Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 20-002_rev

ReaLSAT: A new Reservoir and Lake Surface Area Timeseries Dataset created using machine learning and satellite imagery

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

November 4, 2020

ReaLSAT: A new Reservoir and Lake Surface Area Timeseries Dataset created using machine learning and satellite imagery

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

Lakes and reservoirs, as most humans experience and use them, are dynamic three-dimensional bodies of water, with surface levels that rise and fall with seasonal precipitation patterns, long-term changes in climate, and human management decisions. A global dataset that provides the location and dynamics of water bodies 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 (Reservoir and Lake Surface Area Timeseries) 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 dataset provides information for 669107 lakes and reservoirs between 0.1 and 100 square kilometers in size. The visualization of these water bodies and their surface area time series is also available online†. The aim of this paper is to provide an overview of the dataset and a summary of some of the key insights that can be derived from the dataset.

I. Introduction

A lake, as a singular entity, is much more than a conglomeration of square pixels on a map or a static polygon mapped with boundaries fixed on the landscape. Lakes, as most humans experience and use them, are dynamic three-dimensional bodies of water, with surface levels that rise and fall with seasonal precipitation patterns, long-term changes in climate, and human management decisions. Moreover, changes in lakes are tightly coupled with other ecosystem services. For example, in many locations, humans have modified the landscape to control water level in an effort to boost the net benefits of lakes, namely in the areas of irrigation/water supply and power production, often at a cost to natural biodiversity and downstream habitats.

In the last two decades, remote sensing of inland freshwaters, primarily through Landsat imagery, has provided unprecedented opportunity to not only identify the location of the world's lakes but also their sizes and dynamics. Various efforts have begun to provide the necessary underlying data to undertake global hydrologic and biogeochemical modeling. These efforts can be categorized as either vector-based mapping, which provide static polygons of the world's lakes [1, 2], or raster-based mapping which divides the earth's surface into pixels and document the change in the presence of water over time in those pixels [3, 4]. These two approaches provide an incomplete view of the world's lakes, in that the former gives no indication of the dynamic nature of surface water, while the latter does not delineate specific water bodies and also suffers from large amounts of missing data and labelling errors. Given that humans interact with and manage lakes as discrete systems, and decision making is based on the known variance in water availability, a full integration of these two approaches is needed, while at the same time improving their fidelity to produce time-resolved changes of discrete water bodies. By treating lakes as dynamic and discrete entities, patterns of change can be tied to both local freshwater conservation challenges and global biogeochemical cycling.

The Reservoir and Lake Surface Area Timeseries (ReaLSAT) dataset provides an unprecedented reconstruction of surface area variations of lakes at global scale instead of pixels using Earth Observation (EO) data and novel machine

*khand035@umn.edu
†ReaLSAT is available for download at http://umnlcc.cs.umn.edu/realsat/

learning techniques. Specifically, ReaLSAT provides surface area variations of 669107 water bodies south of 50°N (most of the human interaction with fresh water occurs in this geographical range) of sizes between 0.1 and 100 square kilometers. ReaLSAT was created using existing pixel based land/water classification maps available for the period 1984 to 2015 at monthly temporal scale from the GSW dataset [3]. Even though GSW is considered as the state-of-the-art, it suffers from classification errors and missing data due to the challenges in analyzing EO data. Thus, these maps cannot be used directly to extract water bodies and their variations. ReaLSAT uses a novel physics guided machine learning framework that exploit physical properties of lake dynamics to overcome the challenge of data quality while providing information as dynamic lake polygons instead of pixels. The key methodological advancement is based on physical constraints of bathymetry (topography) that regulate growing and shrinking of lakes. For example, within a given water body, a pixel at higher elevation cannot be filled until all the pixels at lower elevation are filled. This heuristic provides a very robust constraint to both correct and impute missing labels in any existing pixel based dataset. The approach falls under the rapidly growing paradigm of theory-guided data science [5]. Next, we present an overview of the ReaLSAT dataset and also present a variety of analyses that are enabled by it.

II. Coverage

Figure 1 (a) shows the size distribution of 669107 water bodies in the dataset. The size distribution follows the power law distribution which is consistent with the previous literature. Here, the size of a water body refers to the area of the pixels labeled as water at least 10 % in the GSW dataset.

