,162 manually classified lakes to predict assignment as an NL or RSVR, then further classified RSVRs based on NHD Fcodes, isolation, and angularity. Our classification indicates that > 46% of lakes ≥ 4 ha in the conterminous U.S. are reservoir lakes. These data can be easily combined with other LAGOS-US modules and U.S. national databases for the broad-scale study of reservoir lakes and NLs.

Background and motivation

Both NLs and reservoirs provide important and often distinct ecosystem services to humans (Lehner et al. 2011). For reservoirs, water-control structures direct water supply for irrigation, facilitate flood control, aid navigation, create hydropower, and increase tourism, fisheries, or recreation (Thornton et al. 1990; Lehner et al. 2011; Doubek and Carey 2017; Mamun et al. 2020). In the United States, most reservoirs are less than 90 years old (Thornton et al. 1990) and as water control structures have aged, some are now being removed (Habel et al. 2020). As new dams continue to be constructed and old reservoirs altered, there have been calls for dam construction and removal to be based on the most up to date and reliable data available and for a balance between sustainability of fresh waters and human needs (Lehner et al. 2011). For example, although gross emissions of greenhouse gases (GHGs) from reservoirs may account for 1.5% of the global warming potential of GHGs (Deemer et al. 2016), reservoirs also have the potential to positively influence human-impacted nutrient cycles. Harrison et al. (2009) showed the importance of including small NLs and reservoirs in models predicting the global nitrogen removal by surface waters and Tranvik (2009) found that carbon burial rates in reservoirs could be one to two orders of magnitude higher than in NLs. Therefore, accurately quantifying the role of both NLs and reservoirs on such global cycles requires accurately differentiating between these two types of lakes and accurately documenting their location and numbers at broad spatial scales (Lehner et al. 2011).

Defining lakes and reservoirs

Although reservoir lakes are overwhelmingly understudied compared to NLs, there are some studies that document differences between these two types of lakes (Doubek and Carey 2017). For example, compared to NLs, reservoirs tend to be warmer in temperature (Thornton et al. 1990), have larger watershed sizes that are heavily influenced by both nutrient and sediment runoff from their surrounding agricultural landscapes (Knoll et al. 2003), and have larger ratios of basin to lake/reservoir surface area (Lehner and Döll 2004; Doubek and Carey 2017). However, an important recent study found that many differences between NLs and reservoirs depended on latitude, making clear differences between them difficult to quantify (Doubek and Carey 2017). In addition, reservoirs vary in both form and function and range from run-of-the-river high-flow reservoirs to very still and less-connected reservoirs, including entirely artificial water bodies with very angular shapes. Given this diversity, there are many definitions of "reservoir" and little standardization. For example, relatively simple descriptions include "engineered systems" (Thornton et al. 1990), "[hu]man made lakes" (Lehner and Döll 2004), and "constructed impoundments" (Doubek and Carey 2017). However, what constitutes *human influence* is subjective and other characteristics in addition to the presence of an artificial construct or impoundment are important for differentiating reservoirs from other lakes. For example, reservoirs can be characterized by their position and placement within the river network, outlet control presence and type (i.e., water control structure or dam), and origin (Hayes et al. 2017). There is also a variety of dam types and sizes that influence reservoir shapes and sizes. In addition to the challenges in defining reservoirs, there is also a lack of data available to classify lakes as either natural or human-made. Such a classification requires information on the water control structure, including its presence, hydrologic location, height, management, and history. As a result, existing reservoir datasets are mainly for either very large water bodies, for water bodies with very large dams, or for individual reservoirs studied for long time periods (e.g., Birkett and Mason 1995; Lehner 2011). There is currently no classification system differentiating between NLs and reservoir lakes at broad spatial extents and for smaller lake, which severely limits the regional to continental study of reservoirs and their comparisons to NLs.

Although there is yet to be a single agreed-upon and used reservoir definition, we use the definition described in Hayes et al. (2017) as a starting point and refine it based on data that is widely available at broad spatial extents—lake outlines and shape—to indicate function (Box 1). Furthermore, similar to the definition of "lake" used in the LAGOS-US data platform (Cheruvelil et al. 2021), we used the generic term "lake" to refer to both NLs and reservoir lakes in this database and data paper (Box 1).

Data description

The LAGOS-US RESERVOIR v1 data module and its associated User Guide (Polus et al. 2022) fills the above data gaps by classifying all lakes greater than or equal to 4 ha in the conterminous U.S. ($n = 137,465$) into one of three classes: natural lake (NL), reservoir lake class A (RSVR-A), or reservoir lake class

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

BOX 1. Definitions of lake, natural lake, and reservoir used for the LAGOS-US RESERVOIR data module.

Lake—A perennial body of relatively still water with a geographically defined polygon in the high-resolution National Hydrography Dataset (NHD) that is 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). Lakes that are extremely high-intensity and artificial as indicated by the NHD, such as sewage treatment points, aquaculture ponds, and retention ponds, are not included (Cheruvelil et al. 2021).

The above definition applies to all 137,465 lakes ≥ 4 ha in the conterminous U.S. Then, we divided all lakes into one of three categories based on model interpretation of lake outlines and a metric of lake shape that indicates the angularity of the lake using the definitions below. Although it would be preferred to base the classification on lake function and the exact degree of human modification, such detailed information is not available on the hundreds of thousands of lakes in the conterminous U.S. (or globally). Therefore, these definitions are based on a machine learning model that predicts the probability of a lake being human-made versus natural using lake shape. To further differentiate reservoirs, we used a metric of angularity of lake outlines. The three categories in this data module and their definitions are as follows:

Natural lake (NL)—A lake that is likely to be entirely or mostly naturally formed and that does not have a relatively large, flow-altering structure on it or near it based on visual interpretation of imagery. Such lakes may have a small human-made water-control structures on it that appear to be physically small relative to the size of the lake shoreline, or that are downstream of the lake and so are assumed to have minimal impacts on the lake, such as those that can influence water levels only.

Reservoir Class A (RSVR_A)—A lake that is likely to be either human-made or highly human-altered by the presence of a relatively large water control structure that appears to significantly change the flow of water based on a machine learning model prediction with lake outlines as model input.

Reservoir Class B (RSVR_B)—A lake that is likely to be entirely human-made based on a highly angular shape that is rarely, if ever, seen in natural lakes. Angularity is defined as a shape that nearly conforms to a rectangle, defined as the ratio between the lake area and the area of the minimum bounding rectangle area that is close to 1 (Smith et al. 2021). Most angular lakes are also not connected to other water bodies through stream connections (i.e., isolated) and do not fit traditional definitions of reservoirs as dammed rivers. These lakes have been defined as impoundments located outside of river networks by Hayes et al. (2017).

B (RSVR-B). The lake polygons, locations, and identifiers were obtained from the LAGOS-US LOCUS data module (Smith et al. 2021), which obtained the lake base layers from the NHD snapshot that was downloaded January 2017 (NHD; USGS 2017). RESERVOIR provides model-based prediction probabilities for all three lake classes, and characteristics of lakes such as location, geometry, and lake hydrologic connectivity. RESERVOIR can be linked with other LAGOS-US modules using the common lake identifier *lagoslakeid*. RESERVOIR consists of one data table that includes observation-level flags, two metadata tables (source table, data dictionary table; Fig. 1), a polygon layer of all lakes labeled by class, and a detailed User Guide (Polus et al. 2022).

Overview of model approach

We manually classified 12,162 lakes (natural [NL] and reservoir [RSVR] lakes), then used those to train a machine learning model to predict the whether all lakes ≥ 4 ha in the conterminous U.S. were either natural or reservoir lakes based on lake shape. We used lake shape as a determining factor based on the conventional wisdom that reservoir lakes are more dendritic than NLs, which has some recent support from the literature (Doubek and Carey 2017). This process resulted in the machine learning classification of 77,667 NLs and 59,798 RSVRs along with prediction probabilities associated with each lake being one of these two classes. We then further refined this classification using information from the NHD Fcodes, isolation, and angularity. The final classification resulted in 73,053 NLs, 61,042 RSVR_As, and 3370 RSVR_Bs (Fig. 2). RESERVOIR also includes information such as location, lake shape, surface water connectivity class, and lake name. This macroscale dataset of both large and small NLs and reservoir lakes is designed to be combined with other LAGOS-US data modules and national databases using unique lake identifiers to allow for the study of reservoirs at the regional to conterminous U.S. scale.

Metadata tables

LAGOS-US RESERVOIR has a source table (*source_table_rsvr*) and a data dictionary table (*data_dictionary_rsvr*). The source table includes official names, descriptions, citations, and other relevant metadata related to each of the source datasets in which variables from RESERVOIR were obtained. The data

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

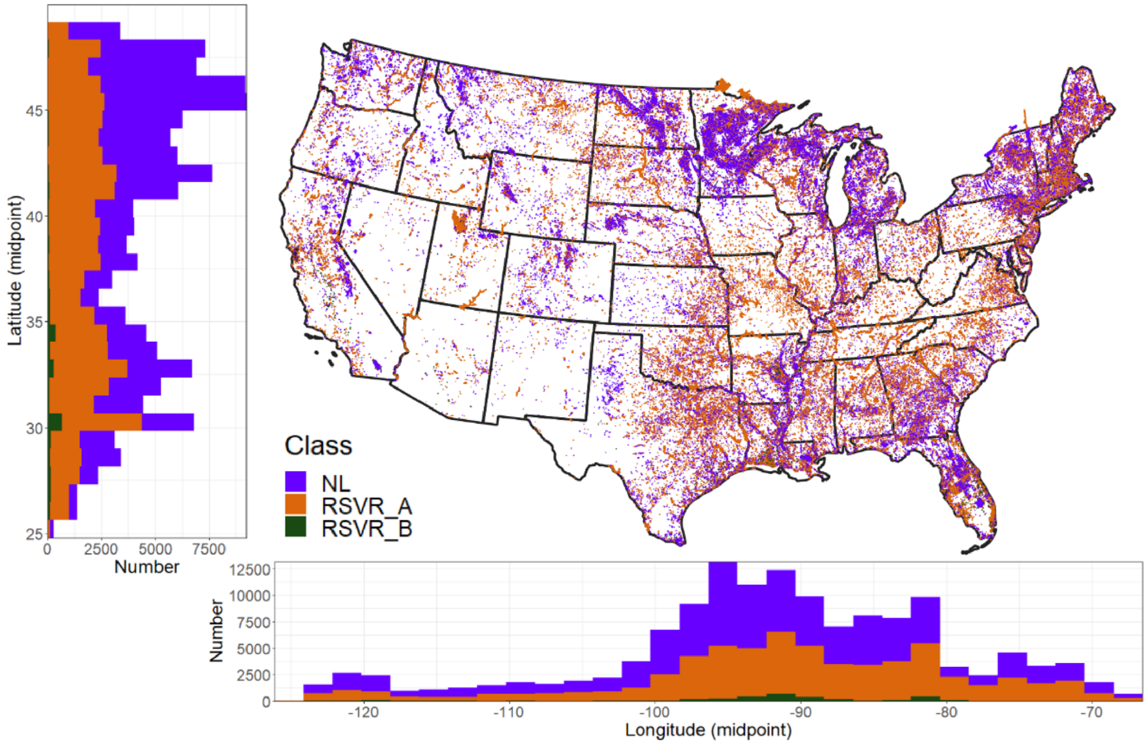


Fig. 1. LAGOS-US RESERVOIR v1 map and histograms depicting locations of 137,465 lakes ≥ 4 ha as NL (purple, n = 73,053), RSVR_A (orange, n = 61,042), or RSVR_B (green, n = 3370) in the conterminous U.S. states (black outlines). RSVR polygons were the first GIS layer to be plotted; therefore, some NL polygons may overlap or hide the true spatial extent of all RSVRs.

