ReaLSAT-R: A new Reservoir Surface Area Dynamics Database created using Machine Learning and Satellite Imagery

Ankush Khandelwal
khand035@umn.edu
University of Minnesota

Anuj Karpatne
karpatne@vt.edu
Virginia Tech

Zhihao Wei
zhihaowei@emails.bjut.edu.cn
Beijing University of Technology

Rahul Ghosh
ghosh128@umn.edu
University of Minnesota

Huangying Kuang
kuang048@umn.edu
University of Minnesota

Kelly Cutler
lind0436@umn.edu
University of Minnesota

Hilary Dugan
hdugan@wisc.edu
University of Wisconsin-Madison

Paul Hanson
pchanson@wisc.edu
University of Wisconsin-Madison

Vipin Kumar
kumar001@umn.edu
University of Minnesota

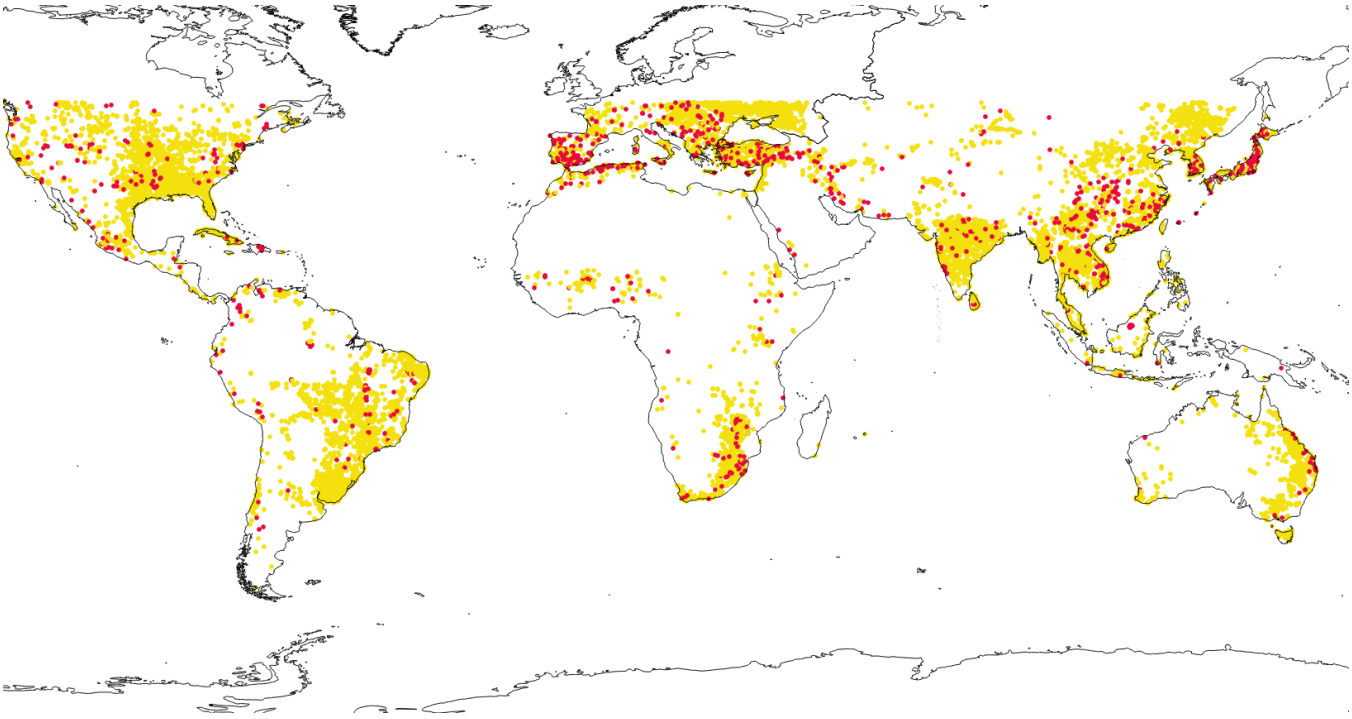

Figure 1: Geographical location of reservoirs in ReaLSAT-R database. Red points (853) represents reservoirs that were reported in GRanD as well. Yellow points (14620) represents reservoirs that are unique to ReaLSAT-R.