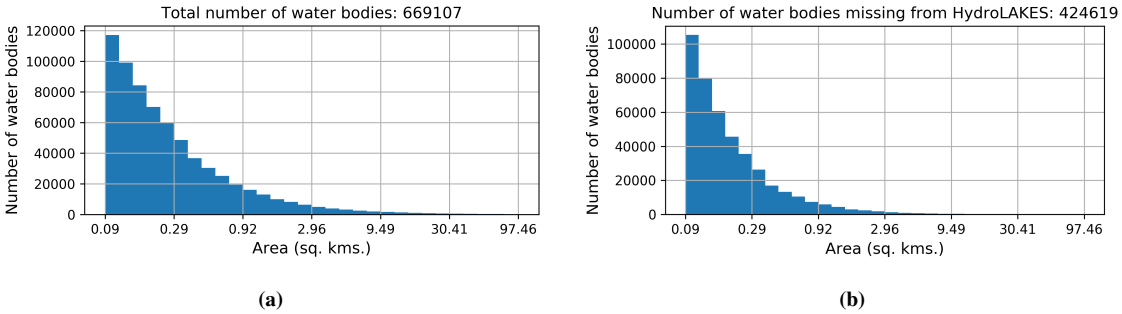
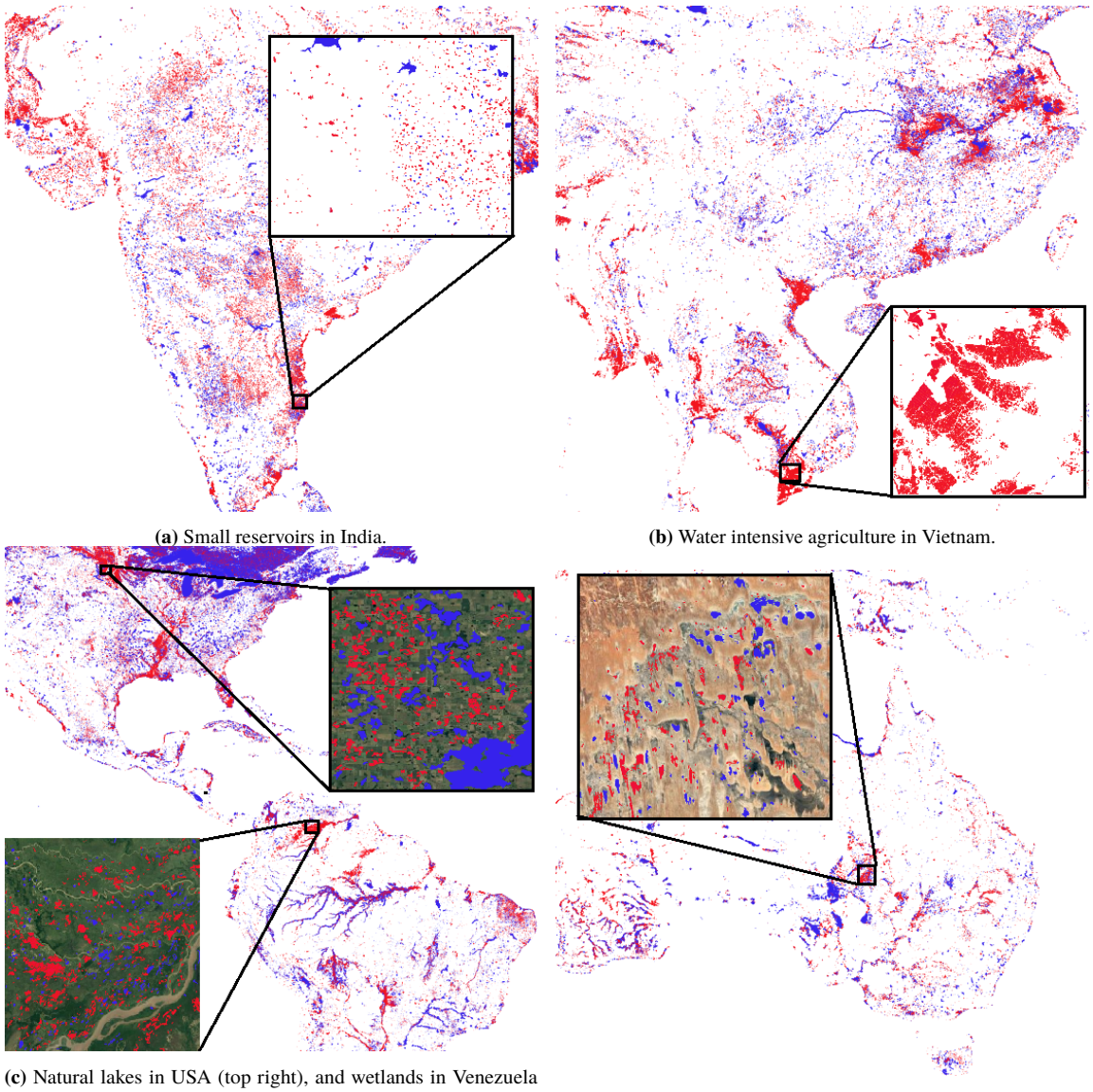


