HydroLAKES

Technical Documentation Version 1.0
December 2016

Prepared by Bernhard Lehner and Mathis Messager

(bernhard.lehner@mcgill.ca)

Department of Geography, McGill University

Montreal, Quebec, Canada

Table of Content

Acknowledgements

1. Overview and background

Lakes are key components of hydrological, biogeochemical, and ecological processes, thus knowledge about their distribution, geometric characteristics, and residence time is crucial in understanding their properties and interactions within the Earth system. However, global information on lakes is scarce and inconsistent across spatial scales and regions. The goal of the HydroLAKES database is to provide a seamless high-resolution map of the world's lakes and their main characteristics in support of a broad range of global-scale assessments and analyses, with a focus on improving our ability to model the important roles that lakes play in the global environment.

The HydroLAKES database was designed as a digital map repository to include all lakes with a surface area of at least 10 ha (Fig. 1). The current version comprises the shoreline polygons of 1,427,688 individual lakes. HydroLAKES aims to be as comprehensive and consistent as possible at a global scale and contains both freshwater and saline lakes, including the Caspian Sea, as well as human-made reservoirs and regulated lakes. The HydroLAKES database was created by compiling, correcting, and unifying several near-global and regional datasets (see Methods), foremost the SRTM Water Body Data (SWBD; Slater et al., 2006) for regions from 56˚S to 60˚N, and CanVec (Natural Resources Canada, 2013) for most North American lakes. Map generalization methods were applied and some polygon outlines were smoothed during the mapping process to ensure spatial consistency of the data. The resulting map scale is estimated to be between 1:100,000 and 1:250,000 for most lakes globally, with some coarser ones at 1:1 million.

To enhance the attribute information provided by HydroLAKES, a spatial co-registration to the global river network database HydroSHEDS (Lehner et al., 2008) was established. Literature estimates of lake depths and/or volumes were compiled for all lakes ≥ 500 km^2. To predict average lake depths and volumes for smaller lakes, a geostatistical model based on surrounding land surface topography was developed (Messager et al. 2016). The addition of high-resolution discharge data allowed the estimation of hydraulic residence times for each lake.

All natural lakes of the HydroLAKES database show a combined surface area of 2.67 x 10^6 km^2 (1.8% of global land area), a total shoreline length of 7.2 x 10^6 km (about four times longer than the world's ocean coastline), and a total volume of 181.9 x 10^3 km^3 (0.8% of total global non-frozen terrestrial water stocks). Mean and median hydraulic residence times for all lakes were computed to be 1834 days and 451 days, respectively.

HydroLAKES is publicly available for download at http://www.hydrosheds.org and is free for scientific, educational, and other uses. The data is licensed under a Creative Commons Attribution 4.0 International License (see section 4). By downloading and using the data the user agrees to the terms and conditions of this license. The copyright © of HydroLAKES is held by the authors, 2016, all rights reserved.

Citations and acknowledgements of the HydroLAKES database should be made as follows:

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O. (2016): Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature Communications: 13603. doi: 10.1038/ncomms13603. Data is available at www.hydrosheds.org.
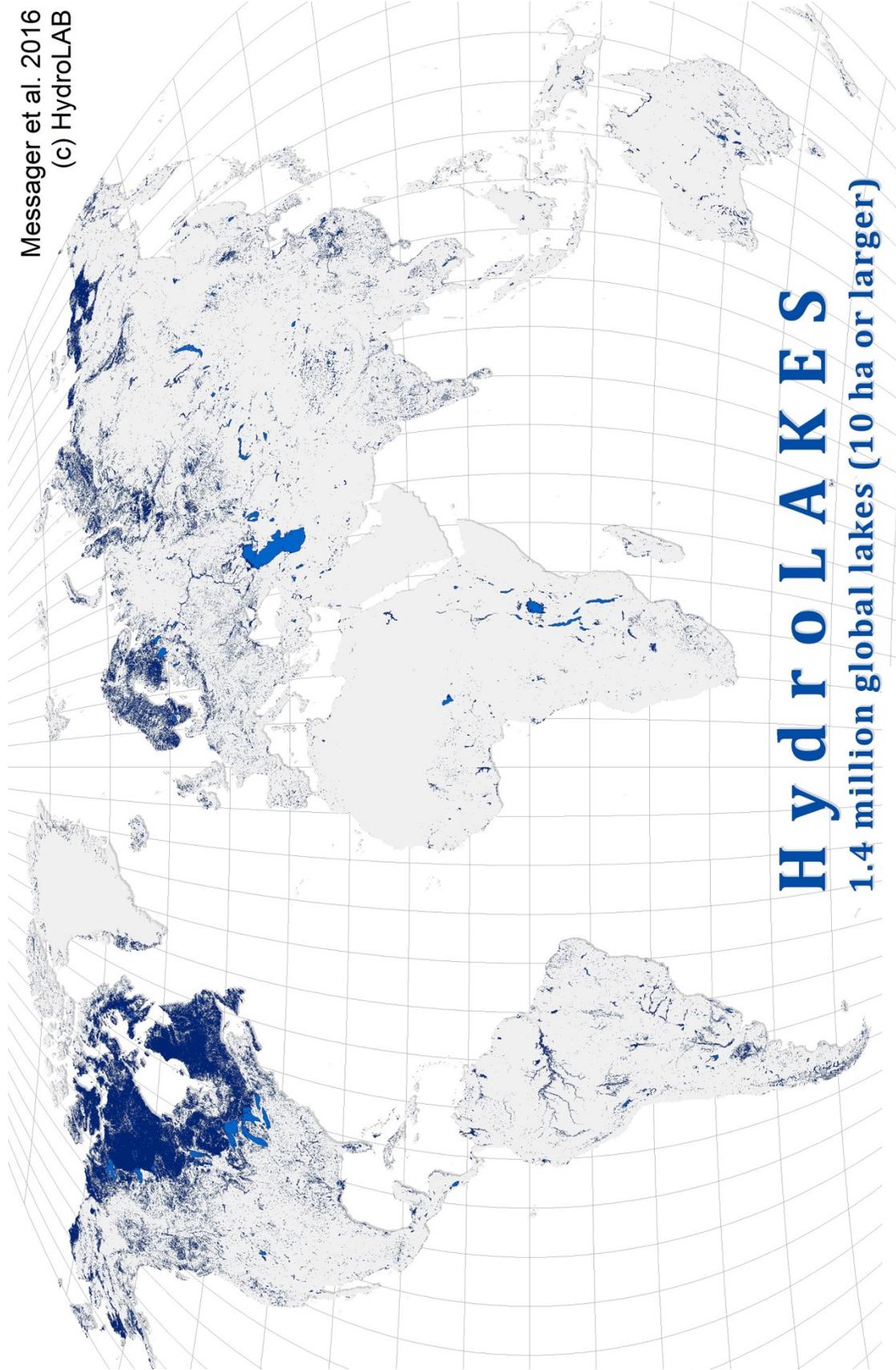
Messager et al. 2016
(c) HydroLAB

H y d r o L A K E S
1.4 million global lakes (10 ha or larger)

Figure 1: Global lake distribution of HydroLAKES.

2. Methods and data characteristics

2.1 Data sources

Table 1 provides an overview of all lake and reservoir datasets used in the development of HydroLAKES. The number of lakes refers to the polygons extracted from each source dataset. More details regarding the data sources and subsequent modifications of the polygon geometry are provided below.

Table 1: Datasets used in the creation of HydroLAKES.

Original dataset	Region	Original format and resolution	Reference	Number of lakes
Canadian hydrographic dataset (CanVec)	Canada (entire country)	Vector; 1:50,000	Natural Resources Canada (2013)	863,550
Shuttle Radar Topographic Mission (SRTM) Water Body Data (SWBD)	56° South to 60° North	Raster; 1 arc-second (~30 m at the equator); vectorized and smoothed	Slater et al. (2006)	282,571
MODerate resolution Imaging Spectro-radiometer (MODIS) MOD44W water mask	Russia above 60° North	Raster; 250 m; vectorized and smoothed	Carroll et al. (2009)	167,435
US National Hydrography Dataset (NHD)	Alaska (entire state)	Vector; 1:24:000	U.S. Geological Survey (2013)	58,496
European Catchments and Rivers Network System (ECRINS)	Europe above 60° North and entire Norway	Vector; varying resolutions (~1:250,000)	European Environment Agency (2012)	50,699
Global Lakes and Wetlands Database (GLWD)	World	Vector; 1:1 million	Lehner and Döll (2004)	3,023
Global Reservoir and Dam database (GRanD)	World	Vector; varying resolutions (1:1 million or better)	Lehner et al. (2011)	1,133
Other (own mapping)	World	Vector; varying resolutions (1:1 million or better)	n/a	781
Total				**1,427,688**

Lake polygons were compiled primarily from SWBD for all areas below 60°N, supplemented by CanVec, MODIS, NHD, and ECRINS datasets for different regions of the world. Data from GLWD, GRanD, and own mapping was added to supplement the global coverage and replace erroneous or inaccurate polygons.

SWBD (Global; 56º South to 60º North)

The SRTM Water Body Data (SWBD) is fully described in Slater et al. (2006). SWBD was generated as a by-product of the Shuttle Radar Topography Mission to provide a uniform and consistent water mask for correcting the SRTM digital elevation model. The primary data source for water identification in SWBD was the orthorectified imagery of radar intensities collected at

1 arc-second resolution (approximately 30 meters at the equator) during the SRTM mission. All lake shorelines were delineated as they appeared at the time of the data collection in February 2000. The creation of SWBD was performed by contractors who used semi-automated extraction protocols in combination with manual supervision and rule-based editing. The contractors used ancillary data sources as guidance and confirmation of the presence or absence of water. These ancillary data were provided by landcover water masks as well as existing maps and charts from 1:50,000 to 1:1 million scales. The landcover water mask was derived from orthorectified, composite Landsat Thematic Mapper (TM) data at a resolution of 28 m, dated from the late 1980s to 1994, i.e. significantly older than the SRTM data. Only in cases where the land/water interface was indiscernible in the SRTM data, or when SRTM data were missing, the landcover may have been used for water boundary delineation.