dictionary table includes the names, units, and other relevant metadata related to each variable in RESERVOIR.

Data table: Lake reservoir

The RESERVOIR data table (*lake_reservoir*) includes variables related to the lake classification (e.g., method, class, model, probability), lake characteristics, common identifiers linking RESERVOIR to other LAGOS-US data modules and to other key U.S. datasets, locational information of lakes and dams, and flags associated with observations. In addition to identifying each lake as either a NL, RSVR_A, or RSVR_B and providing locational information (Fig. 2), the data table includes whether the class was assigned manually using aerial imagery or was predicted by the machine learning model that used lake outlines as model input, the model it was predicted from (based on location in the United States), and the probability associated with each classification. These probabilities give users information about which predictions may be associated with a higher degree of confidence (i.e., may have been correctly classified).

RESERVOIR includes two additional metrics that help to understand and interpret the lake classification. First, a metric of lake shape (shoreline development factor) is included in the data table because it has long been assumed that reservoirs are differently shaped from NLs. Our data support this assumption to some degree by showing that both manually classified and model-classified RSVRs have a higher shoreline development factor (indicating a larger deviation from a perfect circle) than NLs (Fig. 3). However, there is large overlap in shape between NLs and RSVRS, suggesting that it is not possible to differentiate NLs from RSVRs using this simple shape metric alone.

Second, we include a measure of lake surface water connectivity, which is based on upstream and downstream and lake connections. These connectivity metrics are fully described in LAGOS-US LOCUS (Cheruvelil et al. 2021). There are six classes of connectivity determined from the stream network (made up of both permanent and intermittent/ephemeral stream flow; Fergus et al. 2017) using data from NHDPlus HR Beta 2021 snapshot. There are two classes of lakes that have inflow(s) and outflow(s)—“Drainage”, in which there are no upstream lakes ≥ 10 ha or “DrainageLk”, in which there are one or more upstream lakes ≥ 10 ha. There are two classes of lakes that have only inflow(s)—Terminal, in which there are no upstream lakes ≥ 10 ha or “TerminalLk”, in which there is one or more upstream lakes ≥ 10 ha. The last two classes are less connected to other surface waters, with “Headwater” having only an outflow and “Isolated” having no inflows or

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

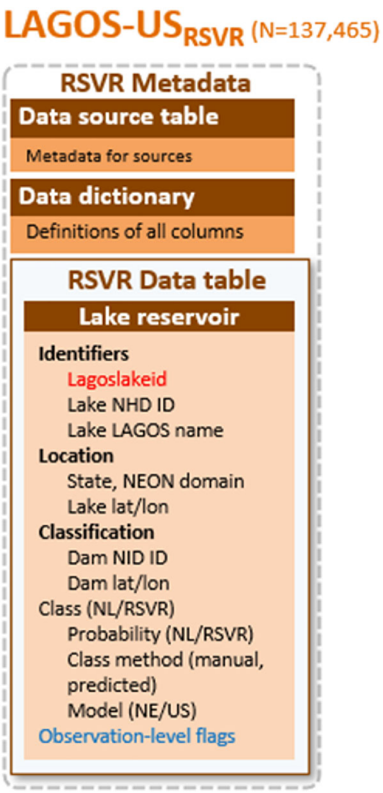


Fig. 2. The RESERVOIR schema. RESERVOIR includes two metadata tables (in the form of a source table and a data dictionary) and a data table (*lake_reservoir*) that includes observation-level flags (blue text). The tables are connected to each other and other LAGOS-US modules via lagoslakeid, depicted with red text. The variables in black text included in the data tables are representative examples, rather than exhaustive. The census population of lakes ≥ 4 ha is $N = 137{,}465$. Not shown: Detailed user guide and polygon layer of natural lakes and reservoirs.

outflows. Because reservoirs are assumed to be created from the damming of rivers and streams, we expected that RSVRs would be connected to streams only, whereas NLs would be well represented in all six of these lake connectivity categories. As expected, the largest number of RSVRs were in the Drainage class. However, beyond this result, our expectations were not met. The next highest number of RSVRs were found in the least connected Isolated class (Fig. 4). Data exploration found that many of these systems are in highly populated areas and appear to be human-made detention and retention ponds, as well as constructed ponds with highly modified stream connections that may be underground or not represented in the NHD stream dataset.

RESERVOIR also includes LAGOS-US lake identifiers, National Inventory of Dams (NID) IDs, NHD IDs, and official lake names, when available. Although RSVRs and NLs are officially named at approximately the same proportions (43% and 44%, respectively; Fig. 5), the most common names given to these two types of water bodies differ (Fig. 6). For NLs, the most common name is “Mud”; whereas, for RSVRs, the most common name is “Long” (Fig. 6). “Horseshoe” was found to be prominent across both classes (Fig. 6). Finally, both NLs and RSVRs share the term “lake” or “pond” as common lake types used during naming. However, RSVRs are also commonly termed “reservoirs” (data not shown).

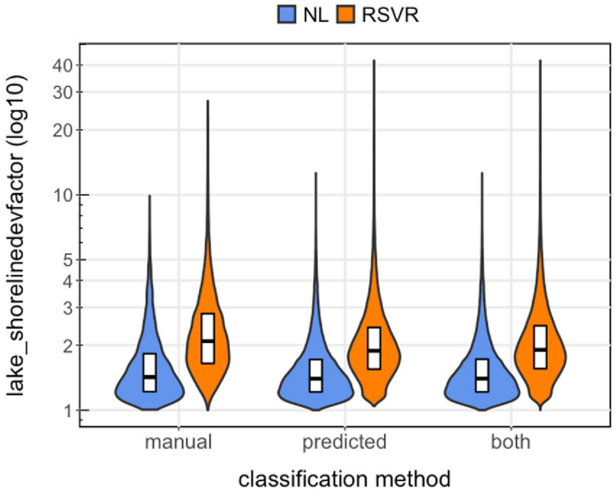


Fig. 3. Violin plots of the shape metric “shoreline development factor” (SDF) plotted following $\log_{10}$ transformation. Violin plots show the kernel density distribution of shoreline development factor for manually classified lakes, model predicted lakes, and all lakes separated by natural lakes (NL, blue) and reservoirs (RSVR, orange). Embedded boxplots show the median value and the interquartile range (25th and 75th percentiles) of the $\log_{10}$ transformed data. 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. For this analysis, we did not differentiate the RSVR classes since the sample size of the B class was extremely small.

Methods

Data sources

The LAGOS-US RESERVOIR data module was created using existing datasets from a variety of sources: the NID (USACE 2015), the National Agriculture Imagery Program (U.S. Department of Agriculture Farm Service Agency Aerial Photography Field Office 2016), Google Earth Imagery, and LAGOS-US LOCUS v1.0 that is based on the high-resolution NHD that was downloaded in January 2017 (LOCUS; Cheruvelil et al. 2021; Smith et al. 2021) with code and images made publicly available (Wang et al. 2021). All 479,950 lake polygons greater than or equal to one hectares within the spatial extent of the conterminous U.S. were obtained from the LAGOS-US LOCUS v1.0 geodatabase (gis_locus.gpkg; Smith et al. 2021). However, the study lake population for LAGOS-US RESERVOIR is a subset of these polygons ($n = 137{,}465$) that are ≥ 4 ha. The minimum lake area of 4 ha was selected because of