Fig. 1 Size distribution of water bodies in ReaLSAT. (a) all water bodies. (b) only water bodies that are missing from HydroLAKES dataset.

In terms of spatial coverage, ReaLSAT contains water bodies that exists below 50 degree North and have surface area between 0.1 and 100 square kilometers. Since the dataset was created using satellite imagery analysis, it has the potential to provide much better coverage compared to existing datasets. For example, we found that ReaLSAT contains 424619 water bodies that are missing from HydroLAKES dataset [2] (a widely used dataset by the hydrology community). Figure 1 (b) shows the size distribution of the water bodies that are missing from HydroLAKES dataset. The size distribution follows a power law very which is very similar to the distribution of the complete dataset which suggests that water bodies of all sizes are missing from HydroLAKES. Figure 2 shows a selection of some of the spatial clusters of the water bodies that are missing from HydroLAKES. As illustrated through these examples, there are different types of water bodies that are missing from HydroLAKES.

III. Dynamics

The above analysis highlights the utility of ReaLSAT in providing additional water bodies that are missing from existing datasets. Apart from the location of these water bodies, ReaLSAT also provides monthly scale surface area variations of these individual water bodies. Since GSW dataset contains a large amount of missing labels as well as incorrect labels, it can not be used directly to create robust monthly surface area variations of water bodies. In order to overcome this challenge, a novel machine learning framework, ORBIT (Ordering Based Information Transfer) [6–9] was used to create the ReaLSAT dataset. This framework (described in more detail in Section V) 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 correct physically inconsistent labels, and impute missing labels. For example, Figure 3 illustrates the utility of the ORBIT framework. Figure 3 (a) show the fraction map of a reservoir in Brazil (latitude:



(a) Small reservoirs in India.

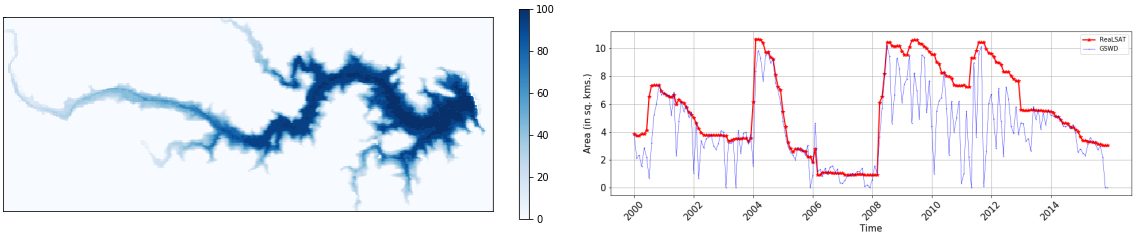
(b) Water intensive agriculture in Vietnam.

(c) Natural lakes in USA (top right), and wetlands in Venezuela (bottom left).

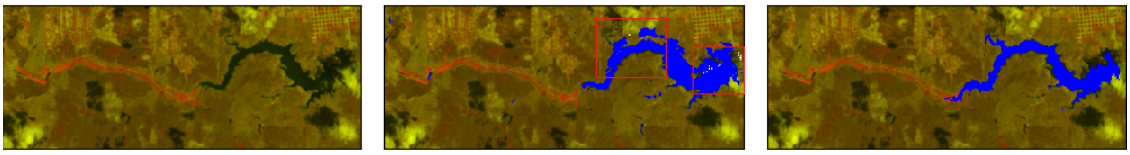
(d) Shallow lakes in Australia.

Fig. 2 A selection of spatial clusters of missing water bodies in HydroLAKES. Water bodies missing from HydroLAKES are shown in red, and water bodies present in HydroLAKES as well as ReaLSAT are shown in blue.