ABSTRACT

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. This paper presents a new database, ReaLSAT-R (Reservoir and Lake Surface Area Timeseries-Reservoirs) that has been created by analyzing spectral data from Earth Observation (EO) Satellites using novel machine learning (ML) techniques. These ML techniques can construct highly accurate surface area extents of water bodies at regular intervals despite the challenges arising from heterogeneity and missing or poor quality spectral data. The ReaLSAT-R database provides information for 15473 reservoirs between 0.1 and 100 square kilometers in size that were created after 1984. The number of reservoirs identified in ReaLSAT-R is substantially larger than those available in GRanD, which is the state of the art database maintained by the ecological community, and contains only the static shape of a

ICLR '20 Workshop on AI for Earth Sciences, April 26th, 2020

reservoir. ReaLSAT-R contains the surface area time series for new reservoirs created since 1984 as well as 3274 others that were created before 1984 (reported in GRanD database), thus permitting a global view on the state of these reservoirs that are being impacted by changing climate and human actions. The visualization of these reservoirs and their surface area time series is available online via ReaLSAT-R web interface. This paper reports the summary of the second version of the database. The database was formerly called as GLADD-R (Global Lake Dynamics Database-Reservoirs) and its first version was presented in KDD '19 Workshop on Data Mining and AI for Conservation.

KEYWORDS

physics guided machine learning, rank aggregation, global reservoir dataset, surface water dynamics, change detection

1 INTRODUCTION

Reservoirs (dams and impoundments) have been constructed for millennia to provide persistent fresh water for agriculture, 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 the 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 [6] is the largest database that provides information on reservoirs globally. The first version of this database (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 area dynamics.

This paper presents a new database, ReaLSAT-R (Reservoir and Lake Surface Area Timeseries-Reservoirs) that has been created by analyzing multi-spectral data from Earth Observation (EO) Satellites using novel machine learning (ML) techniques. Earth Observation datasets, which are available globally at regular time intervals, not only enable monitoring of surface area dynamics of water bodies but can also be used to identify construction of new reservoirs automatically at a global scale. In particular, spatially and temporally explicit EO data can be used to label each pixel on Earth as either land or water at any given timestep using state-of-the-art machine learning approaches. If these labels are perfect, then they can be monitored over time to track changes in surface water at a location or a region of interest. However, despite their promise, machine learning methods suffer from a number of challenges when applied to global scale EO data, leading to erroneous as well as missing class labels. First, EO datasets are generally plagued with noise, outliers and missing data, due to sensor anomalies and atmospheric disturbances such as clouds, aerosols and sun angle. Second, even without the above data acquisition related issues, remote sensing data might not be able to distinguish certain classes, such as algae on water, as they appear similar to land. Third, these challenges become even more severe at global scale due to high heterogeneity in the data, as locations with same input values can belong to either land or water depending on their geographical context. Hence, the state-of-the-art machine learning methods for creating water extent maps show unsatisfactory performance especially in the context of identifying dynamics of water bodies at a global scale [1].

To overcome the aforementioned challenges, a novel machine learning framework, ORBIT (Ordering Based Information Transfer) [2–5] was developed. This framework (described in more detail in Section 2) makes use of the inherent ordering constraint among pixels due to the earth's topography/elevation. The elevation ordering based constraint enables the framework to identify and correct physically inconsistent labels. The ORBIT framework [2, 3] was used to construct the ReaLSAT database.

This paper reports the summary of the second version of the database. The database was formerly called as GLADD-R (Global Lake Dynamics Database-Reservoirs) and its first version was presented in KDD '19 Workshop on Data Mining and AI for Conservation. The version 2.0 of the ReaLSAT-R database provides information for 16651 reservoirs between 0.1 and 100 square kilometers in size that were created after 1984. The number of new reservoirs identified in ReaLSAT-R is substantially larger than those available in GRanD (only 853 out of 15473 reservoirs are reported in GRanD). Note that GRanD contains only static shape of the reservoir. In contrast, ReaLSAT-R web interface contains the surface area time series for all new reservoirs created after 1984 as well as 3274 others that were created before 1984 (reported in GRanD database), thus permitting a global view on the state of these reservoirs that are being impacted by changing climate and human actions.