Particular difficulties in the water body delineation existed for areas of ice and snow due to the associated lack of contrast in the SRTM data. In agricultural areas that appeared to be covered with water, lakes were only depicted if this was supported by the ancillary data. To avoid confusion, the ancillary data were used to intentionally exclude ice, snow, wetlands, agricultural and rice fields, and mangrove swamps. Furthermore, some water bodies were identified from voids in the SRTM radar data due to the specific reflectance property of water surfaces that can lead to weak backscatter signals; yet these voids were difficult to discern from voids due to radar shadow in steep terrain. Besides using ancillary data, editors thus tried to avoid illogical features such as lakes sitting on steep slopes. SWBD includes water bodies of various types, but contractors provided a distinction between lakes (including reservoirs and lagoons) and rivers as part of their editing. Finally, after production and review, the contractors vectorized the final water mask using skeletonization and other boundary extraction tools to convert the raster data into ESRI® Shapefile format. The original SWBD data was provided globally in about 14,000 individual 1x1 degree tiles.

For the creation of HydroLAKES, all lake polygons of the individual tiles were extracted and mosaicked. Lakes that were split into multiple polygons at the edges of tiles were dissolved to form a single polygon per lake. To avoid problems in this dissolving process, caused by small misalignments at tile boundaries that resulted in narrow slivers and offsets, the original polygons were first rasterized at 30 m pixel resolution, then the tiles were mosaicked into seamless grids at a regional scale, and finally the small gaps were removed from the grids using raster-based boundary cleaning processes (expand-and-shrink techniques). The results were then re-vectorized using a customized simplification step to slightly smooth the lake outlines and avoid pixel-shaped polygon boundaries.

CanVec (Canada)

The CanVec database (Natural Resources Canada, 2013) has been produced from various sources, the main one being the National Topographic Data Base (NTDB) of Canada. CanVec contains a variety of waterbody types at a spatial scale of 1:50,000 and is provided as a set of polygon layers in ESRI® geodatabase format. Tests against remote sensing imagery confirmed a very high spatial accuracy of the water body outlines. CanVec provides different water features, including lakes and rivers. It should be noted, however, that the CanVec data is not fully consistent across all provinces; for example, a distinction between permanent and intermittent lakes is available for some provinces but not all.

For the creation of HydroLAKES, a direct use of CanVec as the source of lake polygons was prevented by the fact that the available version 12.0 of CanVec does not consistently distinguish lakes from rivers in all Canadian provinces but rather provides many extensive polygons that represent a conglomeration of multiple lakes and connecting rivers. At the coast, the same issue can involve parts of the ocean surface that is merged with rivers, estuaries or lagoons into single polygons. This problem necessitated an elaborate manual cleaning process: all polygons were visually inspected; those that clearly included river or ocean parts were split into multiple parts; and only those parts representing individual lakes were retained in the HydroLAKES data. A variety of atlases and topographic maps were used to guide this manual cleaning process.

Other data preparation steps included the omission of objects attributed as intermittent lakes in CanVec (as tests showed that most of them were classified as wetlands on other topographic maps). Some corrections of corrupt polygon geometries were conducted, and several cross-border lakes with the US were replaced with alternative (complete) polygons or completed through digitizing of remote sensing imagery.

MODIS (Northern Russia)

The MODerate resolution Imaging Spectro-radiometer (MODIS) MOD44W water mask was created using the SWBD data in conjunction with MODIS 250 m data to provide a global coverage of waterbodies at a 250 m resolution including areas between 60˚N and 90˚N where SWDB is not available (Carroll et al., 2009). Classification of waterbodies was performed by applying regional decision trees for each continent up to 80˚N. A separate classification accounting for ice cover was applied for the region from 80˚N to 90˚N using remote sensing imagery for the months of July and August to represent times of minimal ice cover.

For the inclusion in HydroLAKES, the MODIS water mask was vectorized using customized boundary smoothing procedures. As MODIS data provides only a binary open water surface without further classification into water types such as lakes or rivers, an additional manual post-processing step was performed: all elongated and linear polygon features were visually inspected in order to identify and remove rivers, or to disconnect and remove parts of rivers that were merged with lakes. This step was guided by comparisons with a variety of ancillary information, including atlases, topographic maps, and remote sensing imagery (including Google Earth).