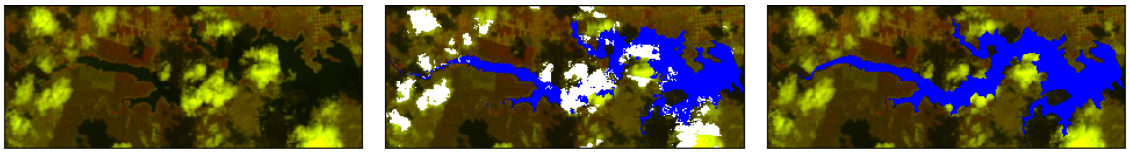
-5.633609, longitude: -35.655042). The fraction map represents the percentage of timesteps each pixels is labelled as water in the ReaLSAT dataset. For the ease of visualization, Figure 3 (b) shows the surface area variations from 2000 onwards. The red timeseries represents area according to ReaLSAT, whereas the blue timeseries represents the surface area variations according to the GSW dataset. As we can see from the two timeseries, ReaLSAT is able to provide a much more smooth and consistent timeseries compared to using GSW labels directly, while capturing surface area variations that are not apparent from the GSW based timeseries. For example, ReaLSAT shows a continuous decrease in the area between years 2011 and 2012 which is not visible in the GSW timeseries. This highlights the ability of ORBIT framework to use robust physical constraints among the pixels to impute missing labels and correct physically inconsistent labels. Figure 3 (c) and (d) show two snapshots to provide visual validation of the corrections made by the ORBIT framework. In Figure 3 (c), regions where ORBIT is able to correct erroneous labels are marked by two red boxes. Similarly, in Figure 3 (d), ORBIT is able to impute missing labels by leveraging labels of other locations. Since a lot of shallow pixels are labelled as water, the ORBIT framework imputes the missing labels in the middle parts of the lake as water.



(a) Fraction Map **(b)** Surface area timeseries



(c) Visual comparison for Nov 11, 2015. left: False Color Composite, middle: GSW labels, right: ReaLSAT labels. Blue represents water, and missing labels are shown in white



(d) Visual comparison for Dec 27, 2010. left: False Color Composite, middle: GSW labels, right: ReaLSAT labels. Blue represents water, and missing labels are shown in white

Fig. 3 An illustrative example to demonstrate the utility of the ORBIT framework using a reservoir in Brazil (latitude: -5.633609, longitude: -35.655042). (c) Highlights the ability of ORBIT framework to correct physically inconsistent labels. (d) Highlights the ability of ORBIT framework to impute missing labels.

Since, ORBIT framework uses bathymetric constrains to correct and impute labels, it will not be applicable for water bodies that violate this assumption. Hence, we provide an additional tag (see section VI.E for details) to distinguish water bodies for which monthly dynamics can be used to perform temporal analysis.

Figure 4 shows the aggregate surface area timeseries of 485861 water bodies (out of 669107) used for analyzing surface area dynamics. Note that the timeseries before the year 2000 has been shown in gray to signify that before 2000, the GSW dataset contains a lot of missing values in major parts of the world and hence, the timeseries signal should be interpreted with caution. Overall, years 2011 and 2012 appears to be the wettest years and 2009 and 2014 are one of the driest years globally.

Figure 5 shows the total area variations for each continent separately. As we can see, all continents have very different area variations over the study period. South America has shown a very long drought pattern from 2002 till 2010, whereas Africa has shown gradual increase in the total surface area during this period (partly due to addition of new reservoirs). Area variations in North America were consistent except for the year 2012. Australia has shown a major increase in surface for the period 2010-2012.

IV. Water Body Type Identification

The above section provides a global overview of surface area changes. In this section, we provide an overview of how surface area variations of individual water bodies can be analyzed to potentially categorize them into different types. For example, Figure 6 show an illustrative example of four different types of water bodies, namely, reservoirs, water intensive agriculture (e.g. rice farms), transient lakes, and mining lakes. As evident by their surface area variations, these water bodies show very different type of temporal pattern. In particular, when new reservoirs are constructed, they show a sudden increase in the area after they become operational, and this sudden increase tends to persist over time. Water intensive farms such as rice farms show very high variability in every season, where the area reaches very high values and return to zero every year. Transient lakes are mostly climate driven and appear in years with high precipitation and disappear in the following years. Mining lakes are characterized by gradual increase in the area as the mining activity progresses over time. Hence, different types of characteristic functions (scoring functions) can be