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

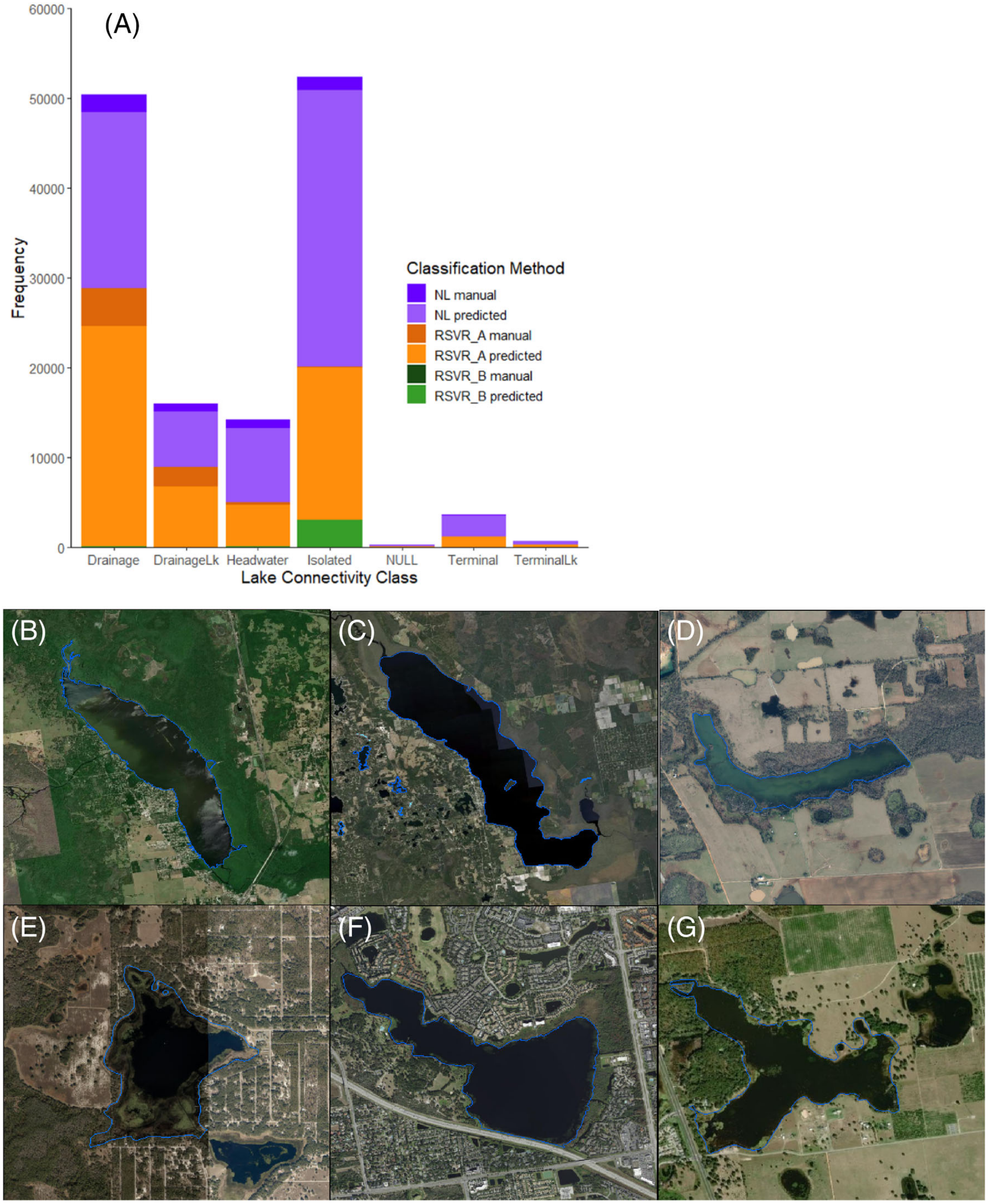


Fig. 4. Stacked bar chart of the frequency of natural lakes (NL) and reservoirs lakes (RSVR_A and RSVR_B) according to lake connectivity class (**A**). Lakes are further differentiated according to whether they were manually classified via aerial imagery (dark purple, dark orange, and dark green) or were predicted with a machine learning model light purple, light orange, and light green). Not included: 292 water bodies classified as "NULL" due to data that prevented quantifying their hydrologic connections. Images demonstrating examples of RSVRs in each of the six lake connectivity classes (**B–G**): **B** = headwater, **C** = drainage, **D** = terminal, **E** = isolated, **F** = DrainageLK, **G** = TerminalLK (service layer credits. Source: Esri, Maxar, GeoEye, earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS user community).

limitations associated with interpreting aerial imagery for smaller water bodies that prevented definitive manual interpretation of the presence of water control structures on these smaller lakes. Details of these sources and how they were used are included in the LAGOS-US RESERVOIR User Guide (Polus et al. 2022).

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

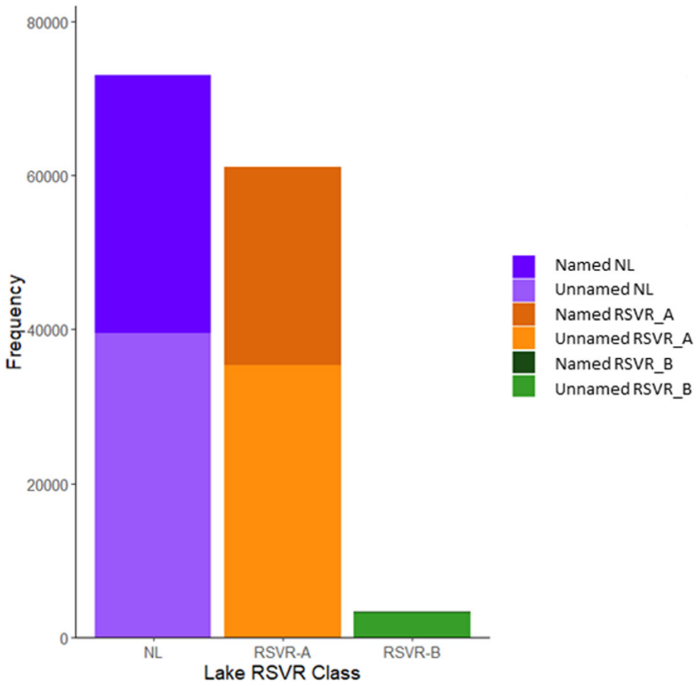


Fig. 5. The frequency of lakes with and without official names, according to whether they are natural lakes (NL) or reservoirs (RSVR_A and RSVR_B).

Classification method overview

There were four main steps to classify all lakes ≥ 4 ha in the conterminous U.S. as either a NL or RSVR (Fig. 7). We describe the details of each step below.

Step 1: Compile and query lake dataset by region

We divided the United States into two model regions (Fig. 8). The first region, the “NE model,” has a high lake density (both NLs and reservoirs) and includes lakes that were predominantly glaciated (Fig. 2). This region includes 17 northeastern and upper midwestern U.S. states obtained from the LAGOS-NE v1 database (Soranno et al. 2015, 2017). The second region, the “U.S. model” includes the remaining 31 states and Washington D.C. This region has lower overall densities of NLs and reservoirs, and includes many states that are dominated by RSVRs (Fig. 8). We selected a subset of lakes within each model region (NE and U.S.) for manual classification. However, the approach was slightly different between the two regions. For the NE model, we selected ∼ 200 NLs and ∼ 200 RSVRs per state for all 17 states; but, for the U.S. model, this approach was prohibitive because of the large geographic area and larger number of states. Therefore, we grouped the remaining states according to the National Ecological Observatory Network domains that have ecological boundaries based on similar climatic characteristics (NEON; Keller et al. 2008). We selected 15 of the 31 states that we thought best represented those NEON domains (Fig. 8). Our goal was to manually classify ∼ 200 NLs and ∼ 200 RSVRs per NEON domain for the U.S. model; however, this threshold was not met in every region due to low numbers of NLs in some NEON domains in the U.S. study region, leading to some domains containing fewer manually classified lakes than others (Fig. 8).

Step 2: Build training dataset

For both regional training datasets, we manually classified lakes using aerial imagery to examine lake outlines and to identify the presence of dams or other water control structures (Polus et al. 2022). We first subset all lakes into two possible categories (potential RSVRs or NLs) by overlaying a GIS point layer of over 90,000 dam locations from the NID (USACE 2015) over our GIS polygon layer of lakes to identify dams that were within 50 m of a lake. We used the 3D distance tool in ArcMap, which considers the elevation of input

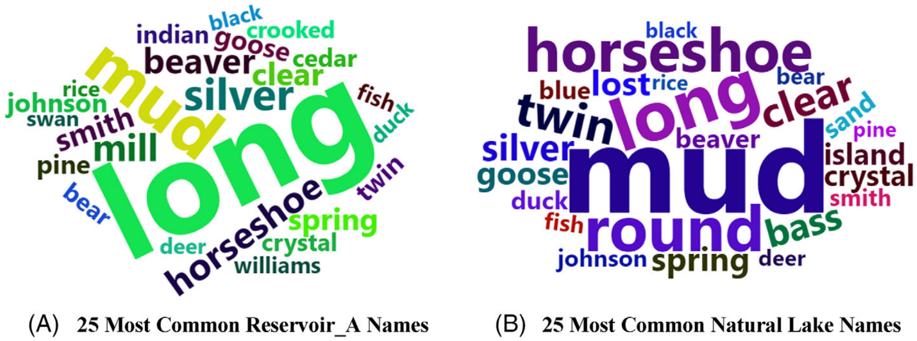
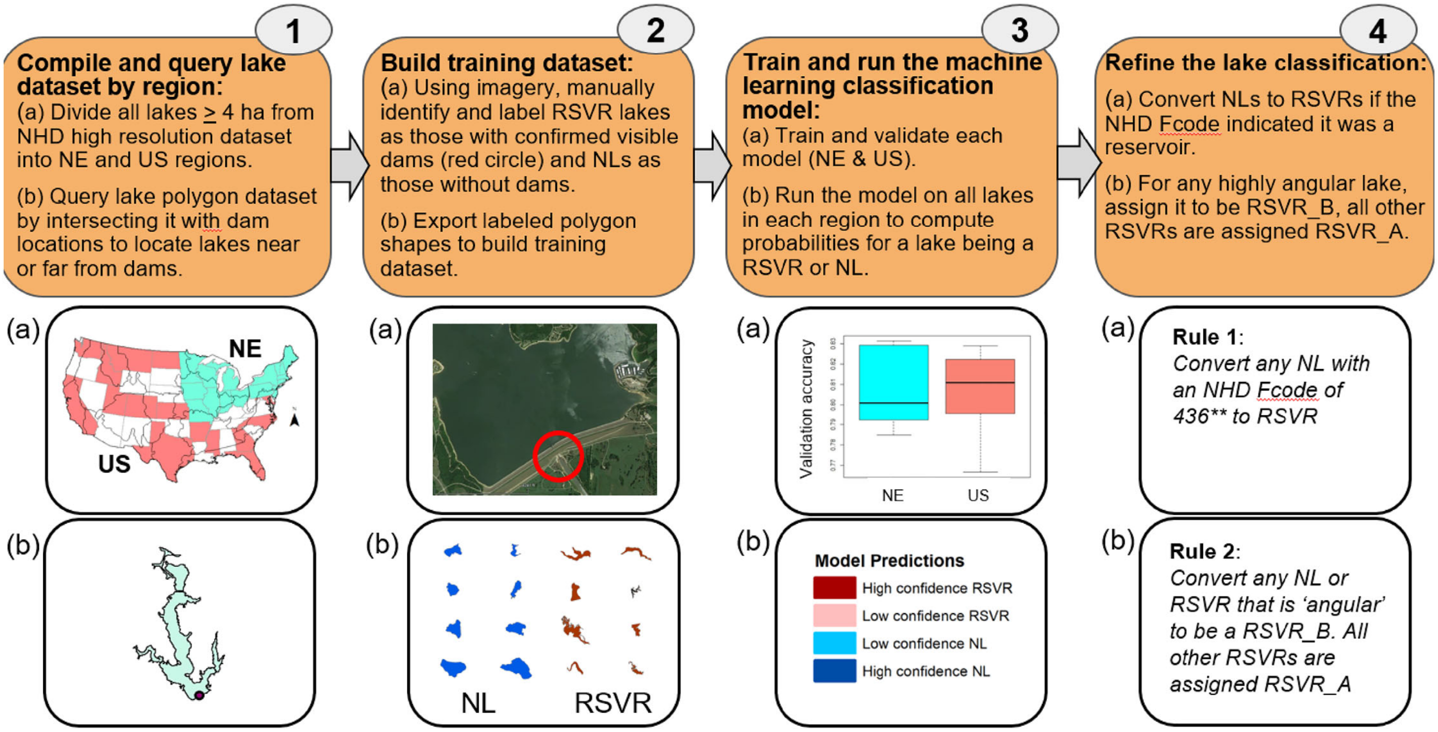


