GLADD-R: A new Global Lake Dynamics Database for Reservoirs created using machine learning and satellite data

Ankush Khandelwal, Anuj Karpatne, Zhihao Wei, Huangying Kuang, Rahul Ghosh, Hilary Dugan, Paul Hanson, Vipin Kumar

April 1, 2019
GLADD-R: A new Global Lake Dynamics Database for Reservoirs

Ankush Khandelwal*, Rahul Ghosh, Zhihao Wei, Huangying Kuang and Vipin Kumar
Department of Computer Science and Engineering, University of Minnesota

Hilary Dugan and Paul Hanson
Center for Limnology, University of Wisconsin-Madison

Anuj Karpatne
Department of Computer Science, Virginia Tech

Reservoirs play a crucial role for human sustenance as they provide freshwater for agriculture, power generation, human consumption, and recreation. A global database of reservoirs that provides their location and dynamics can be of great importance to the ecological community as it enables the study of the impact of human actions and climate change on fresh water availability. Here we present a new database, GLADD-R (Global Lake Dynamics Database-Reservoirs) that provides such information for 1882 reservoirs between 1 and 100 square kilometers in size that were created after 1985. The visualization of these reservoirs and their surface area time series is available at http://umnlcc.cs.umn.edu/GlobalReservoirDatabase/

I. Introduction

Reservoirs (dams and impoundments) have been constructed for millennia to provide persistent fresh water for agriculture and aquaculture, industry, human consumption, flood mitigation, reliable navigation, energy production, waste disposal, and recreation. The need for accessible and high-quality surface water has grown with the changing needs of civilization. Few human alterations of the Earth’s water cycle rival the impacts of reservoir construction, including the unintended negative effects on water quality and contamination, habitability to native species, fish migration, and flooding disasters when infrastructure fails. A global database of reservoirs that provides their location and dynamics can be of great importance to the ecological community as it can enable the study of the impact of human actions and climate change on fresh water availability.

Currently, GRanD database [1] is the largest database that provides information on reservoirs globally. The first version of GRanD (v1.1) was released in 2011 which provided the locations of 6862 reservoirs and a static snapshot of reservoir’s attributes such as dam height, depth of the reservoir, average discharge, average surface area, and reference shape. A new version of the database (v1.3) was released in February 2019 and it provides information for an additional 458 reservoirs (7320 reservoirs in total). The database was created through manual curation effort which impact its completeness and makes it difficult to update over time. Moreover, the database does not provide temporal information about their surface level dynamics.

Here we describe a new database, GLADD-R (Global Lake Dynamic Database for Reservoirs) that has been created by analyzing satellite imagery via automated machine learning techniques. There are five main components of the processing pipeline: a) estimate land/water labels at different timesteps using global satellite imagery data, b) extract individual water bodies from the multi-temporal pixel based land/water label, c) handle missing data and labeling errors to obtain more accurate and complete surface extents of water bodies at each timestep for which land/water labels are available, d) analyze surface area time series of individual water bodies to automatically identify candidate reservoirs that show a sudden increase in their surface area soon after their construction, and e) manually verify candidate reservoirs, i.e., locate the dam wall using high resolution imagery.

While GLADD-R will continue to evolve, the current version, GLADD-R-1.0, was created using LANDSAT based land/water label provided by the JRC product [2] for the period 1984 to 2015 at monthly temporal scale. To handle the issue of missing and incorrect labels, we used new machine learning techniques [3–5] that incorporate physical constraints in lake area dynamics to produce more accurate and complete dynamics of water bodies at global scale.

*khand035@umn.edu
In order to identify sudden increase in the area of the water body, a minimum time-window of two years was needed before and after the date when a dam became operational. Hence, GLADD-R reports dam construction activity between January 1986 and December 2012 even though the surface area dynamics is available from March 1984 till October 2015. As an illustrative example, Figure [a](a) shows surface area dynamics of a reservoir on the Sambito River in Brazil. The sudden increase in area is evident as the surface area of the reservoir increased from 0 to approximately 12 square kilometers in just three months. The surface area time series also makes it easy to identify seasonal changes and reduction in the size of the reservoir over time. Figure [a](c) highlights the utility of the label correction step. The red line represents surface area time series created using JRC labels without correction. The area variations show a lot of spurious fluctuations due to missing data and labeling errors which were removed in the corrected time series.