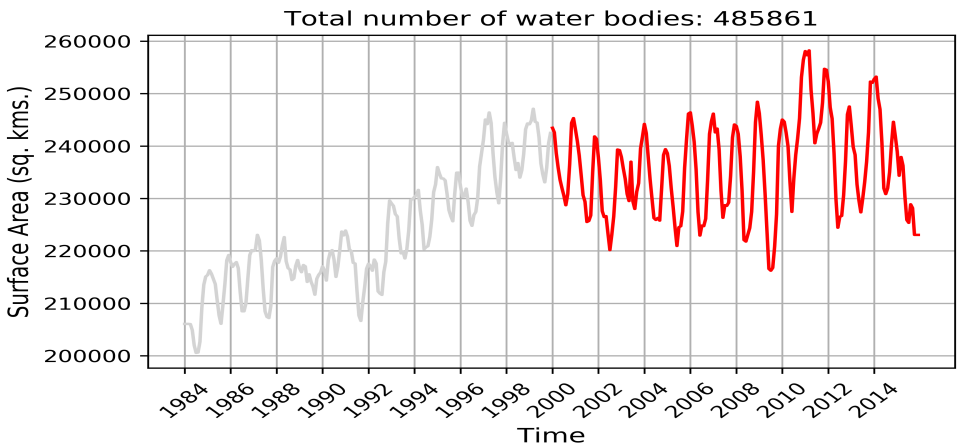


Fig. 4 Aggregate surface area dynamics of 485861 water bodies in ReaLSAT.

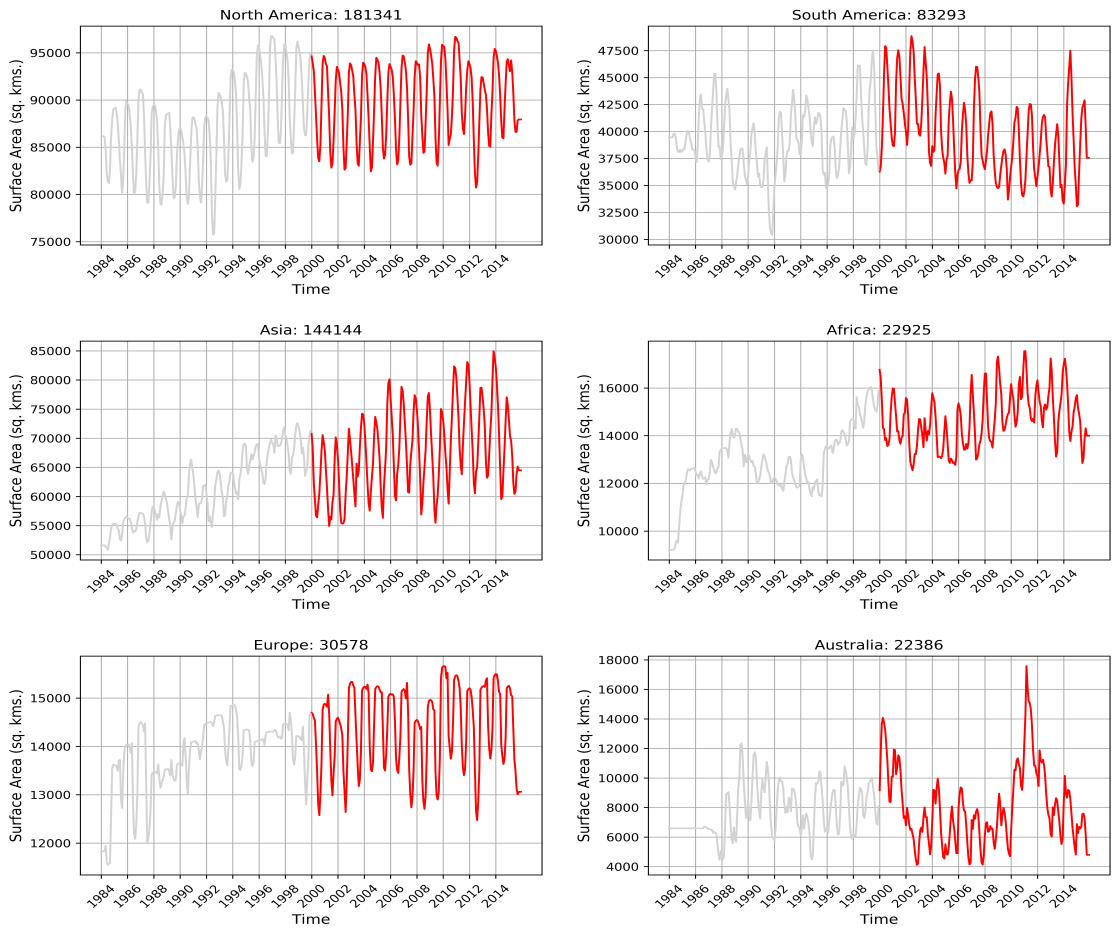


Fig. 5 Aggregate surface area dynamics of water bodies across different continents.

developed to capture the above observations and enable automated detection of different types of water bodies.