Fig. 6. Word clouds depicting the 25 most common official lake names for (**A**) reservoirs (RSVR_A) and (**B**) natural lakes ≥ 4 ha in the conterminous U.S. for both natural lakes and reservoirs (RSVR_A), names often include three parts—a descriptive adjective, a primary name, and a lake type. For these word clouds, only the primary name is depicted (such as “mud,” “long,” “round,” “twin,” “horseshoe”). The descriptive adjective portion of the names (such as big/little, south/north) and the lake type (such as a “lake,” “pond,” “reservoir,” “tank,” “impoundment”) were not included in this analysis so that primary names could be analyzed and visualized. A word cloud was not visualized for RSVR_B water bodies since just 113 were named and almost all of these were uniquely named.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Compile and query lake dataset by region:
(a) Divide all lakes ≥ 4 ha from NHD high resolution dataset into NE and US regions.
(b) Query lake polygon dataset by intersecting it with dam locations to locate lakes near or far from dams.

Build training dataset:
(a) Using imagery, manually identify and label RSVR lakes as those with confirmed visible dams (red circle) and NLs as those without dams.
(b) Export labeled polygon shapes to build training dataset.

Train and run the machine learning classification model:
(a) Train and validate each model (NE & US).
(b) Run the model on all lakes in each region to compute probabilities for a lake being a RSVR or NL.

Refine the lake classification:
(a) Convert NLs to RSVRs if the NHD Fcode indicated it was a reservoir.
(b) For any highly angular lake, assign it to be RSVR_B, all other RSVRs are assigned RSVR_A.

Rule 1: *Convert any NL with an NHD Fcode of 436** to RSVR*

Rule 2: *Convert any NL or RSVR that is 'angular' to be a RSVR_B. All other RSVRs are assigned RSVR_A*

Fig. 7. Flow diagram depicting the four steps to predict natural lakes (NL) and reservoir lakes (RSVRs) in conterminous U.S. NHD is National Hydrography Dataset.

data (i.e., the dam points) to maintain a realistic view of local topography. We identified candidate RSVRs in each region by selecting water bodies that were within 50 m of a dam and we identified candidate NLs by the absence of a dam within 50 m. We tested other distances to screen the data and found that larger distances included too many NLs that were not reservoirs, and smaller distances excluded too many NLs that were reservoirs. Regardless of the distance chosen, we used this step only to screen the data to find candidate RSVRs and NLs that were then manually identified using imagery.

The manual processing of the above two classes of candidate lakes involved visual interpretation of aerial imagery using Google Earth's historical imagery tool that allowed inspection during all times of the year including leaf-off to allow better identification of dams on lake shorelines. This manual step was important because although most dams within 50 m from a lake from the NID dataset were found to be what appeared to be true dams on lake shorelines, there were some instances where dams were located on inflowing (or outflowing) streams very close to the lake shoreline, were small relative to the size of the lake shoreline, and that simply appeared visually to slow down water flow, rather than result in significant damming of water based on a combination of size relative to the size of the lake, and location of the dam relative to the lake shoreline and incoming streams. Therefore, we classified such lakes as NLs. In fact, although RSVR shapes are thought to be dendritic, with one straight side (where the dam is assumed to be), while NL shapes are thought to be more rounded (Fig. 9), we found this simplification was only sometimes the case (Fig. 10). Lakes form via different natural processes that may influence the presumed "circular" shape of NLs. Our assumption was that these water bodies would be ecologically more similar to NLs than RSVRS because the water control structure was very small relative to the size of the lake, so likely would not greatly alter the lake basin nor affect the natural hydrology and ecology (see definitions in Box 1; Fig. 10). Because the NID dataset did not indicate whether a lake basin was present prior to the construction of the dam, we were not able to use a more quantitative metric for this manual assessment. Furthermore, simply using visual interpretation has the advantage of being a method that can be replicated in other regions of the world that lack detailed data on dams, which is part of the justification for developing a visual-only metric of RSVR classification.

Each of the manually classified water bodies in the two training datasets (NE and U.S.) was exported as an individual polygon file depicting the physical lake outline that included the label of manually classified NL or RSVR (n = 12,162). The sample sizes for the training and test datasets are in Table 1, along with the total number of water bodies classified by each model. Sixteen percent (n = 7127) and 6% (n = 5035) of the polygons were manually classified for the NE and U.S. models,

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

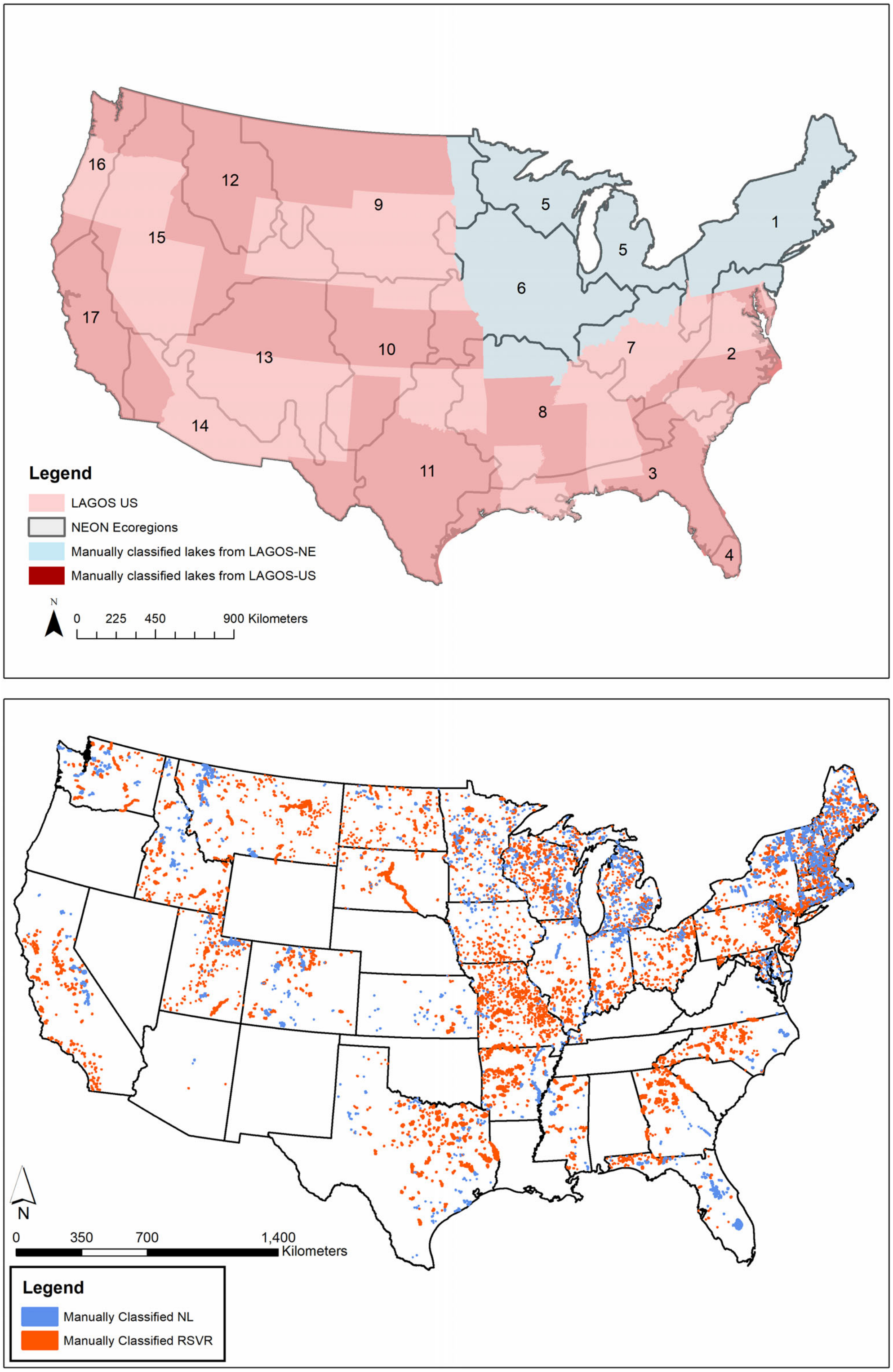


Fig. 8. Map showing the spatial extent of the NE and U.S. models (upper panel) and the lakes that were manually classified as either NL (blue) or RSVR (red) (lower panel). NEON domain boundaries are outlined in black and gray. The blue area is the boundary of the NE model; the red area is the boundary of the U.S. model. Darker red areas are states for which a subset of the lakes was manually classified. For the NE model, all states had ~ 400 lakes manually classified. Light red areas indicate areas for which no manual classification was done.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