US NHD (Alaska)

The US National Hydrologic Database (NHD) provides water body coverage for the United States based on maps at a spatial scale of 1:24,000 (U.S. Geological Survey, 2013). This database was compiled from USGS hydrologic digital line graph files and EPA reach files, as well as individual state contributions. Visual comparisons with CanVec data indicate that the process of digitizing the line graph files may have introduced some smoothing of the lake outlines.

NHD data was only used in the creation of HydroLAKES for Alaska to maximize the global consistency of SWBD data up to 60°N. Aside from selecting only polygons equal or larger than 10 ha, no further geometric modifications were applied.

ECRINS (Europe above 60˚ North; and all of Norway)

The data from the European Catchments and RIvers Network System (ECRINS, described in European Environment Agency, 2012) provides water body polygons for Europe at varying spatial scales, estimated to be mostly at 1:250,000 or finer. ECRINS lake data was derived from the Corine Land Cover 2006, XFD Art 13, as well as CCM River and Catchment data.

Most lake polygons provided by ECRINS were included in HydroLAKES without further modifications. Some obvious inconsistencies, such as polygons with pixelated shorelines (likely the result of vectorization processes of coarser remote sensing imagery) were smoothed or replaced with alternative polygons.

GLWD (Global)

The Global Lakes and Wetlands Database (GLWD; Lehner and Döll, 2004) provides a comprehensive global dataset of lake and reservoir shoreline polygons. Most lakes contained in the database are larger than 1 km^2 in surface area, or 0.5 km^3 in storage volume for reservoirs. The original polygons were mostly sourced from the Digital Chart of the World database (at a scale of 1:1 million) supplemented in few cases by other global lake repositories to correct for errors. The spatial resolution, accuracy and precision of the lake polygons contained in GLWD are generally of lower quality than those contained in HydroLAKES. This is partly due to known projection problems which can lead to significant shifts and distortions in the shoreline shapes.

GLWD lake polygons were only added to HydroLAKES as a secondary source to fill some data gaps, or where they provided improved quality to an erroneous HydroLAKES polygon. Some corrections of the shorelines may have been applied, including slight shifts in location, guided by comparisons with remote sensing imagery (including Google Earth).

GRanD (Global)

The Global Reservoir and Dam database (GRanD version 1.1; Lehner et al., 2011), distributed by the Global Water Systems Project, provides a high-resolution and extensively validated global dataset of reservoir polygons and their associated dams. The majority of reservoir polygons contained in GRanD are sourced by SWBD, with some regions covered by independently digitized shorelines.

For the creation of HydroLAKES, the information of 6796 large GRanD polygons (≥ 10 ha) was used to differentiate human-made reservoirs and regulated lakes from natural lakes. Each respective HydroLAKES polygon was assigned the corresponding GRanD-ID to allow linkage between the two databases. In most cases, polygon outlines of HydroLAKES and GRanD are coinciding (as both are primarily sourced from SWBD). In some instances where GRanD provides better or unique polygon information (e.g. for reservoirs built after the year 2000), the polygons from GRanD were used in HydroLAKES. In cases of GRanD showing lower quality, e.g. in Canada where HydroLAKES is based on high-quality national data, some polygons of GRanD were not included but the reservoir attribute and corresponding GRanD-ID was assigned to the HydroLAKES polygon instead. One additional very large reservoir (Eastmain Reservoir, Canada) was identified and flagged in HydroLAKES that is not included in GRanD v1.1.

2.2 Creation and characteristics of HydroLAKES polygon database

<u>General</u>

To create the HydroLAKES database, lake polygons were compiled from a multitude of sources (see Table 1). If the original data were provided in raster format, they were first vectorized using boundary smoothing procedures. Main processing steps in the creation of HydroLAKES included manual identification and removal of river and wetland polygons; removal of duplicates and overlapping polygons; dissolving of segmented polygons into individual lake entities; correction of corrupt or incorrect polygon geometry; removal of small islands (less than 3 ha) within lakes; smoothing of water body shorelines to reduce inconsistencies between datasets of different initial resolution; and establishing a 10 ha (0.10 km^2) cut-off based on lake surface area. More detailed processing steps are provided below and in Table 2.

<u>Definition of lakes and quality of detection</u>

When creating a lake database, the distinction between lakes, rivers and wetlands is a difficult and important issue. In various definitions (e.g., that of the RAMSAR Convention), lakes are generally included within the broader category of wetlands and are then only distinguished based on their permanency or perennial status and depth. Distinctions between rivers and lakes may be similarly difficult based on variations in flow velocity, channel width or depth. While the capability and quality of HydroLAKES in distinguishing "lakes only" has not been verified, the underpinning source datasets used in compiling HydroLAKES are believed to be less prone to confusion than traditional remote-sensing based products which require the development of individual classifications. In general, high uncertainties are expected in any dataset for the transition zones between lakes, rivers and wetlands, or where periodic changes in hydrological processes lead to temporal alterations in these features, such as isolated oxbow lakes turning into active rivers during flood cycles, or lakes converting into wetlands during drier periods.

