Examining the Spatiotemporal Dynamics of Lake Oroville's Area with NDWI Indices from Landsat 8 Image Data

James Lu
Halicioğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093
jqlu@ucsd.edu
Samuel Aguire
Halicioğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093
saguirre@ucsd.edu

Abstract

Surface water has an important relationship with both the climate and human well being, thus it is important for us to monitor and measure bodies of water as the climate continues to change. Prior approaches in this field have had a larger scope and monitored changes throughout large bodies of water throughout the globe, however they lack more granular observations of smaller and more specific bodies of water in the United States. We attempt to close this gap by analyzing other important bodies of water throughout the United States, and specifically we analyzed Lake Oroville in Northern California. We used images captured with varying wavelengths on the electromagnetic spectrum to identify differences between water and non-water areas, and to analyze how these bodies of water change over time.

1 Introduction

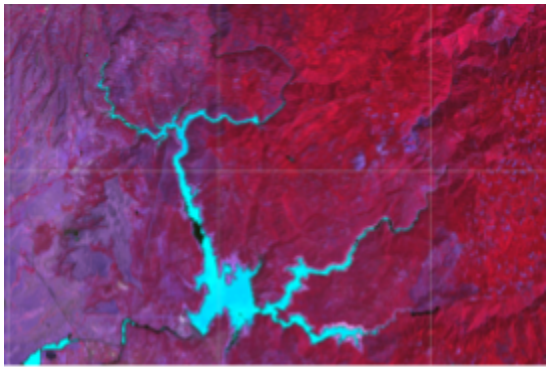


Figure 1: Mosaic image of Lake Oroville

1.1 Intro

Quality of life is said to be captured by an individual's state of being. However, this notion is challenged by the limitation and availability of resources, such as water. In fact, there has been a growing discussion on the topic of water restriction in recent years. More specifically, in the state of California the concern of water shortages, alongside a growing fear of water limitations, has been imposed on its residents. Of course, there are a number of variables that contribute to the state's ongoing issue. For instance, California boasts a substantial agriculture industry, one that demands a substantial amount of water. Nevertheless, in the past couple of decades these concerns have only seemed to elevate as California's water reserves continue to deplete. An essential resource required for the sustainability of the state's immense agricultural industry, and its dense population.

Indeed, there has been a dramatic rise in the development of data driven solutions in an effort to mitigate such environmental distress. Across many domains, a myriad of scientists and researchers strive to implement solutions that quantify and document the long-term changes of surface water at a global scale. Therefore, the objective of this statistical analysis report is to not only supplement existing literature, but to provide a unique perspective on a localized area of interest. More precisely, we examined the spatial and temporal changes of Lake Oroville, a reservoir in Northern California, using 80 Landsat 8 atmospherically corrected surface reflectance images.

1.2 Literature Review

Our efforts draw great influence from the work of Pekel et al. [4], a comprehensive statistical analysis that quantifies and measures the long-term variability of global surface water. The work of these authors is significant because it provides this variability in surface water at a high-resolution and more importantly, at an incredibly large scale. In fact, Pekel et al. [4] provide this analysis at a 30 meter spatial resolution over a 32 year span using an immense amount of satellite imagery data. More specifically, the authors use the entirety of Landsat 5 Thematic Mapper, the Landsat 7 Enhanced Thematic Mapper-plus and the Landsat 8 Operational Land Imager orthorectified, top-of-atmosphere reflectance and brightness temperature images (Pekel et al., 2016). Given the volume of data being used, solutions in the cloud computing space were utilized for processing such amounts of data.

Additionally, not only does our work exclusively examine a localized area of interest, Pekel et al. [4] also utilize a combination of sophisticated techniques for such a large scale analysis. For example, in the task of water detection, a task that is very much non-trivial at the given scale, the authors use techniques such as expert systems, visual analytics, and evidential reasoning. Methods which they describe as being less commonly used within the field of remote sensing (Pekel et al., 2016). These techniques displayed great practicality in their work, but also demonstrated to be useful for the task of water detection in our work as well. For example, rudimentary forms of evidential reasoning and visual analytics were employed in our work specifically to determine what images were suitable for analysis.