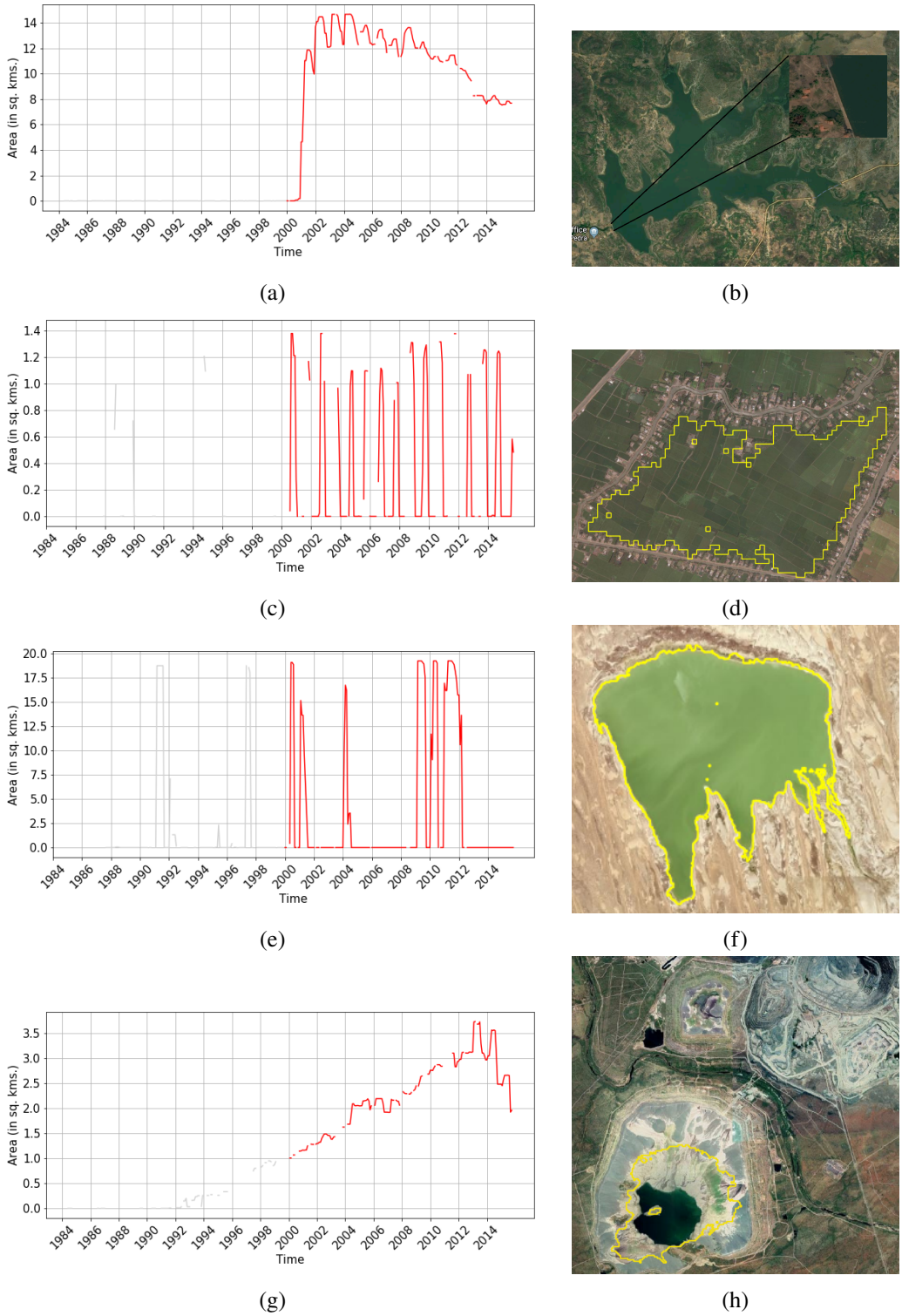


Fig. 6 An illustrative example of surface area dynamics of different types of water bodies. (a-b) Reservoir, (c-d) Farm, (e-f) Transient lake. (g-h) Mining lake. Timeseries charts do not show timesteps with more than 90% missing data.