In terms of distinguishing lakes from rivers, HydroLAKES has undergone extensive manual inspections to remove polygons or parts of polygons that resemble river courses rather than lakes, in particular for the underpinning CanVec data of Canada (see section 2.1 above). Other source datasets used in the creation of HydroLAKES either provided lake-only polygons or, in the case of SWBD and GLWD, included a distinction between lakes and rivers. However, a clear spatial separation between lakes and rivers into unique polygons is highly ambiguous in many places as transitions between them may be fluent or subtle.

In terms of distinguishing lakes from wetlands, it is not the goal of HydroLAKES itself to determine a process-based separation between these features. Rather, HydroLAKES relies on the given distinction provided in the utilized source datasets, all of which contained an explicit "lake" category that by design did not include wetlands. In particular, the CanVec and US NHD datasets were both generated from topographic maps in which wetlands and lakes were distinguished. Similarly, the SWBD polygons used for regions from 56°S to 60°N and the ECRINS database for European lakes over 60°N are both the result of extensive manual post-processing intentionally designed to remove wetlands (e.g., see Slater et al., 2006). SWBD developers compared radar data to ancillary reference data (e.g., water masks from NGA or Landsat Thematic Mapper) to ensure that delineated lakes were not wetlands.

Scale

The resolution of the underpinning source data ranged from 1:24,000 to 1:1 million for the vector data, and from 30 m pixels to 250 m pixels for the raster data. Due to these inconsistencies in scale and the various polygon transformations, smoothing and generalization steps during the map consolidation process, the resulting resolution of the global HydroLAKES database cannot be strictly defined. However, regional comparisons with maps at a variety of known resolutions, as well as tests using shoreline scaling laws as developed by Winslow et al. (2014) suggest the following scales as best approximation: about 1:100,000 for Canada and Alaska (i.e., accounting for two thirds of global lake numbers); about 1:250,000 for Europe and all areas below 60 degrees northern latitude (i.e., accounting for most of the global landmass); and about 1:1 million for the remaining areas (i.e., northern Russia and Greenland).

Co-registration with HydroSHEDS

A spatial co-registration between HydroLAKES and the river network of the HydroSHEDS database (Lehner et al., 2008) was established by linking each lake to the most downstream river pixel that drains the lake. This pour point (or lake outlet) is typically near the lake's shoreline but can also occur near the center of a lake polygon for terminal lakes in endorheic basins.

To create a single pour point for each lake, the cell accumulation values (i.e., the number of upstream pixels as provided by HydroSHEDS, representing a proxy for watershed area) were analyzed within each lake, and the pixel with the maximum value per lake was identified as lake pour point. Where multiple pixels with equal maximum cell accumulations were identified for the same lake, the pixel with highest modeled discharge value was selected, and if there were still multiple pixels, a random choice was made among them. This ultimately resulted in one pour point pixel per lake. Finally, the precise coordinates of the pour point location were calculated as the centroid of the intersection between the lake polygon and the pour point pixel, assuring that all pour points are located inside their corresponding lake polygon.

After creating the lake pour points, HydroSHEDS information, such as upstream catchment area or modeled discharge estimates, were added to the HydroLAKES database by extracting the respective HydroSHEDS values at the pour point locations.

Co-registration with GRanD

HydroLAKES was also co-registered to the Global Reservoir and Dam (GRanD) database (Lehner et al., 2011) by identifying 6796 corresponding polygons and attributing them with the IDs of the GRanD database. These IDs can be used to join additional information from the GRanD database to HydroLAKES, if required.

Lake volume estimates

Messager et al. (2016) used HydroLAKES in combination with the elevation data provided by EarthEnv-DEM90 (Robinson et al., 2014) at 90 m resolution to calculate estimates of average depths for every lake polygon. The selected prediction model applies size-specific multiple regression equations using lake surface area and the average terrain slope within a 100 m buffer surrounding the lake. The equations have been developed based on bathymetric data records for

more than 7000 lakes globally, and the results were tested against independent validation data of more than 5000 lakes. The validation confirmed satisfying regional results, but caution is advised when interpreting the volume of singular lakes as individual errors and uncertainties can be large.

For all lakes with surface areas between 0.1 and 500 km^2 the modeled mean lake depth was multiplied by the lake surface area to obtain lake volume. Due to their increasingly complex bathymetry, lake depth estimate for lakes greater than 500 km^2 were taken from 170 literature sources, with Herdendorf's (1982) global compilation as a main source. Additionally, the storage capacities of 6797 large reservoirs or regulated lakes were added mostly from the Global Reservoir and Dam (GRanD) database (Lehner et al., 2011).