Another paper by Özelkan [3] provided us with a few methodologies and ideas to incorporate into our own analysis. The work of Özelkan [3] centers around the utilization of NDWI indices derived from Landsat-8 OLI multispectral satellite images. However, they develop six different NDWI models for the task of water detection. By using green, near infrared, shortwave infrared 1, and shortwave infrared 2 wavelengths, at both 15m and 30m spatial resolution. Özelkan [3] aimed to demonstrate which of these six NDWI models produced the best results and accuracy of Atikhisar Dam (Özelkan, 2019).

Finally, our work shares a few parallels to that of Hui et al. [1], in which they attempt to model the spatial and temporal changes of a localized area of interest, Poyang Lake (Hui et al., 2008). Although our work does not directly follow in the footsteps of Hui et al. [1], examining their literature helped us form a conceptual framework for the given problem.

In any case, our work attempts to supplement existing literature of such authors by conducting an statistical analysis of a localized area of interest, Lake Oroville.

1.3 Data Description

We use satellite images taken from the USGS Landsat 8 Level 2, Collection 2, Tier 1 surface reflectance dataset. There are 9 spectral bands available for us to use captured every 2 weeks with a 30 meter resolution. The dataset comprises images that were captured from March 2013 till the present day, enabling us to analyze the alterations in the surface water of Lake Oroville during the preceding decade. With the various spectral bands captured by the satellite we can distinguish between surface water and non surface water areas at a 30 m resolution.

In order to compare and quantify how accurate our NDWI model's estimated surface water was, satellite imagery was collected from JRC Monthly Water History, v1.4 from the Google Earth Engine Catalog. This dataset was developed by Pekel el al. [4] and demonstrated to be the most reliable way to compare our results. Due to the fact that their findings, research, and methodologies are well respected in the satellite imagery domain. Therefore, we reasoned that it would provide for validation data.

2 Methods

2.1 ETL

Google Earth Engine was our primary method for data collection, as it provides a flexible and robust API to extract, transform, and load satellite image data. This process was initiated by using built in functions to select and filter Landsat 8 images based on a desired time frame. We selected the earliest date USGS Landsat 8 Level 2, Collection 2, Tier 1 allowed, which was March 18, 2023. Our end date was selected in the same fashion, providing at the time was October 24, 2022. It was reasoned that selecting the complete timeframe would be fitting to acquire a precise comprehension of the spatiotemporal dynamics of Lake Oroville's surface area.

The presence of cloud disturbance posed a significant obstacle during the initial stages of our analysis, requiring us to mitigate the potential issues by minimizing cloud noise to accurately classify water and non-water features within a given image. Therefore, we attempted to filter the data based on the percentage of cloud covering in an image. This was initially done by experimenting with a cloud cover percentage parameter, however it was ultimately decided that using a threshold of 0.5 would produce clean enough data to analyze. As lowering or increasing this threshold did not improve the quality of images with respect to cloud disturbance by a significant amount.

Subsequently, it was determined that the optimal approach to address the challenges arising from cloud disturbance was to selectively eliminate such images from our analysis. After the filtering of satellite data was performed we then proceeded to generate false color images by computing the Normalized Difference Water Index (NDWI). As this would then allow us to classify between water and non-water surfaces in a given image.

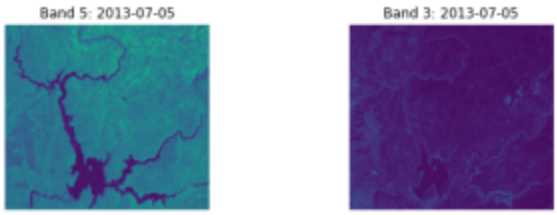


Figure 2: Visualization of Lake Oroville using bands B3 and B5

$$NDWI_{(B3,\,B5)} = \frac{B3 - B5}{B3 + B5}$$

$$NDWI_{(B3,\,B6)} = \frac{B3 - B6}{B3 + B6}$$

$$NDWI_{(B3,B7)} = \frac{B3 - B7}{B3 + B7}$$

This was done by selecting visible green (SR_B3), near infra-red (SR_B5), shortwave infrared 1 (SR_B6) and shortwave infrared 2 (SR_B7) bands. Then using Google's API to compute the normalized difference among all images. These three combinations of bands would serve as our models for our estimated surface area. As mentioned previously, this idea was inspired by Özelkan [3], where they used the same types of bands, but produced six different models by using 15 meter and 30 meter spatial resolutions. In our case we produced three different NDWI models at a 30 meter resolution.