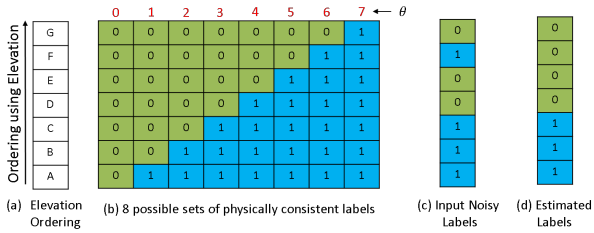


Fig. 7 An illustrative example showing elevation ordering based label correction process

V. ORBIT (Ordering Based Information Transfer) Framework

In this section, we provide an overview of the ORBIT framework used to create ReaLSAT dataset. The 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 due to topographic constraints all the locations in the water body 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 7 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 7 (b) shows 8 possible sets of physically consistent labels for 7 locations shown in Figure 7 (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 7 (c) shows the erroneous input labels and 7 (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 [6] 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 [6], 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 [9] uses the elevation ordering constraint 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.

VI. 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.

A. 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. The GSW dataset product 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 the new dataset as well.

B. 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 8 illustrates the database generation process on a small region in USA. Figure 8 (a) 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. Figure 8 (b) shows the binary mask created by thresholding the fraction image. Finally, this binary mask is used to extract individual connected components (sets of contiguous pixels) as shown in Figure 8 (c). In this image, each connected component is shown in a different color.

(a) **(b)** **(c)**

Fig. 8 An illustrative example of the water body construction process.

Using these reference shapes, we extract pixel-based land/water label maps at monthly scale for each lake individually. A buffer region is added around the reference shape of a water body to capture its complete dynamics. To avoid including other nearby lakes in the buffer, we further prune the buffer region using an automated approach as described in [7].

C. River Segment Identification

Since, GSW is a pixel based dataset, water bodies extracted in the previous step also contains river segments. Since, rivers do not follow the topographic constraints, the ORBIT framework is not applicable for river segments. Hence, we remove river segments from the initial candidate set of water bodies. To identify river segments, we used their geometric characteristics. In particular, we calculate how many erosion operations (referred to as e) would be required to completely erode a water body. For water bodies that are narrow and long, it would take a very few erosion operations to erode a water body even if the total number of pixels are large. On the other hand, if we assume a water body to be a perfect square, then it would take $\frac{\sqrt{N}}{2}$ erosion operations to completely erode a water body with N pixels. Thus, a morphological score, $\frac{\sqrt{N}}{2e}$, can be used to rank water bodies, where higher values would correspond to rivers. We ranked the water bodies based on this score and visually inspected water bodies to identify a suitable threshold of 5. Note that, there could still be some river segments that are part of ReaLSAT because no threshold would completely eliminate rivers without excluding too many lakes and reservoirs.

D. 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.

E. Identification of Water Bodies with Reliable Surface Area Dynamics

The ORBIT framework makes an assumption that a water body exists as a single bowl (or a set of bowls that are always connected). This assumption enables the use of topographic constraints to correct and impute missing labels. However, not all water bodies follow this assumption, and hence for such water bodies, surface area variations provided by ReaLSAT should be used with caution. To identify these water bodies, we used four different scoring functions -

- *Connected Component Score (CS)*: The aim of this score is to detect if a water body exists as a set of disjointed components which will indicate the violation of the topographic constraint. In particular, for any given water body, we calculated the number of pixels that do not belong to the two biggest components at any given timestep as a penalty for that timestep. To calculate a single value for a water body, we aggregate the penalties as following:

$$CS = \frac{\sum_{t=1}^{T} P[t]}{\sum_{t=1}^{T} W[t]} \tag{1}$$

  where, $P[t]$ is the penalty at timestep t and $W[t]$ is the total number of water pixels at timestep t. An ideal water body will have a CS score of 0.
- *Disagreement Score (DS)*: While ORBIT has the ability to correct the physical inconsistencies in the input labels from the GSW dataset, its effectiveness will get impacted if input labels have too many errors. Hence, for each water body we calculated the disagreement between the fraction map created using GSW dataset and the fraction map created after ORBIT based correction. A fraction map is basically represents fraction of timesteps a pixel is labelled as water. Water bodies that show high disagreement might have significant data issues (such as algae, very shallow topography), and hence their area variations should be used with caution. Specifically, we calculated the Disagreement Score as follows -

$$DS = \frac{1}{N} \sum_{i} \sum_{j} \frac{|F_{GSW}^{i,j} - F_{ReaLSAT}^{i,j}|}{max(F_{GSW}^{i,j}, F_{ReaLSAT}^{i,j})} \tag{2}$$