Discharge and residence time estimates

To assign an average discharge estimate for every lake, long-term (1971–2000) average naturalized runoff and discharge data was obtained from the global integrated water model WaterGAP (Döll et al., 2003; model version 2.2 as of 2014). The data was spatially downscaled from its original 0.5 degree (~50 km) pixel resolution to the 15 arc-second (~500 m) resolution of the global HydroSHEDS river network (Lehner et al., 2008) by using geo-statistical approaches (Lehner and Grill, 2013). Preliminary tests against average discharges reported at about 3000 gauging stations provided by the Global Runoff Data Center, Koblenz, Germany, show good overall correlations, yet individual uncertainties may be high (Lehner et al., in prep.).

To assign the downscaled discharge estimates to individual lakes, the values were extracted at the location of each lake pour point. An estimate of the average residence time for each lake was then calculated as the ratio between lake volume and discharge.

Other comments and uncertainties

It should be noted that nearly all datasets used in the development of HydroLAKES were either generated by radar technology (SWBD), analysis of long-term imagery composites (MODIS), or based on topographic maps (e.g., CanVec, US NHD). The effect of clouds is only minor on these types of remote sensing imagery, and absent for topographic maps.

While HydroLAKES is assumed to achieve complete coverage of larger lakes, discrepancies with other existing lake datasets may be due to varying definitions of what constitutes a lake as opposed to rivers or wetlands. Also, some lakes may have changed in their extent (or even disappeared) in recent times, or undergone strong seasonal fluctuations that are not properly represented by the temporal snapshot provided in HydroLAKES. Finally, there may be some confusion in interpreting connected pools in close vicinity as one or multiple lakes. As a prominent example, Lakes Michigan and Huron have been split into two lakes—despite their natural connection—in order to conform with the general convention of treating them as separate units.

The level of completeness of HydroLAKES at the lower size limit of 10 ha is difficult to verify. Messager et al. (2016) provide a statistical extrapolation towards smaller lakes which indicates that HydroLAKES achieves virtually full completion for lakes above 35 ha and close to full completion for lakes between 10 and 35 ha.

3. Data specifications, format and distribution

3.1 Vector data format and distribution

HydroLAKES consists of two separate GIS layers:

- 'HydroLAKES_polys_v10' contains all lake shoreline polygons
- 'HydroLAKES_points_v10' contains all lake pour points

Each HydroLAKES layer is provided in two different formats:

- Within an ESRI® geodatabase
- As a stand-alone ESRI® shapefile

All versions contain the same attribute information, i.e. all lake polygons or pour points have the same columns in their respective attribute tables (see Table 2 below).

The geodatabase version requires ESRI® ArcGIS (or compatible) software to be opened. All geometric, projection, and attribute information is contained within the geodatabase.

The shapefile version is provided as it is readable by a larger variety of software products, including open source GIS packages. Each HydroLAKES shapefile consists of five main files (.dbf, .sbn, .sbx, .shp, .shx) and projection information is provided in an ASCII text file (.prj).

NOTE: Users without GIS software or without the option to interpret shapefiles may import the file 'HydroLAKES_points_v10.dbf' (in dBASE IV format) in alternative spreadsheet or database programs that are capable of reading 1.4 million rows. This file contains all attribute information of HydroLAKES, and the pour point locations can be plotted using the provided XY coordinates.

HydroLAKES data is available electronically in compressed zip file format from http://www.hydrosheds.org. To use the data files, the zip files must first be decompressed. Each zip file includes a copy of the HydroLAKES Technical Documentation.

3.2 Data projection

The HydroLAKES layers are provided in geographic (latitude/longitude) projection, referenced to datum WGS84. In ESRI® software this projection is defined by the geographic coordinate system GCS_WGS_1984 and datum D_WGS_1984. The projection information is provided as part of the distributed data layers.

3.3 Attribute table

Table 2 shows the column structure and information contained in the attribute table associated with both the geodatabase and the shapefile format of HydroLAKES. Note that in the geodatabase format the fields 'OBJECTID' as well as 'Shape_Length' and Shape_Area' are added by default by the ArcGIS software—these fields are not officially part of HydroLAKES.

Table 2: Attribute table of HydroLAKES polygon and point layers.

Column	Description
Hylak_id	Unique lake identifier. Values range from 1 to 1,427,688.
Lake_name	Name of lake or reservoir. This field is currently only populated for lakes with an area of at least 500 km^2; for large reservoirs where a name was available in the GRanD database; and for smaller lakes where a name was available in the GLWD database.
Country	Country that the lake (or reservoir) is located in. International or transboundary lakes are assigned to the country in which its corresponding lake pour point is located and may be arbitrary for pour points that fall on country boundaries.
Continent	Continent that the lake (or reservoir) is located in. Geographic continent: Africa, Asia, Europe, North America, South America, or Oceania (Australia and Pacific Islands)
Poly_src	Source of original lake polygon: CanVec; SWBD; MODIS; NHD; ECRINS; GLWD; GRanD; or Other More information on these data sources can be found in Table 1.
Lake_type	Indicator for lake type: 1: Lake 2: Reservoir 3: Lake control (i.e. natural lake with regulation structure) Note that the default value for all water bodies is 1, and only those water bodies explicitly identified as other types (mostly based on information from the GRanD database) have other values; hence the type 'Lake' also includes all unidentified smaller human-made reservoirs and regulated lakes.
Grand_id	ID of the corresponding reservoir in the GRanD database, or value 0 for no corresponding GRanD record. This field can be used to join additional attributes from the GRanD database.
Lake_area	Lake surface area (i.e. polygon area), in square kilometers.
Shore_len	Length of shoreline (i.e. polygon outline), in kilometers.