ReaLSAT-R-2.0, was created using LANDSAT based land/water classification maps available for the period 1984 to 2015 at monthly temporal scale [7]. Even though this pixel based classification product (henceforth referred to as GSW dataset) is considered as the state-of-the-art, it suffers from significant classification errors and missing data due to aforementioned challenges in analyzing EO data. Moreover, GSW dataset provides information at pixel level and thus surface area variations of individual water bodies are not readily available. ORBIT framework provides a robust way to create more accurate and complete classification maps for individual water bodies using erroneous and incomplete pixel based classification products. As an illustrative example, Figure 2 shows labels before and after correction for lake Naivasha in Kenya in February 2012.

Reservoirs show a very specific pattern in their surface area variation which can be used to identify them automatically. Specifically, reservoirs show a sudden increase in their surface area after they become operational and this increase tends to persist over time. Thus, after obtaining high quality classification maps for individual water bodies, time series analysis can be performed to automatically distinguish reservoirs from natural lakes. As an illustrative example, Figure 3 (top) shows surface area dynamics of a reservoir on the

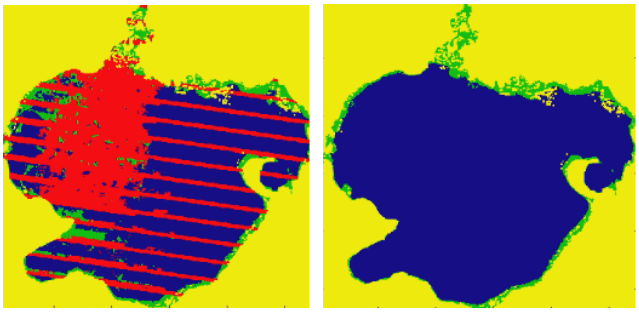

Figure 2: An illustrative example showing utility of the label correction step. Blue color represents water, green represents land, yellow represents pixels out of buffer region, and red represents missing labels. (left) Classification labels from GSW dataset. (right) Classification labels after correction using ORBIT framework

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.

The rest of the paper is as follows: Section 2 provides an overview of the ORBIT framework. Section 3 provides a summary of the processing pipeline used to create the ReaLSAT-R database. Section 4 provides some highlights of the ReaLSAT-R database and finally Section ?? concludes the paper with upcoming future updates.

2 THE ORBIT FRAMEWORK

The ORBIT framework makes use of the inherent ordering constraint among instances/pixels to improve the accuracy of classification maps. The key idea is the following - if a location is filled with water then by laws of physics all the locations in the basin that have lower elevation should also be filled with water. Thus, physically inconsistent labels that do not adhere to this physical constraint can be detected.

Figure 4 illustrates the utility of this constraint using a toy example. Given an elevation ordering (π) and a set of potentially erroneous labels at any given time step t, the aim is to estimate correct labels that are physically consistent with the elevation ordering. For a given elevation ordering of N instances, there are only N + 1 possible sets of labels that are physically consistent. For example, Figure 4 (b) shows 8 possible sets of physically consistent labels for 7 locations shown in Figure 4 (a). In the absence of any external information about these labels, ORBIT framework adopts the maximum likelihood estimation approach. Specifically, it makes an assumption that majority of the input labels are correct and hence selects the set of physically consistent labels that matches the most with input labels. For example, Figure 4 (c) shows the erroneous input labels and 4 (d) shows the selected set that matches the most with input labels. In this illustrative example, location F is detected as erroneous and its label is changed from water to land.

Note that good quality elevation information is not explicitly available for most water bodies in the world. To overcome this challenge, ORBIT framework uses a rank aggregation based strategy

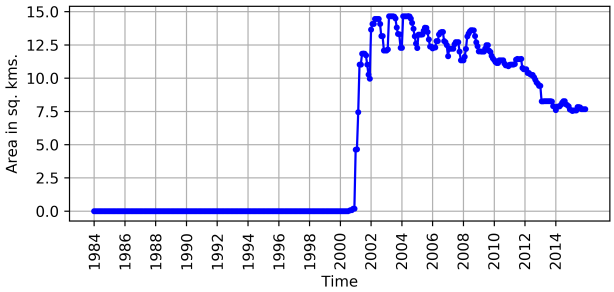

Figure 3: An illustrative example of surface area dynamics of a reservoir on Sambito river in Brazil (latitude: -6.180322 , longitude: -41.978494). (top) Surface area time series using ReaLSAT-R methodology. (bottom) High resolution aerial imagery of the reservoir. The zoomed-in inset shows the dam of the reservoir.