  Where F_{GSW} is the fraction map created using input labels from GSW dataset, $F_{ReaLSAT}$ is the fraction map created using labels from ReaLSAT dataset, i and j are the pixel locations of the fraction map, and N is the total number of dynamic pixels in the water body. The score penalizes errors more on the boundary of the water body (low value of the denominator) compared to the center of the water body (generally have larger fraction values).
- *Transient Score (TS)*: Shallow lakes that appear for very small amount of time do not always have a well defined topographic structure, and thus could lead to incorrect changes by ORBIT framework. To capture the transient nature of water bodies, we used a simple score that counts number of timesteps a water body was completely dry.
- *Static Score (SS)*: While transient water bodies exists for very small time periods, a large number of water bodies that have some portion that always remains water. These water bodies tend to have much better topographic structure and thus ORBIT based corrections are more effective for these water bodies. To calculate this aspect, we simply use the ratio of pixels that are always water over pixels that are water at least once in the study period. A high score would represent a very static water body and a low score would represent a highly dynamic water body.

We visually inspected water bodies by ranking them according to these four different scores and identified the appropriate thresholds to tag water bodies for which ORBIT corrections would be effective. Specifically, we used the following set of rules to tag such water bodies

- $SS = 0$ & $TS > 0.5$ & $CS < 0.1$ & $DS < 0.2$: This rule selects water bodies that are highly dynamic (SS=0), very transient (TS>0.5), but have very low connected component based penalty (CS<0.1) and also have low disagreement between input fraction map and fraction map after ORBIT correction (DS<0.2).
- $SS = 0$ & $TS <= 0.5$ & $CS < 0.15$ & $DS < 0.3$: This rule selects water bodies that are highly dynamic (SS=0), are relatively less transient (TS<=0.5). The thresholds for CS and DS are relaxed because these water bodies are less transient.
- $SS > 0$ & $CS < 0.2$ & $DS < 0.2$: This rule selects water bodies that are relatively stable (have some portion that is always water i.e. SS>0), The thresholds for CS and DS are even more relaxed because these water bodies not only less transient but also have portions that are always water.

- $SS >= 0.5$ & $CS < 0.2$: This rule selects water bodies that are highly stable (SS>=0.5).

Out of 669107 total water bodies, 485861 of them satisfy one of the above rules. While ReaLSAT provides area variations of all 669107 water bodies, only the selected water bodies were used to analyze the surface dynamics in section III.

VII. Data Availability

ReaLSAT provides the location, reference shape, and monthly surface extent maps (which are used to create monthly surface area timeseries) for 669107 water bodies. the location and time series information of these reservoirs is available at http://umnlcc.cs.umn.edu/realsat/.

References

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

[2] Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O., "Estimating the volume and age of water stored in global lakes using a geo-statistical approach," *Nature communications*, Vol. 7, No. 1, 2016, pp. 1–11.

[3] Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S., "High-resolution mapping of global surface water and its long-term changes," *Nature*, Vol. 540, No. 7633, 2016, p. 418.

[4] Donchyts, G., Baart, F., Winsemius, H., Gorelick, N., Kwadijk, J., and Van De Giesen, N., "Earth's surface water change over the past 30 years," *Nature Climate Change*, Vol. 6, No. 9, 2016, pp. 810–813.

[5] Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V., "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on knowledge and data engineering*, Vol. 29, No. 10, 2017, pp. 2318–2331.

[6] Khandelwal, A., Mithal, V., and Kumar, V., "Post classification label refinement using implicit ordering constraint among data instances," *2015 IEEE International Conference on Data Mining*, IEEE, 2015, pp. 799–804.

[7] Khandelwal, A., Karpatne, A., Marlier, M. E., Kim, J., Lettenmaier, D. P., and Kumar, V., "An approach for global monitoring of surface water extent variations in reservoirs using MODIS data," *Remote sensing of Environment*, Vol. 202, 2017, pp. 113–128.

[8] Khandelwal, A., Karpatne, A., and Kumar, V., "ORBIT: Ordering Based Information Transfer Across Space and Time for Global Surface Water Monitoring," *arXiv preprint arXiv:1711.05799*, 2017.

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