The images were then exported and downloaded locally using geemap, a third party library for interactive mapping with Google Earth Engine (Wu, 2020). The dataset developed by Pekel et. al [4] was also extracted and loaded from Google Earth Engine in a similar fashion.

2.2 EDA

These bands were also collected as individual images for exploratory analysis purposes.

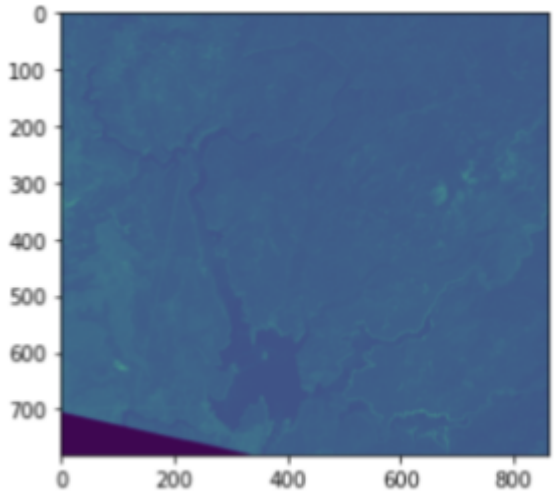


Figure 3: Missing data for band B3, date 2013-06-03

While performing our exploratory analysis we found that the first image of our dataset contained missing pixel values for an image on the date 2013-06-03. This was clearly made evident when plotting the distribution of pixel values of the B3 band. These missing pixel values imposed a problem when attempting to calculate the area of the lake, as they were being included in the calculations. Due to the fact that when computing the NDWI of the given image, the missing pixels had the same range of values for water surfaces. Therefore, we did not consider such images in our calculations.

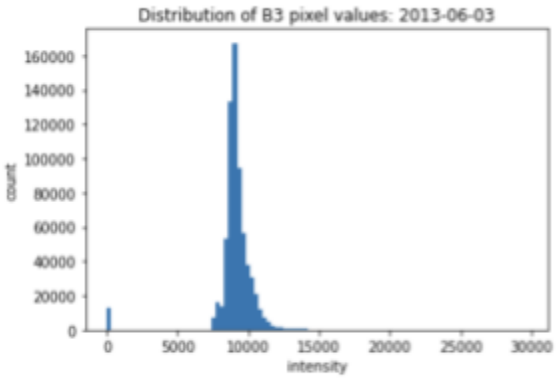


Figure 4: We can see that there are pixel values with 0

We also explored the distribution of pixel values for bands SR_S3, SR_S5, SR_B6, and SR_B7. This allowed us to conceptualize and form approaches to segment water features in the next section. As well as visually understand why a certain model performs better than the others. Notably, when taking a look at bands 6 and 7 from figure 5, they have the same distribution in some of the satellite images.

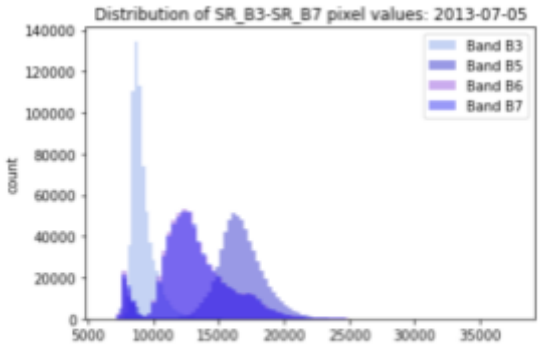


Figure 5: We can see the distribution of pixel values for bands 3-7 on the given date

When looking at the distribution for bands SR_B3 and SR_B5 in figure 5 we see a clear separation between the two histograms. We then proceeded to take a closer look at the average values among all images for these two bands. We observed some slight differences, however it is important to note that the distributions from figure 6 still contain images with cloud noise. Such images contain very different distributions to those without cloud disturbance. This is clearly evident when examining figure 7, which shows that the two histograms do not separate well enough to even attempt to classify the two features (water and non-water) for the given problem. Therefore, we began leaning towards histogram-based thresholding in order to classify water and non-water features and thus reinforced the idea of excluding images with such noise.