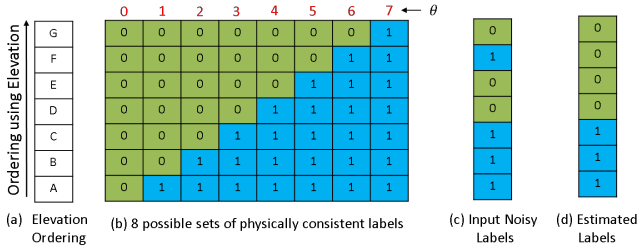

Figure 4: An illustrative example showing elevation ordering based label correction process

[5] to simultaneously estimate inherent elevation ordering and physically consistent labels using an Expectation-Maximization framework.

Furthermore, in most situations, a water body grows and shrinks smoothly (except sudden events such as floods) i.e. surface extents of nearby dates are likely to be very similar. Hence, incorporating temporal context in the label correction process can lead to further

improvement in the label accuracy. Current state-of-the-art methods mainly enforce the temporal consistency either for each pixel individually (e.g. majority filters in time) or use a given pixel's temporal and spatial neighborhood to obtain temporal consistent labels. As shown in [5], these methods perform poorly when noise and missing data is also spatially and temporally auto-correlated which is very common in our application. Moreover, existing methods tend to remove real changes in labels as well because they enforce labels in nearby time steps to be same.

ORBIT framework [2] uses elevation ordering to enforce temporal consistency in total area values instead of consistency in labels of individual pixels. Temporal consistency in total area (surface extent) is a more realistic constraint and it also preserves real dynamics better than existing methods.

3 PROCESSING PIPELINE

As mentioned earlier, existing pixel based classification products do not provide information for individual water bodies separately. In this section, we describe the processing pipeline that was used to create high quality surface area dynamics of individual water bodies from erroneous pixel based information. The high quality surface area dynamics was then used to distinguish reservoirs from natural lakes.

3.1 Pixel based land/water label generation

This step involves classification of EO data to produce land/water label at different timesteps. In the current version, we used the GSW dataset as the source of pixel based classification maps. [7]. The GSW 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. The GSW dataset is the state-of-the-art classification product at LANDSAT scale. The algorithm uses a decision tree framework to assign each pixel to one of the three categories. Instead of training decision rules from the data directly, the authors used visual analytics and human in the loop strategy to identify cluster hulls in the feature space (which includes raw multispectral image bands and derived indices used by remote sensing community) to delineate regions belonging to different classes. These cluster hulls were then converted into equations for the decision tree. Ancillary data such as glacier masks, lava mask, mountain shadow mask, and cloud mask, were used to remove potentially false water labels. It is worth noting that ORBIT framework is not dependent on GSW dataset. If a new multi-temporal product is released in future, ORBIT framework can be applied on top of the new dataset as well.

3.2 Lake Polygons Database Generation

To identify locations and reference shape of lakes around the world, we performed connected component analysis on the GSW dataset's "occurrence" layer. The "occurrence" 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. The threshold value of 10 was used to avoid spuriously labelled pixels from being considered as potential water bodies. Once the binary layer is obtained, we performed a connected component analysis and considered each connected component as a water body in our database. Figure 5 illustrates the database generation process on a small region in USA. The top image shows the GSW dataset's "occurrence" layer. The color scheme goes from light blue to dark blue as the "occurrence" layer value increases from 0 to 100. The middle figure shows the binary mask created by thresholding the top image. Finally, this binary mask is used to extract individual connected components (sets of contiguous pixels). In this image, each connected component is shown in a different color.