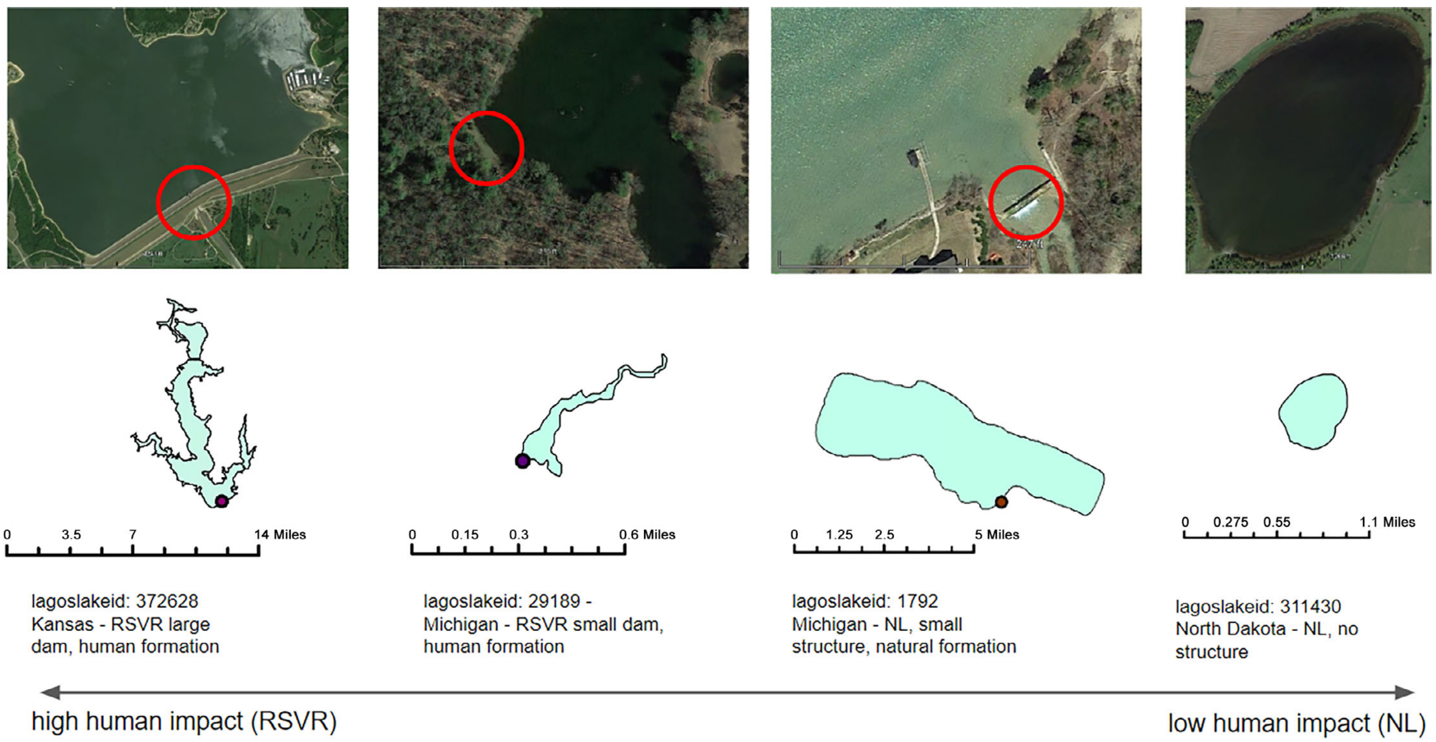


Fig. 9. Images and polygons depicting examples of two reservoirs (left) and two natural lakes (right) that demonstrate a range in the degree of human impacts caused by water control structures and our process for manually (i.e., visually) classifying reservoirs and natural lakes. Left to right: A highly modified lake with a large dam creating a characteristically dendritic reservoir on a mainstem river, a less modified lake that includes a dam on an incoming stream that results in a reservoir, a natural lake that includes a water level control structure at one location, and a natural lake with no structure that is characteristically round.

respectively. Users can be confident in the results of the U.S. model based on only a 5% training dataset because the probabilities associated with correct classification were similarly high between the two models (see *Technical Validation*).

Step 3: Train and run the machine learning classification model

We trained the two machine learning models (NE and U.S.) using the 12,162 manually classified lakes described in *Step 2*. Each model analyzed the shape of the input lake boundaries (via individual polygons; Fig. 10) to quantify the probability that the geometry of an unclassified lake better fit that of a NL or RSVR (lake_prob_nl and lake_prob_rsvr). These two probabilities sum to one (e.g., a lake with lake_prob_nl = 0.25 has a lake_prob_rsvr = 0.75). Model output includes two predictions for every lake. Therefore, a lake with a high probability of being a RSVR subsequently has a lower probability of being a NL (Fig. 11; Polus et al. 2022). Given the complexity of lakes and reservoirs, as well as the scale and scope of this dataset, there may be situations where the probabilities are approximately equal. Therefore, we provide prediction probabilities and overall model accuracy estimates so that users may assess the confidence of the model predictions for each lake and choose their own probability cutoff.

We used the *ResNet18* machine learning model, which is a pretrained deep convolutional neural network model based on image geometry (Krizhevsky et al. 2012; He et al. 2016). We chose this pretrained model because it was trained on a very large image dataset and is quite sensitive to shapes, and we wanted to avoid overfitting the model. All lake polygon PNG file images were rescaled to be 224 × 224 pixels (without changing the aspect ratio), a requirement of the *ResNet18* model, and were exported using the ArcGIS data-driven pages tool. After inputting all classified lake polygon images into the *ResNet18* model, we fine-tuned (trained) the fully-connected layer of the network to fit our manually classified "labeled data." The outputs of the models were the probabilities of lakes to be classified as NLs and RSVRs (Fig. 11).

The models (NE and U.S.) were trained on a compute cluster with a GeForce GTX TITAN X graphics card and CUDA 10.2. The deep learning library we used was PyTorch 1.5.0 (Paszke et al. 2019) coded in Python 3.7.6. We used a standard cross-validation procedure (Stone 1974) to train and evaluate our models. The data were loaded with the DataLoader class provided by PyTorch. We randomly split our labeled data into a training set (90% of data) and a validation set (10% of data) that was used to evaluate the models. Because there were

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

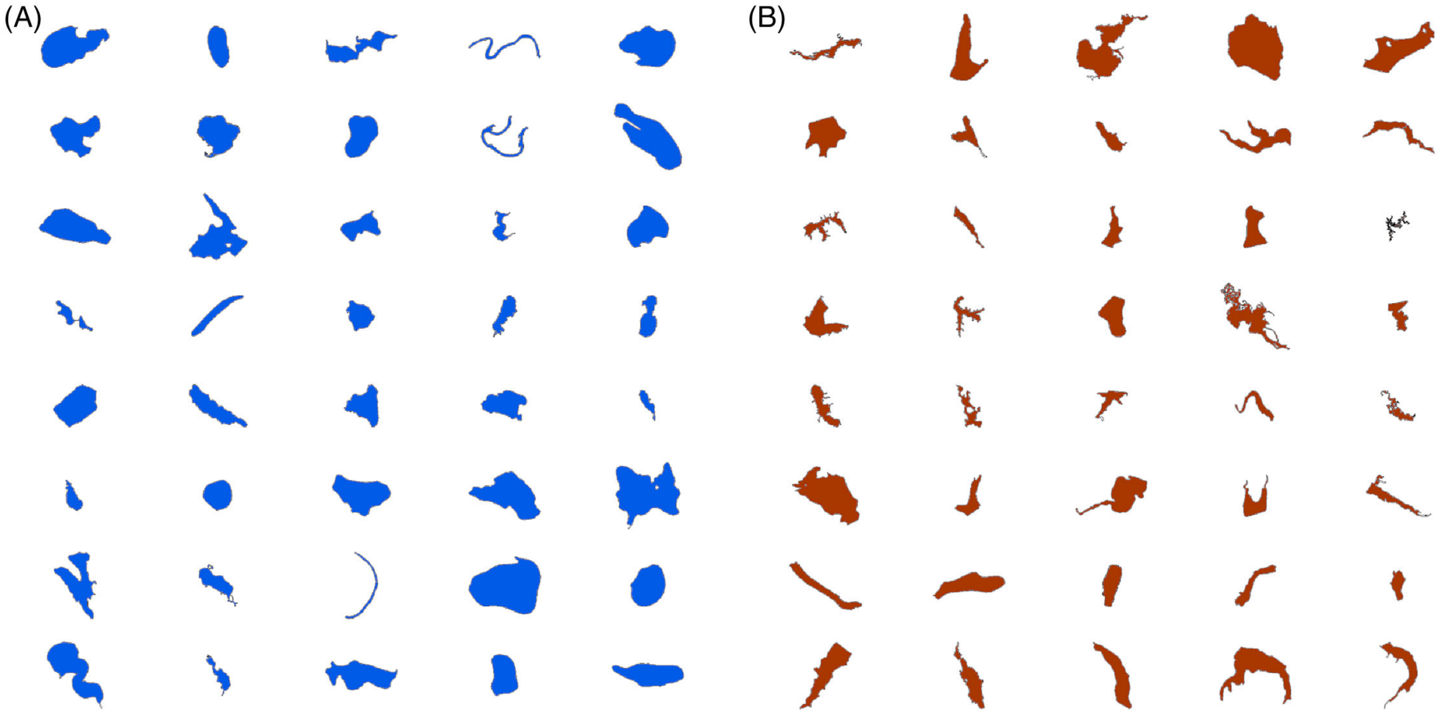


Fig. 10. Randomly selected (**A**) natural lake (blue) and (**B**) reservoir (red) PNG files used in training the classification model and demonstrating the similarity in shapes of NLs and RSVRs. The model used outlines only; filled polygons are displayed for easier visualization.

different numbers of RSVRs and NLs in the training datasets, we downsampled the data to make the two classes balanced. When training the models, the epoch number was set to be 50, and the batch size was 32. In each epoch, the data were randomly shuffled. When finishing the training, we predicted the labels for the validation data.

After completing this model-training process, we compared the predicted NL and RSVR labels against the manually classified labels and computed the model accuracy as the number of correct predictions divided by the total number of predictions. We repeated the cross-validation procedure 10 times to obtain average overall model accuracies. The mean validation accuracy for both models was approximately 0.8, indicating that both models were able to predict a lake belonging to either NL or RSVR 80% of the time (Fig. 12).

Once we had well-performing models, we reran the models using the full dataset (i.e., we did not need the 10% validation dataset) to make predictions for lakes that were not in the training datasets. We calculated a metric of the difference between the probability of the lake being RSVR (lake_prob_rsvr) and the probability of the lake being NL (lake_prob_nl) so that users can assess the likelihood of classification. When the difference is large, the classification has a high probability of being correct (e.g., 0.9 for NL and 0.1 for RSVR would indicate a high likelihood that the lake is a NL). When the probabilities are both around 0.50, with a difference close to 0, then the correct classification is basically a coin-flip.