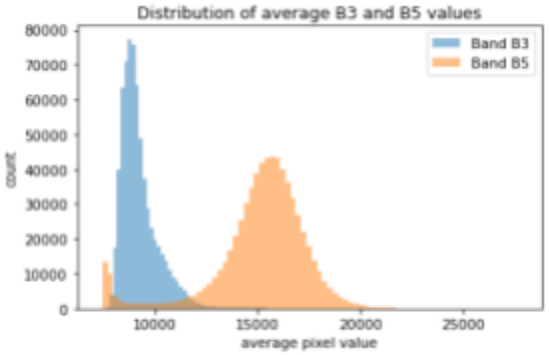


Figure 6: Distribution of average B3 and B5 values using all images in dataset.

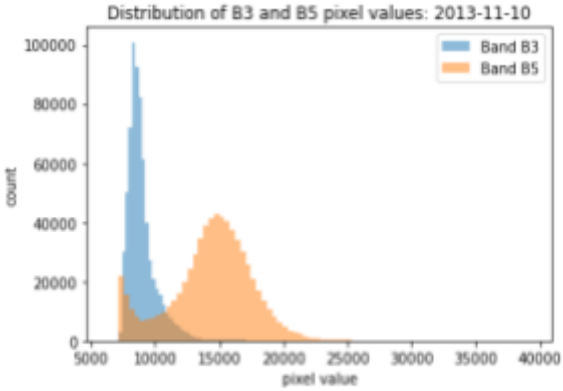


Figure 7: Distribution of a cloudy image, very different in comparison to non-cloudy images.

2.3 Water Detection

In order to quantify the spatiotemporal changes of Lake Oroville, we attempted a few methods to completely isolate water and non-water surfaces. Thus, the first attempt was to use an unsupervised learning technique such as k-means clustering in order to classify our images. However, this approach did not produce results that could be used for accurate results. As there was still much noise in certain images.

So the approach we took was to use a thresholding algorithm offered by Sci-kit Imaging. In our implementation specifically, we utilized minimum thresholding in order to determine the thresholds for water and

non-water from our NDWI images. This method computes the histogram for an image and smoothes it until there are only two maximas and sets the threshold as the minima between the two. Thus, we utilized this method for each image individually. This method created a clear separation between the water and non-water areas, however for some unused images there were noise and other factors from the original images that created issues and inconsistencies after processing and thresholding.

With our valid thresholded images we estimated the surface area of Lake Oroville by counting the pixels classified as water in our processed images. In order to assess the accuracy of our model, we compared our values to a validation dataset created by Pekel et. al. [4] which also classifies water and non-water areas. We also distinguished permanent water from seasonal / ephemeral water in Lake Oroville by comparing all of the images pixel by pixel. If a pixel appears in more than 70% of images then it can be considered permanent water and any pixel under this threshold is considered ephemeral water.