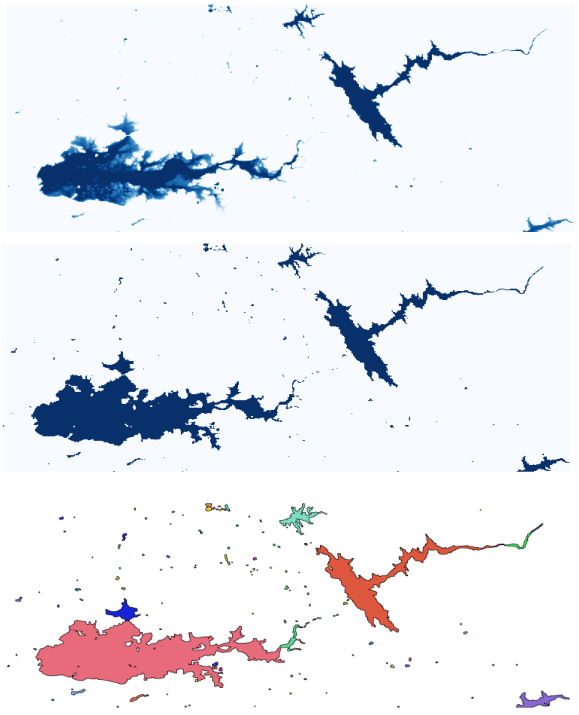

Figure 5: An illustrative example showing the database generation process on a small region in USA. (top) GSW dataset's "occurrence" layer. (middle) binary mask created by thresholding the top image. (bottom) connected component image where each component is being shown in a different color.

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].

3.3 Label Correction

If the pixel-based land/water labels 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. ORBIT framework was used

to correct erroneous labels as well impute missing labels. This is the most crucial step for achieving robust land/water labels.

3.4 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, ReaLSAT-R reports dam construction activity between 1984 and December 2012 even though the surface area dynamics is available from March 1984 till October 2015.

3.5 Manual Verification of Candidate Reservoirs

All candidate reservoirs were manually verified by visual inspection using high resolution satellite imagery. To facilitate the manual verification process, a web interface was developed that displayed all the candidate reservoirs. The interface provided the ability to zoom-in on a high resolution satellite imagery background layer which was used to identify the dam structure or the impoundment wall. 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. Finally, if the annotators were not able to decide whether a candidate is a reservoir or not, it was marked as unknown and hence was excluded from the final set of reservoirs. The interface enabled the user to visualize and tag a candidate reservoir in 30 seconds on an average. The manual verification was done by three co-authors in a span of few days.

Note that the threshold (described in previous subsection) was chosen somewhat arbitrarily. If we choose a strict threshold, it would lead to fewer candidates being rejected by the annotators but would also lead to a smaller number of reservoirs detected. On the other hand, if we choose a less strict threshold, we would be able to get a larger number of true reservoirs but it would take more human work for verification. So, for this version of the database, we chose a threshold such that manual verification does not become too cumbersome.

4 REALSAT-R: HIGHLIGHTS

ReaLSAT-R-2.0 provides location and surface area dynamics of 15473 reservoirs built between 1984 and 2012 globally. Out of 15473 reservoirs reported in ReaLSAT-R, only 853 were also reported in GRanD. Thus, ReaLSAT-R provides information about an additional 14620 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 6 (top) shows the distribution of these reservoirs across different continents while Figure 6 (bottom) shows the cumulative distribution of the number of reservoirs constructed after 1984 in different continents. The majority of dam construction has occurred in Asia and South America since 1984, and the rate of construction in North America has declined significantly.