Column	Description
Shore_dev	Shoreline development, measured as the ratio between shoreline length and the circumference of a circle with the same area. A lake with the shape of a perfect circle has a shoreline development of 1, while higher values indicate increasing shoreline complexity.
Vol_total	Total lake or reservoir volume, in million cubic meters (1 mcm = 0.001 km^3). For most polygons, this value represents the total lake volume as estimated using the geostatistical modeling approach by Messager et al. (2016). However, where either a reported lake volume (for lakes ≥ 500 km^2) or a reported reservoir volume (from GRanD database) existed, the total volume represents this reported value. In cases of regulated lakes, the total volume represents the larger value between reported reservoir and modeled or reported lake volume. Column 'Vol_src' provides additional information regarding these distinctions.
Vol_res	Reported reservoir volume, or storage volume of added lake regulation, in million cubic meters (1 mcm = 0.001 km^3). 0: no reservoir volume
Vol_src	1: 'Vol_total' is the reported total lake volume from literature 2: 'Vol_total' is the reported total reservoir volume from GRanD or literature 3: 'Vol_total' is the estimated total lake volume using the geostatistical modeling approach by Messager et al. (2016)
Depth_avg	Average lake depth, in meters. Average lake depth is defined as the ratio between total lake volume ('Vol_total') and lake area ('Lake_area').
Dis_avg	Average long-term discharge flowing through the lake, in cubic meters per second. This value is derived from modeled runoff and discharge estimates provided by the global hydrological model WaterGAP, downscaled to the 15 arc-second resolution of HydroSHEDS (see section 2.2 for more details) and is extracted at the location of the lake pour point. Note that these model estimates contain considerable uncertainty, in particular for very low flows. -9999: no data as lake pour point is not on HydroSHEDS landmask
Res_time	Average residence time of the lake water, in days. The average residence time is calculated as the ratio between total lake volume ('Vol_total') and average long-term discharge ('Dis_avg'). Values below 0.1 are rounded up to 0.1 as shorter residence times seem implausible (and likely indicate model errors). -1: cannot be calculated as 'Dis_avg' is 0 -9999: no data as lake pour point is not on HydroSHEDS landmask

Column	Description
Elevation	Elevation of lake surface, in meters above sea level. This value was primarily derived from the EarthEnv-DEM90 digital elevation model at 90 m pixel resolution by calculating the majority pixel elevation found within the lake boundaries. To remove some artefacts inherent in this DEM for northern latitudes, all lake values that showed negative elevation for the area north of 60°N were substituted with results using the coarser GTOPO30 DEM of USGS at 1 km pixel resolution, which ensures land surfaces ≥0 in this region. Note that due to the remaining uncertainties in the EarthEnv-DEM90 some small negative values occur along the global ocean coastline south of 60°N which may or may not be correct.
Slope_100	Average slope within a 100 meter buffer around the lake polygon, in degrees. This value is derived from the EarthEnv-DEM90 digital elevation model at 90 m pixel resolution. Slopes for each pixel were computed with latitudinal corrections for the distortion in the XY spacing of geographic coordinates by approximating the geodesic distance between cell centers. For 12 lakes located above the northern limit of the EarthEnv-DEM90 digital elevation model (83°N), slopes were computed from the GTOPO30 DEM of USGS at 1 km pixel resolution. -1: slope values were not calculated for the largest lakes (polygon area ≥ 500 km^2)
Wshd_area	Area of the watershed associated with the lake, in square kilometers. The watershed area is calculated by deriving and measuring the upstream contribution area to the lake pour point using the HydroSHEDS drainage network map at 15 arc-second resolution. -9999: no data as lake pour point is not on HydroSHEDS landmask
Pour_long	Longitude of the lake pour point, in decimal degrees.
Pour_lat	Latitude of the lake pour point, in decimal degrees.

4. License, disclaimer and acknowledgement

4.1 License agreement

The HydroLAKES database (version 1.a) is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0/. By downloading and using the data the user agrees to the terms and conditions of this license. Notwithstanding this free license, we ask users to refrain from redistributing the data in whole in its original format on other websites without the explicit written permission from the authors. HydroLAKES is publicly available for download at http://www.hydrosheds.org. The copyright © of HydroLAKES is held by the authors, 2016, all rights reserved.