Step 4: Refine the lake classification

The final step in our classification effort was to use two additional rules to refine the classification. The first rule stated that for any lake that was classified as an NL by our model, but that had an NHD Fcode indicating it is a reservoir (43,600–43,626), we reclassified as a RSVR. There were ~ 4000 out of a total of 77,667 NLs that fell in this category. We inspected many of the named RSVRs in this category and found that they did indeed meet our definition of reservoirs. Note that the vast majority of

Table 1. Summary statistics of northeast (NE) and United States (U.S.) model results.

Model	Number of lake polygons in training dataset	Number of states used in model	Number of lake polygons in test dataset
NE	7118	17	45,443
U.S.	5044	15	79,860

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

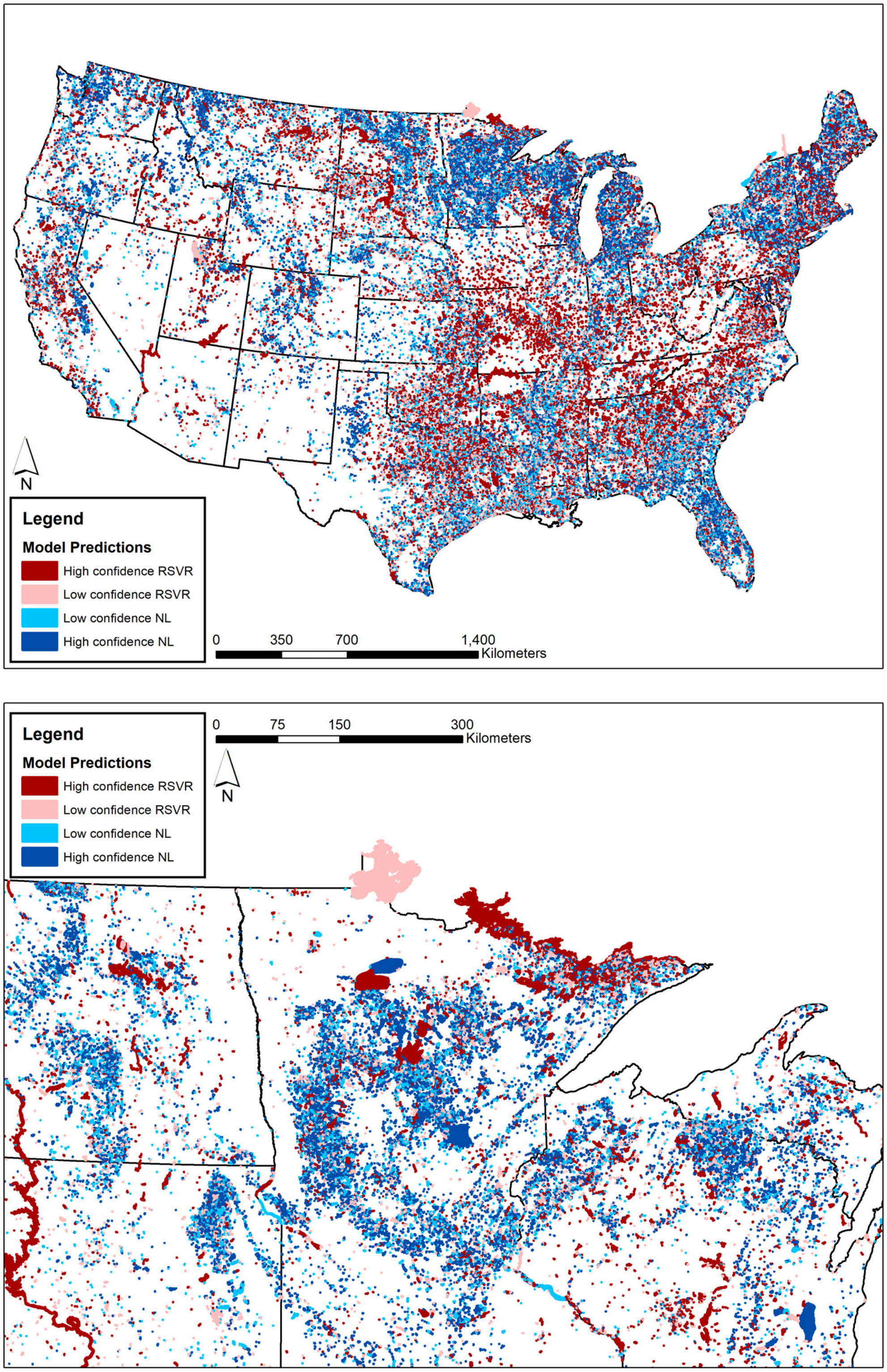


Fig. 11. Map showing probabilities associated with the classification of all lakes ≥ 4 ha as either natural lakes (high confidence NL: lake_prob_NL = 0.75–1.00, dark blue; low confidence NL: lake_prob_NL = 0.50–0.74, light blue) or reservoirs (high confidence RSVR: lake_prob_NL = 0.00–0.25, dark red; low confidence NL: lake_prob_nl = 0.26–0.49, light red) by the machine learning models. Note that the 1 and 0 indicates lakes that were manually classified in step 2 (i.e., visually). Darker shading indicates higher confidence of classification. Below are large-scale maps of the southeastern U.S. (left) and the north-central region of the U.S. (right) showing finer resolution imagery of model classifications.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

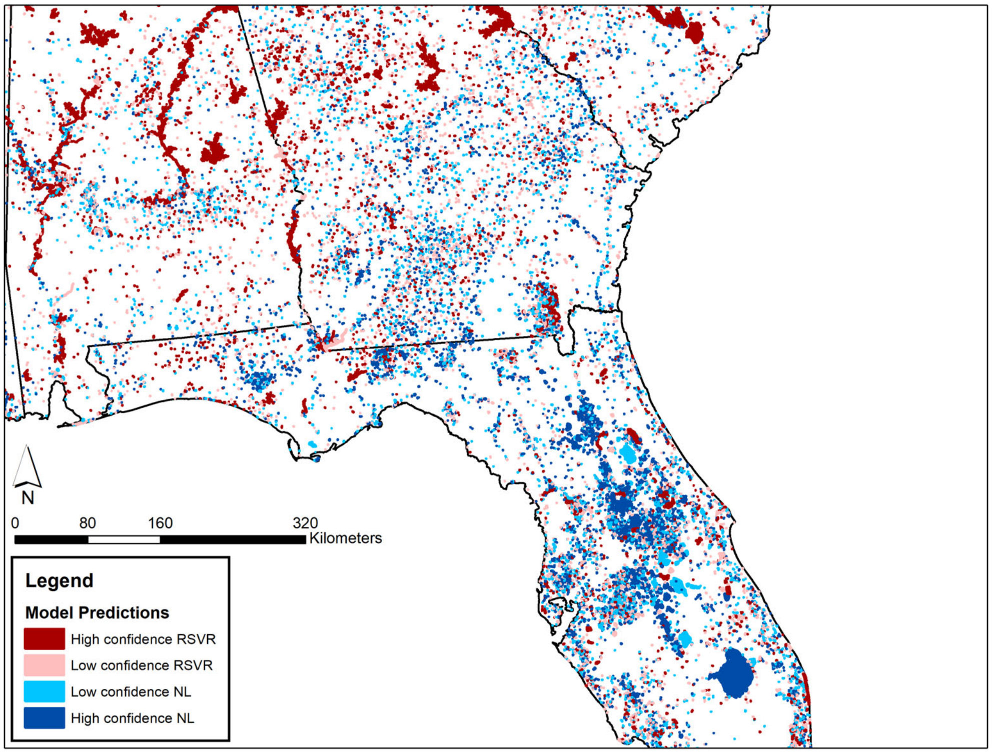


Fig. 11. Continued

lakes in the NHD are classified as "natural" (~ 125,000 out of our total dataset of 137,465). Therefore, NHD Fcodes can be used to identify lakes that were incorrectly predicted to be natural when they are reservoirs, but not to identify NLs that are incorrectly labeled as reservoirs (see also *Technical Validation*). Moreover, since the subset of lakes coded as "natural" by NHD likely includes reservoir, the actual false negative rate for machine classification as a reservoir is unknown.

The second rule related to the angularity of the lake shoreline. Using this rule, we subdivided the RSVR class into 2 subclasses. Subclass RSVR_A includes reservoirs with non-angular shapes whereas subclass RSVR_B from includes reservoirs that are angular and often isolated from rivers. To identify those reservoirs in RSVR_B, we used one of the data flags that was created in the LAGOS-US LOCUS v1 data module that indicates whether the shape of the lake is strongly angular, which is indicative of being artificial (Cheruvelil et al. 2021; Smith et al. 2021). An angular lake is defined as one with a shape that nearly conforms to a rectangle using the ratio between the lake area and the area of the minimum bounding rectangle area that is close to 1 (Smith et al. 2021). Many of the lakes with the highest angular values are also isolated (3042 out of 3370) and so are likely not to fit the traditional definition of reservoirs that assumes they result from a dammed river. We also found 1334 lakes that our model classified as NLs but that were flagged as angular, which we reclassified as RSVR_B. Even though such lakes are in the NHD dataset as "NLs," we did not include such angular lakes in the manually classified datasets because these lakes are not particularly common and do not fit our definitions.

Technical validation

Although it is not possible to conduct technical validation of the source datasets that were used in this study, instead we describe the different steps of our workflow in which we attempt to ensure data quality of the variables that we have created. Furthermore, as for the other LAGOS-US data products that have been created, we do not always definitely differentiate between "good" or "bad" data during technical validation, and instead ensure that our intended processes are doing what we intended through data checks during all input or processing stages, and to identify where caution should be taken by future users by creating data flags for them (see next section). Finally, we validate our dataset against other known data products for large reservoirs.