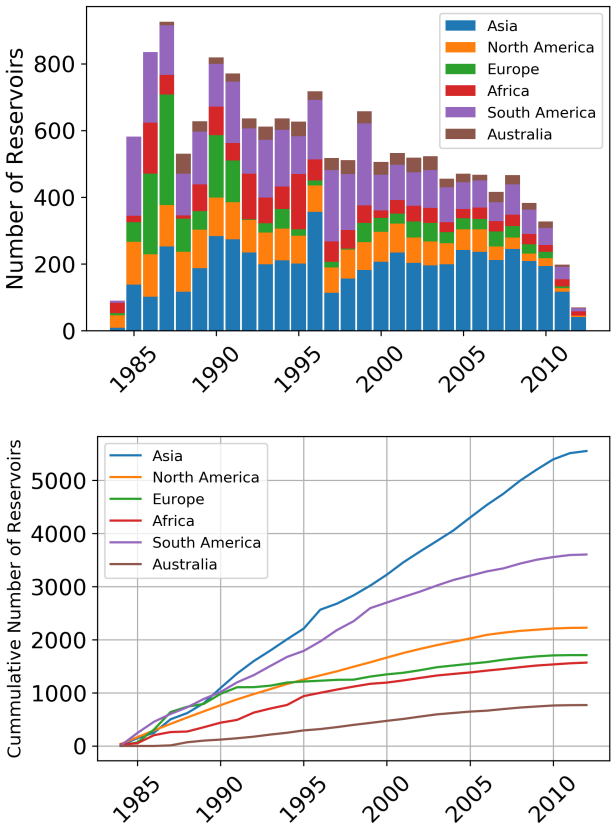

Figure 6: Distribution of reservoirs in ReaLSAT-R-2.0. (top) Year-wise distribution of reservoirs in ReaLSAT-R-2.0 across different continents. (bottom) Time series of cumulative count of reservoirs in ReaLSAT-R-2.0 across different continents.

While GRanD database only provides static information about the extent of reservoirs, ReaLSAT-R also provides surface area at monthly scale from March 1984 to October 2015. Figure 7 shows the aggregate surface area variation of reservoirs in ReaLSAT-R-2.0. Due to the high prevalence of missing data in GSW dataset 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 1984, we processed a subset of 3274 reservoirs that were reported in GRanD and were created before 1984. At global scale, we can see that surface area in reservoirs continued to increase after 2000 as more reservoirs were constructed and approximately

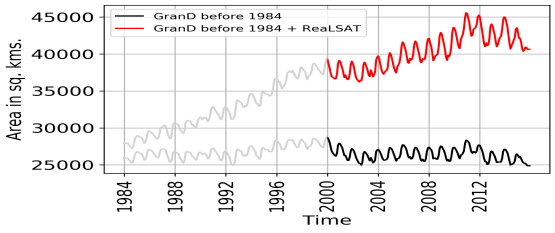

Figure 7: Aggregate surface area dynamics of reservoirs globally. The black line represents aggregate surface area of a subset of reservoirs (3274) reported in GRanD that were built before 1984 with size between 0.1 and 100 sq. kms. The red line represents the aggregate surface area of 3274 old reservoirs and additional 15473 reservoirs created after 1984 that are part of ReaLSAT-R-2.0.

15,000 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.

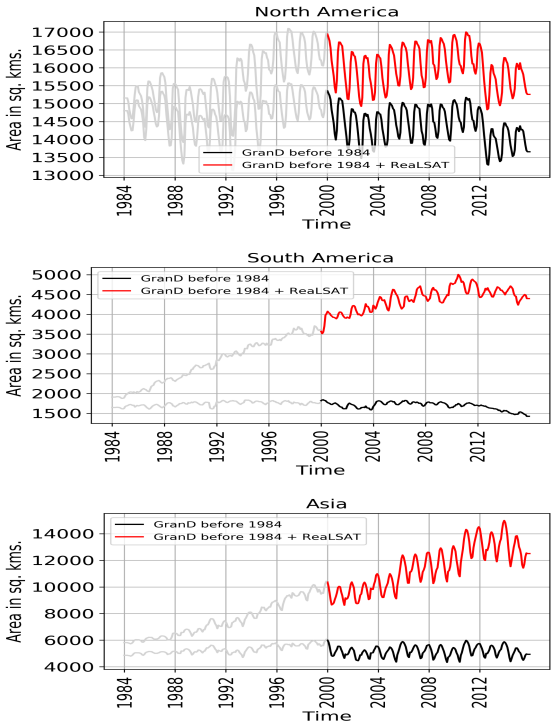

Figure 8: Aggregate surface area dynamics of reservoirs across different continents. The black line represents aggregate surface area of a subset of reservoirs reported in GRanD that were built before 1984 with size between 0.1 and 100 sq. kms. The red line represents the aggregate surface area of the old reservoirs and additional reservoirs created after 1984 that are part of ReaLSAT-R-2.0.