4.2 Disclaimer of warranty

The HydroLAKES database and any related materials contained therein are provided "as is," without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose, noninterference, system integration, or noninfringement. The entire risk of use of the data shall be with the user. The user expressly acknowledges that the data may contain some nonconformities, defects, or errors. The authors do not warrant that the data will meet the user's needs or expectations, that the use of the data will be uninterrupted, or that all nonconformities, defects, or errors can or will be corrected. The authors are not inviting reliance on these data, and the user should always verify actual data.

4.3 Limitation of liability

In no event shall the authors be liable for costs of procurement of substitute goods or services, lost profits, lost sales or business expenditures, investments, or commitments in connection with any business, loss of any goodwill, or for any direct, indirect, special, incidental, exemplary, or consequential damages arising out of the use of the HydroLAKES database and any related materials, however caused, on any theory of liability, and whether or not the authors have been advised of the possibility of such damage. These limitations shall apply notwithstanding any failure of essential purpose of any exclusive remedy.

4.4 Acknowledgement and citation

We kindly ask users to cite HydroLAKES in any published material produced using the data. If possible, online links to the hosting website (http://www.hydrosheds.org) should be provided.

If HydroLAKES represents a crucial component to the research of a user, or if an important research result or conclusion depends on it, we kindly request that the user offers co-authorship to (a representative of) the authors of HydroLAKES. If in doubt, please contact the corresponding author at bernhard.lehner@mcgill.ca.

Citations and acknowledgements of the HydroLAKES data should be made as follows:

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O. (2016): Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature Communications: 13603. doi: 10.1038/ncomms13603. Data is available at www.hydrosheds.org.

5. References

Carroll, M. L., J. R. Townshend, C. M. DiMiceli, P. Noojipady, and R. A. Sohlberg (2009), A new global raster water mask at 250 m resolution, *International Journal of Digital Earth*, *2*(4), 291-308, doi: 10.1080/17538940902951401. Data available online (http://glcf.umd.edu/data/watermask/), accessed May 2014, edited.

Döll, P., F. Kaspar, and B. Lehner (2003), A global hydrological model for deriving water availability indicators: model tuning and validation, *Journal of Hydrology*, *270*(1), 105-134.

European Environment Agency (2012), ECRINS (European Catchments and RIvers Network System): Lakes. Version 1, Data available online (http://projects.eionet.europa.eu/ecrins), accessed June 2013, edited.

Herdendorf, C. E. (1982), Large Lakes of the World, *Journal of Great Lakes Research*, *8*(3), 379-412, doi: http://dx.doi.org/10.1016/S0380-1330(82)71982-3.

Lehner, B., and P. Döll (2004), Development and validation of a global database of lakes, reservoirs and wetlands, *Journal of Hydrology*, *296*(1), 1-22, doi: 10.1016/j.jhydrol.2004.03.028.

Lehner, B., and G. Grill (2013), Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, *Hydrological Processes*, *27*(15), 2171-2186, doi: 10.1002/hyp.9740.

Lehner, B., K. Verdin, and A. Jarvis (2008), New global hydrography derived from spaceborne elevation data, *EOS*, *89*(10), 93-94, doi: 10.1029/2008EO100001.

Lehner, B., C. R. Liermann, C. Revenga, C. Vörösmarty, B. Fekete, P. Crouzet, P. Döll, M. Endejan, K. Frenken, and J. Magome (2011), High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, *Frontiers in Ecology and the Environment*, *9*(9), 494-502, doi: 10.1890/100125.

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. *Nature Communications*, 13603, doi: 10.1038/ncomms13603. Data is available at www.hydrosheds.org.

Natural Resources Canada (2013), CanVec Hydrography: Waterbody Features. Version 12.0. Data available online (ftp://ftp2.cits.rncan.gc.ca/pub/canvec/), accessed May 2013, edited.

Robinson, N., J. Regetz, and R. P. Guralnick (2014), EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data, *ISPRS Journal of Photogrammetry and Remote Sensing*, *87*, 57-67, doi: 10.1016/j.isprsjprs.2013.11.002.

Slater, J. A., G. Garvey, C. Johnston, J. Haase, B. Heady, G. Kroenung, and J. Little (2006), The SRTM data "finishing" process and products, *Photogrammetric Engineering & Remote Sensing*, *72*(3), 237-247, doi: http://dx.doi.org/10.14358/PERS.72.3.237.

U.S. Geological Survey (2013), National Hydrography Geodatabase: Alaska. Data available via http://nhd.usgs.gov/ or The National Map viewer (http://viewer.nationalmap.gov/viewer/nhd.html?p=nhd), accessed July 2013, edited.

Winslow, L. A., J. S. Read, P. C. Hanson, and E.H. Stanley (2014), Lake shoreline in the contiguous United States: quantity, distribution and sensitivity to observation resolution. *Freshwater Biology*, 59, 213-223, doi:10.1111/fwb.12258.