GLADD-R-1.0 provides information for reservoirs of size between 1 and 100 square kilometers. We are in the process of analyzing reservoirs with size between 0.1 and 1 sq. kms. and those larger than 100 sq. kms. These will be available in the next version of GLADD-R.

**II. GLADD-R: Highlights**

GLADD-R-1.0 provides location and surface area dynamics of 1882 reservoirs built between January 1984 and December 2012 globally. Out of 1882 reservoirs reported in GLADD-R, only 415 were also reported in GRanD. Thus, GLADD-R provides information about an additional 1467 reservoirs that are not in GRanD. This highlights the utility of the automated machine learning approach to creating such a database on a global scale with minimum manual effort.

Figure 2 (a) shows the distribution of these reservoirs across different continents while Figure 2 (b) shows the cumulative distribution of the number of reservoirs constructed after 1985 in different continents. The majority of dam
construction has occurred in Asia and South America since 1986, and the rate of construction in North America has declined significantly. Figure 2(c) shows the continent-wide distribution of the 1467 reservoirs that are unique to GLADD-R and Figure 2(d) shows the corresponding cumulative distribution.

![Fig. 2](image)

While GRanD database only provides static information about the extent of reservoirs. In contrast, GLADD-R-1.0 also provides surface area at monthly scale from March 1984 to October 2015. Figure 3 shows the aggregate surface area variation of reservoirs globally. Due to the high prevalence of missing data in JRC label before 2000, the surface area at different time steps during this period can be much lower than the actual area. Hence, the dynamics before 1999 are shown in light grey color to signify less data availability. To provide a baseline of reservoir storage from reservoirs created prior to 1986, we processed a subset of 4142 reservoirs that were reported in GRanD and were created before 1986. At global scale, we can see that surface area in reservoirs continued to increase after 2000 as more reservoirs were constructed and approximately 8,000 sq. kms. of surface area has been added. Furthermore, there has been a reduction in surface area after 2012 until the end of the study period.

Figure 4 shows the aggregate surface area dynamics for each continent separately. Different continents show very different variations in surface area over the study period. Asia has the most number of dams and also the largest aggregate surface area. Even though South America has a greater number of dams, reservoirs in North America have more total surface area. All continents show strong seasonality in area, and all continents other than Europe show the decreasing trend from 2011-2015.
Fig. 3 Aggregate surface area dynamics of reservoirs globally. The black line represents aggregate surface area of a subset of reservoirs (4142) reported in GRanD that were built before 1986 with size between 1 and 100 sq. kms. The red line represents the aggregate surface area of 4142 old reservoirs and additional 1882 reservoirs created after 1985 that are part of GLADD-R.

Fig. 4 Aggregate surface area dynamics of reservoirs across different continents. The black line represents aggregate surface area of a subset of reservoirs reported in GRanD that were built before 1986 with size between 1 and 100 sq. kms. The red line represents the aggregate surface area of the old reservoirs and additional reservoirs created after 1985 that are part of GLADD-R.
III. Methods

GLADD-R was created by using automated machine learning approaches for analyzing satellite imagery datasets. The processing pipeline has the following five components:

A. Pixel based land/water label generation

This step involves classification of satellite imagery data to produce land/water label at different timesteps. In the current version, we used an existing dataset, henceforth referred to as JRC dataset, which has been created through collaboration between European Commission’s Joint Research Center and Google [2]. The JRC dataset was created by analyzing the entire LANDSAT archive from March 1984 till October 2015. For each month a global land/water mask is available where pixels are labeled as either land, water or unknown.

B. Lake Polygons Database Generation

To identify locations and reference shape of lakes around the world, we performed connected component analysis on the JRC dataset’s “recurrence” layer. The recurrence layer provides a number between 0 and 100 for each pixel, which represents the percentage of months the pixels was observed as water. We first binarized the layer by selecting pixels with percentage value greater than 10. Using the value 10 ensured that spuriously labelled pixels will be not selected. Once the binary layer is obtained, we performed a connected component analysis and assigned each connected component as a water body in our database. Using these reference shapes, we extracted pixel-based land/water label at monthly scale for each lake individually. To avoid including other nearby lakes in the buffer, we further prune the buffer region using an automated approach as described in [3].