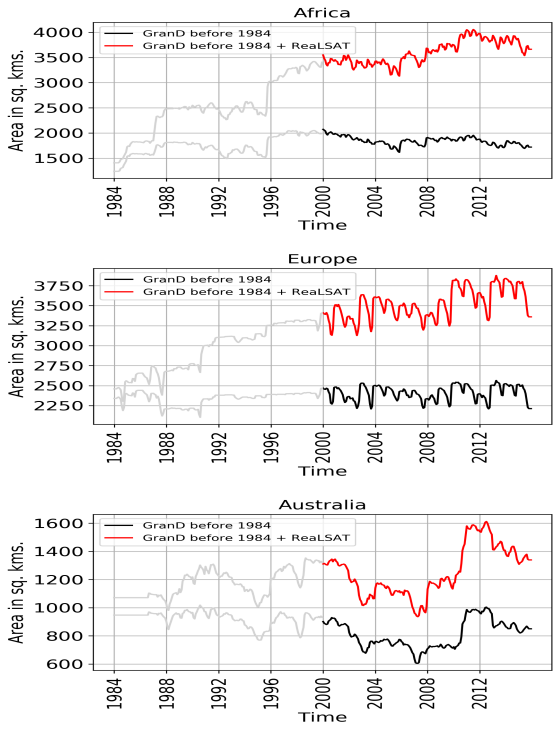

Figure 9: Aggregate surface area dynamics of reservoirs across different continents. The black line represents aggregate surface area of a subset of reservoirs reported in GRanD that were built before 1984 with size between 0.1 and 100 sq. kms. The red line represents the aggregate surface area of the old reservoirs and additional reservoirs created after 1984 that are part of ReaLSAT-R-2.0.

Figure 8 and Figure 9 show 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.

5 DATA AVAILABILITY

ReaLSAT-R provides the location, reference shape, and monthly surface extent maps (which are used to create monthly surface area time series) for each reservoir. Location and time series information of these reservoirs is available at

http://umnlcc.cs.umn.edu/realsat/reservoirs/. This online interface provides locations and reference shapes of all the reservoirs in ReaLSAT-R-2.0. 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 3274

reservoirs reported in GRanD that were built before 1984 with size between 0.1 and 100 sq. kms.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation Awards 1029711 and 1838159. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

REFERENCES

[1] Anuj Karpatne, Ankush Khandelwal, Xi Chen, Varun Mithal, James Faghmous, and Vipin Kumar. 2016. Global monitoring of inland water dynamics: State-of-the-art, challenges, and opportunities. In *Computational Sustainability*. Springer, 121–147.

[2] Ankush Khandelwal. 2019. *ORBIT (Ordering Based Information Transfer): A Physics Guided Machine Learning Framework to Monitor the Dynamics of Water Bodies at a Global Scale.* Ph.D. Dissertation. University of Minnesota.

[3] Ankush Khandelwal, Anuj Karpatne, and Vipin Kumar. 2017. ORBIT: Ordering Based Information Transfer Across Space and Time for Global Surface Water Monitoring. *arXiv preprint arXiv:1711.05799* (2017).

[4] Ankush Khandelwal, Anuj Karpatne, Miriam E Marlier, Jongyoun Kim, Dennis P Lettenmaier, and Vipin Kumar. 2017. An approach for global monitoring of surface water extent variations in reservoirs using MODIS data. *Remote sensing of Environment* 202 (2017), 113–128.

[5] Ankush Khandelwal, Varun Mithal, and Vipin Kumar. 2015. Post classification label refinement using implicit ordering constraint among data instances. In *2015 IEEE International Conference on Data Mining*. IEEE, 799–804.

[6] B Lehner, C Reidy Liermann, C Revenga, C Vorosmarty, B Fekete, P Crouzet, P Doll, M Endejan, K Frenken, J Magome, et al. 2011. Global reservoir and dam database, version 1 (GRanDv1): reservoirs, revision 01. *NASA Socioeconomic Data and Applications Center (SEDAC), Palisades* (2011).

[7] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 7633 (2016), 418.