Data flags

We created two data flags that describe potential data quality issues for future users to consider (Table 2). First, we flagged RSVRs with the connectivity class of "Isolated." It is likely that some of these isolated RSVRs are human-constructed water bodies that do not meet our definition of reservoir (Box 1), but have high levels of human modification

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

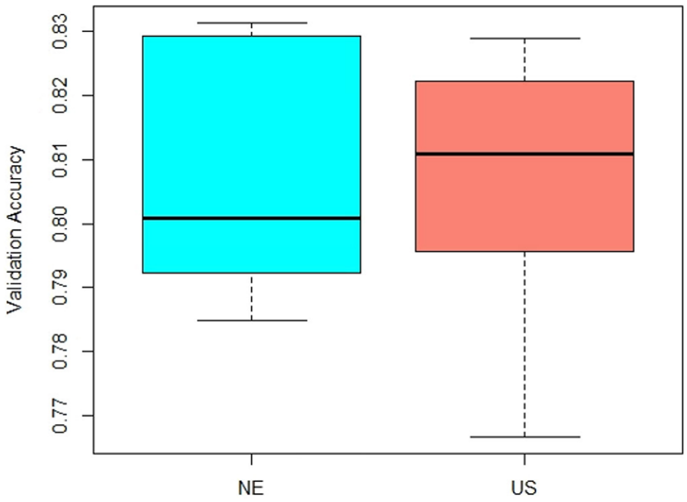


Fig. 12. Model validation accuracy for the 10 model iterations for the two models (NE and U.S.).

Table 2. Summary showing the number and percent of lakes with cautionary data flags. The complete dataset includes 137,465 NLs and RSVRs.

Flag	Number of occurrences	Percent of flags in each class (RSVR or NL)
lake_rsvr_rsvrisolated_flag	15,475	26%
lake_rsvr_nlneardam_flag	5485	7%

(as evidenced by straight lines; Fig. 13). In fact, many of these systems are human-constructed ponds in highly populated areas. Second, we flagged NLs that were less than 50 m from a dam in the NID, which should not be the case but is likely a reflection of the fact that we allowed some lakes with water control structures to be classified as NLs (Fig. 13). Nevertheless, this (and the previous) flag provides cautionary information on potentially misclassified waterbody polygons.

Validation

We validated LAGOS-US RESERVOIR using both the Reservoir Morphology Database (Rodgers 2017) and the Global Reservoir and Dam (GRanD), The Reservoir Morphology Database includes 904 unique LAGOS lakes that our machine learning model classified as a reservoir in each case with a median probability of reservoir classification of 90.5%. The 2019 GRanD v1.3 classifies 2320 LAGOS lakes ≥ 4 ha as reservoirs (Lehner et al. 2011; Messager et al. 2016). Of the lakes common to both databases, 1992 (85.9%) were classified as reservoirs (either RSVR-A or RSVR-B) by LAGOS-US RESERVOIR. Some of the 14.7% of lakes differently classified between GRanD and RESERVOIR are a result of our definition of reservoir, rather than inaccuracy in our model predictions. For example, Lake Cayuga, the longest of the Finger Lakes in New York, has a water regulating lock but was naturally formed. Therefore, although the lake is included in GRanD, we consider it a NL in RESERVOIR. It should be noted that our validation step was conducted for lakes that are present in the GRanD dataset and that are generally larger (for shared lakes: median surface area = 682 ha, 1st quartile = 276 ha, 3rd quartile = 2397 ha), and so future efforts should focus on external validation for smaller lakes.

Data use and recommendations for reuse

Users of RESERVOIR should make note of four important points. First, we strongly encourage the use of the detailed User Guide (Polus et al. 2022) and the data flags as described above to identify lakes that might warrant further inspection. For example, it is likely that the RSVR_B class includes reservoirs with high levels of human modification. In fact, we analyzed the land use within a 500 m buffer surrounding such water bodies and found them to be dominated by non-natural land uses (i.e., agriculture or urban land cover) rather than natural land uses (i.e., forest, grassland, or wetland). The median non-natural cover was

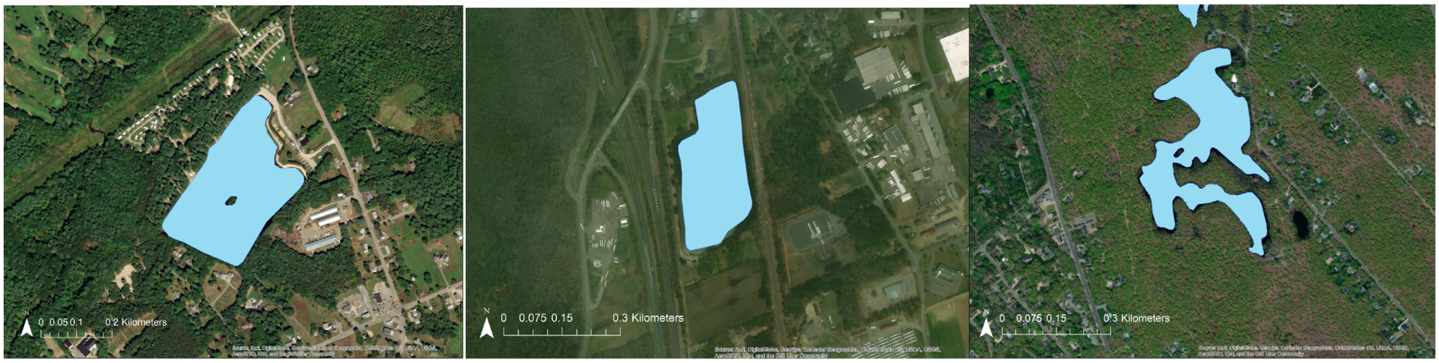

N
0 0.05 0.1 0.2 Kilometers
N
0 0.075 0.15 0.3 Kilometers
N
0 0.075 0.15 0.3 Kilometers

Fig. 13. Examples of RSVRs with "isolated" connectivity class. These lakes are not connected to any other lake, river, or stream. See text for further details.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. A comparison of pre-existing datasets that contain data and classification for lakes and reservoirs. Dataset divided into three categories: LAGOS, conterminous, and global datasets.

SHORT NAME, full name and/or description	Lake type included in dataset	Classification to differentiate waterbody type	Total number of lakes (reservoirs; natural lakes)	Min lake size (area, ha unless otherwise indicated)	Year(s) created	Cite
LAGOS Datasets						
LAGOS-US RESERVOIR, LAGOS-US RESERVOIR v1.0: A database classifying conterminous U.S. lakes 4 ha and larger as natural lakes or reservoirs	Lakes, reservoirs	Yes (RSVR, NL)	137,465 (59,861; 77,604)	4	2019–2020	This study
LAGOS-US LOCUS, LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S.	Lakes, reservoirs	No	479,950 (59,861; 77,604*)	1 (4*)	2017–2020	Cheruvelil et al. (2021)
CONTERMINOUS Datasets						
RMD, reservoir morphology database for reservoirs greater than 101 ha	Reservoirs	No	3828 (3828; 0)	101	2016	Rodgers (2017)
NHD high res, National Hydrography Dataset	Lakes, reservoirs	Partial	> 500,000 (variable)	NA	Continually updated	[1]
NHDPlusV2, geo-spatial hydrologic framework dataset	Lakes, reservoirs	Partial	> 500,000 (variable)	NA	2012	[2]
NHDPlus HR, geo-spatial hydrologic framework dataset	Lakes, reservoirs	Partial	> 500,000 (variable)	NA	2012-present	Horizon systems corporation, upcoming[3]
LakeCat, Lake-catchment dataset-characterizing landscape features for lake basins	Lakes, reservoirs	No	378,000 (NA)	0.1	~ 2017	Hill et al. (2018)
NLA-2007 National Lake Survey of the National Aquatic Resource Surveys	Lakes, reservoirs	Yes	1157 (636; 521)	4	2007	Pollard et al. (2018)
NLA-2012 National Lake Survey of the National Aquatic Resource Surveys	Lakes, reservoirs	Yes	1287 (684;615)	4	2012	Pollard et al. (2018)
The National Eutrophication Survey, NES: Lake characteristics and historical nutrient concentrations	Lakes, reservoirs	No	775 (NA)	6	1972–1975	Stachelek et al. (2018)

(*Continues*)

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. Continued

SHORT NAME, full name and/or description	Lake type included in dataset	Classification to differentiate waterbody type	Total number of lakes (reservoirs; natural lakes)	Min lake size (area, ha unless otherwise indicated)	Year(s) created	Cite
GLOBAL Datasets						
MSSL Global Lakes database, MGLD: Database for identifying “closed” lakes	Lakes, reservoirs	Yes	1409 (226; 857)	10,000	1995	Birkett and Mason (1995)
Dataset of large reservoirs	Reservoirs	No	713 (713; 0)	Storage *volume* of ≥ 0.5 km^3	1997	Vörösmarty et al. (1997)
The world register of dams	Reservoirs	No	25,000 (25,000; 0)	All reservoirs with *dam height* ≥ 15 m	1998	ICOLD (1998)[4]
Survey of the state of World Lakes	Lakes, reservoirs	No	752 (NA)	NA	2002	ILEC (2002)[5]
HydroLAKES database	Lakes, reservoirs	Partial	1,427,688 (NA)	10	2016	Messager et al. (2016)
GRaLT: Global reservoir and Lake training	Lakes, reservoirs	Yes	3800 (1900; 1900)	10	2017	Fang et al. (2019)

*For LAGOS-US data modules, the number of natural lakes and reservoirs that are between 1 and 4 ha in surface area is unknown. Numbers listed are for waterbodies > 4 ha.

[1] https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution.

[2] https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution.

[3] http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php.

[4] https://www.icold-cigb.org/GB/world_register/world_register_of_dams.asp.

[5] http://www.ilec.or.jp/database/database.html.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

~ 90% for RSVR_B lakes vs. less than 30% for both NL and RSVR_A lakes (Polus et al. 2022). Second, users may want to use the prediction probabilities to select the RSVRs that are very likely to be correctly classified as such (e.g., lake_rsvr_probrsvr > 0.85) rather than the 0.50 cutoff that is included in this module. Third, this module does not include reservoirs smaller than 4 ha. Fourth, two of the data sources used to create this module are dynamic. The dam data in NID changes through time with the building and removal of dams, and there are continuous updates to the NHD. Therefore, the snapshot used for the dam locations and the lake polygons in RESERVOIR may not exactly match future iterations of the NHD and NID.