C. Label Correction

If the pixel-based land/water label were accurate and complete, just counting the number of water pixels for each month would have provided area and its variation at the lake level. However, these maps tend to suffer from large amounts of missing data and labelling errors. Thus, these land/water label cannot be used directly to obtain robust surface area dynamics. To overcome this challenge, we developed a new machine learning methodology that uses physical principles governing the lake dynamics to correct and impute labels [3-5]. The key idea behind the methodology is that growing and shrinking of lakes is determined by their bathymetry and surrounding topography. In other words, a pixel with higher elevation cannot be filled until all the pixels with lower elevation are filled. This constraint provides a very robust way to correct as well as to impute missing labels. Since in real world scenarios, bathymetry information is seldom available, the proposed machine learning technique estimates the bathymetry from the multi-temporal imperfect land/water labels and uses it to improve the labels iteratively [4]. This step is the most crucial step for achieving robust land/water labels. As an illustrative example, Figure 5 shows labels before and after correction for lake Naivasha in Kenya in February 2012.

![Fig. 5](image)

**Fig. 5** An illustrative example showing utility of the label correction step. Blue color represents water, green represents land, yellow represents pixels out of buffer region, and red represents missing labels.
D. Identification of Candidate Reservoirs

Once the improved land/water labels are obtained for each lake, we count the number of water pixels for each month to create surface area time series for each water body. For each timestep in a time series, a score is computed which reflects the sudden and persistent increase in surface area values around that timestep. The maximum score across all timestep of a timeseries is used as an indicator of the reservoir construction activity. All the water bodies that had the score greater than a certain threshold are considered as candidate reservoirs. To ensure reliable estimation of sudden increase in the area of the water body, a minimum time-window of two years is used before and after the timestep under consideration. Due to this constraint, GLADD-R reports dam construction activity between January 1986 and December 2012 even though the surface area dynamics is available from March 1984 till October 2015.

E. Manual Verification of Candidate Reservoirs

All candidate reservoirs were manually verified by visual inspection using high resolution satellite imagery. Specifically, we looked for a dam wall or an impoundment wall to prune our candidate set of reservoirs. In some cases, especially reservoirs built for mining, agriculture, or just as a lake in a residential neighborhood, such a barrier is not visible. But even in these cases, we were able to verify the sudden appearance of the reservoir using Google Timelapse.

IV. Ecological and Hydrological Significance

There are thousands of reservoirs on Earth, and very little empirical data for making sound management decisions based on understanding the ecological dynamics of these systems. Just like large rivers, there are significant physical, biological, and economic consequences of changing water levels in reservoirs. Floods destroy infrastructure and deteriorate water quality, while droughts reduce habitat for important fisheries and diminish water supplies. The GLADD-R database provides the necessary data to build mechanistic, and predictive models for estimating reservoir water level. With associated reservoir depth information, gathered from GRanD, water level models can be leveraged to predict water volume and residence time. In this sense, GLADD-R is complementary to GRanD. Also, GLADD-R has identified new reservoirs, for which GRanD may be interested in collecting relevant attributes from respective countries and/or local authorities.

V. Data Availability

Location and timeseries information of these reservoirs is available at http://umnlcc.cs.umn.edu/GlobalReservoirDatabase/. This online interface provides locations and shapes of all the reservoirs in GLADD-R. For each reservoir, its surface area time series can be visualized by clicking on the point on the map. The viewer also makes it easy to see the time lapse view of the reservoir, which allows instant visual verification of the year of reservoir construction. The viewer also provides the surface area time series of 4142 reservoirs reported in GRanD that were built before 1986 with size between 1 and 100 sq. kms.

VI. Future Updates

As we improve the methodology (pixel labelling, label correction, and reservoir detection), more reservoirs of size between 1 and 100 sq. kms. will be added in the next version. We are also in the process of analyzing water bodies of size between 0.1 and 1 sq. kms. and those of size greater than 100 square kilometers. The next version of GLADD-R will provide the location and dynamics of reservoirs in these size ranges as well. Our other future goals include providing continuous monitoring of reservoirs globally using LANDSAT, MODIS, and Sentinel imagery datasets. This will also enable us to identify new reservoirs created after the JRC product period. In addition, we aim to provide the dynamics of all the reservoirs that were created before the JRC product period using the information from the GRanD database (our web visualization provides information of 4142 such reservoirs currently).

References