LAGOS-US RESERVOIR v1 will be the first research-ready dataset of its scope and scale. However, those wanting to use the data for local or single-state purposes will want to complete additional manual checking of the data prior to use. RESERVOIR can be linked with other LAGOS-US modules as well as additional national-scale datasets via common identifiers to enable scientists to conduct a wide range of reservoir and NL studies. This module will facilitate the study of both NLs and reservoirs at regional- to conterminous-scale in the U.S. for lakes as small as 4 ha, allowing scientists and environmental managers to better understand the similarities and differences between NLs and reservoirs, estimate the role of both in global cycles, and predict lake responses to global changes.

Comparison with existing datasets

RESERVOIR is the first to classify lakes as either NL or RSVR at a very broad spatial extent and for all lakes above a relatively small area threshold (i.e., ≥ 4 ha). Several datasets exist that include both NLs and reservoirs (Table 3). However, the majority of national or global lake datasets do not differentiate between these two types of lakes. Additionally, they do not define their classification criteria of reservoirs, or use "lake" and "reservoir" interchangeably without defining what constitutes each (e.g., ICOLD 1998, ILEC 2002; Table 3). For example, although the MSSL Global Lakes Database classifies them, it does not classify NLs explicitly (Birkett and Mason 1995). The HydroLAKES Database (Messager et al. 2016) does not independently differentiate between NLs and reservoirs; however, it is co-registered with the Global Reservoir and Dam (GRanD) database (Lehner et al. 2011; Messager et al. 2016). Finally, the Survey of the State of World Lakes (ILEC 2002) is unique in its global scale but does not differentiate between NLs and reservoirs.

The NHD includes Fcodes that indicate lake types (e.g., lake/pond or reservoir). In constructing the LAGOS-US database some subcategories of NHD reservoirs that are typically outside the purview of limnological research were prefiltered out (e.g., Fcode 43601: aquaculture, Fcode 43608: swimming pool) whereas others were not (e.g., Fcode 43615: water storage reservoirs; Cheruvelil et al. 2021). In addition, the lack of a reservoir NHD Fcode does not preclude the possibility that a lake is a reservoir. Only 8622 lakes in RESERVOIR are coded as reservoirs by the NHD and 4043 of those are classified as NLs. Inspection of these lakes with conflicting classifications found that they are generally small lakes (median size = 7 ha). Generally, our machine learning classification appears to perform best on visually classifying large reservoirs and the large dams associated with their creation, which aligns with a lack of false negatives when compared to the > 101 ha surface area Reservoir Morphology Database (Rodgers 2017). We conducted a manual evaluation of the 100 largest named lakes with conflicting RESERVOIR and NHD classifications. While some lakes lacked sufficient information to determine the correct class, the vast majority were humanmade reservoirs rather than NLs. In fact, just one lake was confirmed to be a NL, which appeared to have been enlarged through damming. Therefore, we gave preference to the NHD classification, when available.

Beyond the NHD, only three existing datasets include a classification that identifies all lakes as either a NL or reservoir (Birkett and Mason 1995; Pollard et al. 2018; Fang et al. 2019). However, these three studies include a relatively small number of lakes or include only very large lakes (Table 3). There are also some reservoir-only datasets; however, those often focus on large reservoirs with sizable impoundments or dams. For example, the World Register of Dams, a global database containing 25,000 reservoirs, limits the reservoirs in its database by only containing those with a dam height ≥ 15 m (ICOLD 1998). The Reservoir Morphology Database of the United States includes detailed information about reservoirs such as dam height, dates for construction of dams, as well as the storage, discharge and volume, but is only for reservoirs ≥ 101 ha (Rodgers 2017; Table 3). Therefore, RESERVOIR will facilitate studying a wider range of reservoirs based on waterbody and dam size, as well as the regional-conterminous U.S. study of reservoirs and NLs.

References

Birkett, C. M., and I. M. Mason. 1995. A new global lakes database for a remote sensing program studying climatically sensitive large lakes. J. Great Lakes Res. **21**: 307–318. doi:10.1016/S0380-1330(95)71041-3

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

Deemer, B. R., and others. 2016. Greenhouse gas emissions from reservoir water surfaces: A new global synthesis. BioScience **66**: 949–964. doi:10.1093/biosci/biw117

Doubek, J. P., and C. C. Carey. 2017. Catchment, morphometric, and water quality characteristics differ between reservoirs and naturally formed lakes on a latitudinal gradient in the conterminous United States. Inland Waters **7**: 171–180. doi:10.1080/20442041.2017.1293317

Fang, W., and others. 2019. Recognizing global reservoirs from Landsat 8 images: A deep learning approach. IEEE

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

J. Sel. Top. Appl. Earth Obs. Remote Sens. **12**: 3168–3177. doi:10.1109/JSTARS.2019.2929601

Fergus, C. E., J.-F. Lapierre, S. K. Oliver, N. K. Skaff, K. S. Cheruvelil, K. Webster, C. Scott, and P. Soranno. 2017. The freshwater landscape: Lake, wetland, and stream abundance and connectivity at macroscales. Ecosphere **8**: e01911. doi:10.1002/ecs2.1911

Habel, M., and others. 2020. Dam and reservoir removal projects: A mix of social-ecological trends and cost-cutting attitudes. Sci. Rep. **10**: 19210. doi:10.1038/s41598-020-76158-3

Harrison, J. A., and others. 2009. The regional and global significance of nitrogen removal in lakes and reservoirs. Biogeochemistry **93**: 143–157. doi:10.1007/s10533-008-9272-x

Hayes, N. M., B. R. Deemer, J. R. Corman, N. R. Razavi, and K. E. Strock. 2017. Key differences between lakes and reservoirs modify climate signals: A case for a new conceptual model. Limnol. Oceanogr.: Lett. **2**: 47–62. doi:10.1002/lol2.10036

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residual networks, p. 630–645. *In* B. Leibe, J. Matas, N. Sebe, and M. Welling [eds.], *Computer vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, v. **9908**. Springer. doi:10.1007/978-3-319-46493-0_38

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

ICOLD. 1998. World register of dams.

ILEC. 2002. International Lake Environment Committee and United Nations Environment Programme. Data Book of World Lake Environments: A Survey of the State of World Lakes. ILEC: Kusatsu, Japan. Available from http://www.ilec.or.jp/database.html

Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman. 2008. A continental strategy for the national ecological observatory network. Front. Ecol. Environ. **6**: 282–284. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

Knoll, L. B., M. J. Vanni, and W. H. Renwick. 2003. Phytoplankton primary production and photosynthetic parameters in reservoirs along a gradient of watershed land use. Limnol. Oceanogr. **48**: 608–617. doi:10.4319/lo.2003.48.2.0608

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks, p. 1097–1105. *In*, *Advances in neural information processing systems*. NIPS.

Lehner, B., and P. Döll. 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. **296**: 1–22. doi:10.1016/j.jhydrol.2004.03.028

Lehner, B., and others. 2011. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. Front. Ecol. Environ. **9**: 494–502. doi:10.1890/100125

Mamun, M., S. Kwon, J. E. Kim, and K. G. An. 2020. Evaluation of algal chlorophyll and nutrient relations and the N:P ratios along with trophic status and light regime in 60 Korea reservoirs. Sci. Total Environ. **741**: 140451. doi:10.1016/j.scitotenv.2020.140451

Messager, M. L., B. Lehner, G. Grill, I. Nedeva, and O. Schmitt. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. **7**: 13603. doi:10.1038/ncomms13603

Paszke, A., and others. 2019. Pytorch: An imperative style, high-performance deep learning library, p. 8026–8037. *In*, *Advances in neural information processing systems*. NIPS.

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

Polus, S. M., and others. 2022. LAGOS-US RESERVOIR: Data module classifying conterminous U.S. lakes 4 hectares and larger as natural lakes or reservoirs. Environmental Data Initiative. https://doi.org/10.6073/pasta/f9aa935329a95dfd69bf895015bc5161

Rodgers, K. D. 2017. A reservoir morphology database for the conterminous United States: U.S. Geological Survey Data Series 1062. https://doi.org/10.3133/ds1062

Smith, N. J., K. E. Webster, L. 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. Environmental Data Initiative. https://doi.org/10.6073/pasta/e5c2fb8d77467d3f03de4667ac2173ca

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

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

Stachelek, J., C. Ford, D. Kincaid, 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

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Series B Stat. Methodol. **36**: 111–133. doi:10.1111/j.2517-6161.1974.tb00994.x

Thornton, K. W., B. L. Kimmel, and F. E. Payne. 1990, *Reservoir limnology: Ecological perspectives*. John Wiley & Sons.

Tranvik, L. J., and others. 2009. Lakes and reservoirs as regulators of carbon cycling and climate. Limnol. Oceanogr. **54**: 2298–2314. doi:10.4319/lo.2009.54.6_part_2.2298

U.S. Army Corps of Engineers. 2015. Federal Emergency Management Agency: National Inventory of Dams (NID). Available from: https://nid.sec.usace.army.mil/ords/f?p=105:1

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

U.S. Department of Agriculture Farm Service Agency Aerial Photography Field Office. 2016. National Geospatial Data Asset (NGDA) National Agriculture Imagery Program (NAIP). Available from: https://gis.apfo.usda.gov/arcgis/rest/services

U.S. Geological Survey. 2017. National Hydrography Dataset (ver. USGS NHD) Plus High Resolution (HR) Beta for Hydrologic Unit (HU) 4–2001. Available from: https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products

Wang, Q., Polus, S., and P. J. Hanly. 2021. LAGOS-US RESERVOIR Data Module Code. Zenodo. https://doi.org/10.5281/zenodo.5584528

Acknowledgments

We thank Dana Infante, Arthur Cooper, and Jonathan Doubek for their ideas at the inception of this project and their dam expertise; Claire Boudreau who was an REU student during summer 2018 at the inception of this project; Laura Danila who was an REU student during summer 2019 and worked through early algorithms; Pang-Ning Tan and Jiayu Zhou for advising the team on algorithms and working closely with Qi Wang; Allie Shoffner for her editorial help; Jake Namovich and Arika Hawkins for their hourly help with manual lake classification; and Katherine Webster for her help with documentation and QAQC and contributing to lake name code. We also thank two anonymous reviewers and the editors for helpful suggestions on earlier drafts that greatly improved the manuscript.

Submitted 04 November 2021
Revised 22 November 2022
Accepted 23 November 2022

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10299 by Nanjing Institution Of Geography And Limnology, Wiley Online Library on [16/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License