3 Results

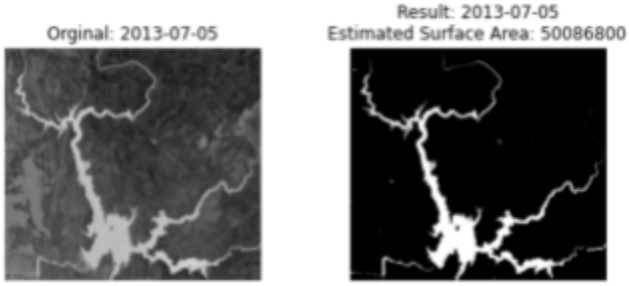


Figure 8: Example of a successful thresholded image

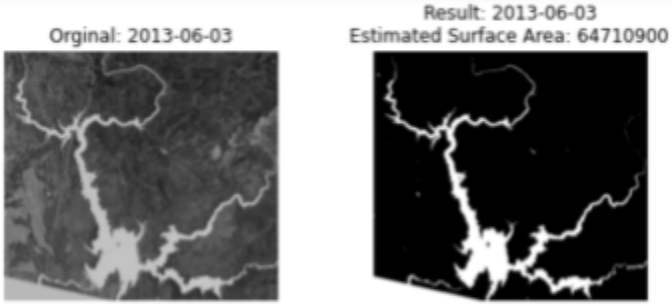


Figure 9: Example of thresholded image with missing data

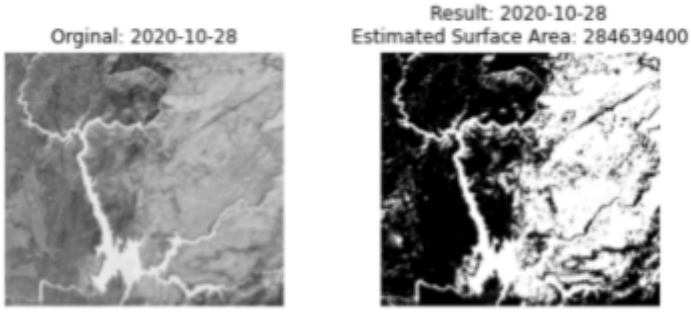
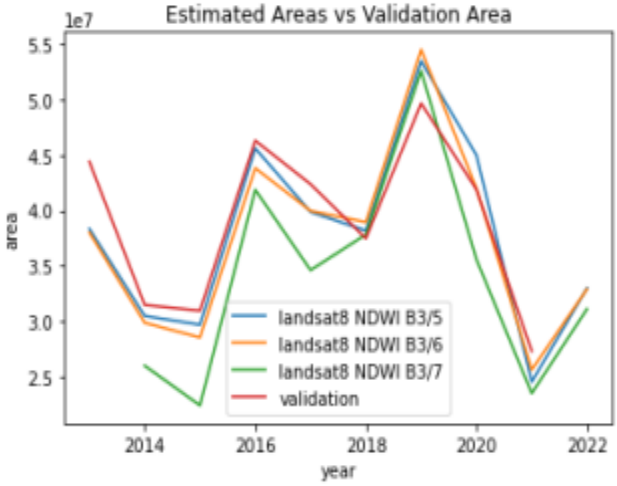
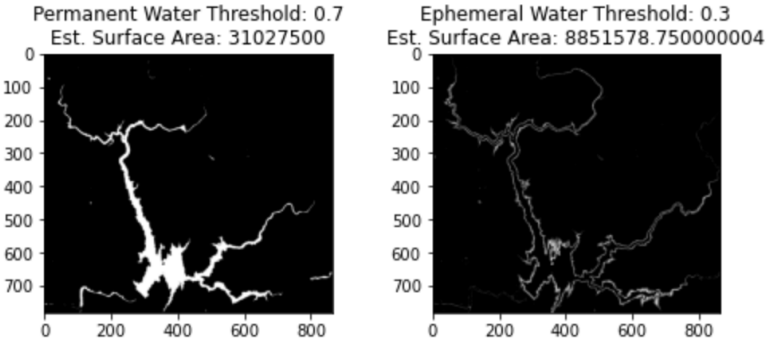


Figure 10: Example of thresholded image with noise





	year	b3_b7_est_surface_area_sq_ft	b3_b5_est_surface_area_sq_ft	b3_b6_est_surface_area_sq_ft	validation_surface_area_sq_ft
0	2014	25961400	30456800	29835900	31446500
1	2015	22370400	29661900	28510500	30932600
2	2016	41852250	45592700	43806300	46293800
3	2017	34592100	39852338	39948043	42361100
4	2018	37810200	38164838	38937729	37455600
5	2019	52591950	53481375	54515300	49664400
6	2020	35561700	44975443	41972914	41877450
7	2021	23462486	24501343	25591275	27277875

Table 1: Displays estimated areas for our three models

4 Discussion

As seen in our results section, our current model produces results that fall into one of three categories: successful thresholding, failure due to missing data, or failure due to noise. Most of our data can be successfully thresholded with our current methods, however there are some processed NDWI images that were generated with missing data thus implying water where there should not be. Another failed scenario occurs when the images captured by LANDSAT are muddied by clouds or other noise that cause specific bands to be nearly unusable. We also found a general upward trend of water surface area until about 2021, where there was a sharp drop-off. This can be observed in Table n and Figure n, however we do not have much confidence in the validity of these results as the amount of image data varies from year to year, which would ultimately affect our surface area estimates.

Our results are similar to other granular analysis of small bodies of water up until 2020, as an analysis of inland water in Sri Lanka [2] found a similar upwards trend in water surface area. However, our results require more image data to increase the confidence level of our predictions. Our approach is limited upon the availability of image data for Lake Oroville, and the time limitation for this quarter. Our approach can be improved upon by 1. obtaining more image data and 2. preparing the images through a pre-processing pipeline in order to mitigate noise and other issues as much as possible. More in depth analysis on the channels of water from Lake Oroville as well as the intra-annual changes and patterns would also be an interesting topic to further explore.

5 References

[1] Fengming Hui, Bing Xu, Huabing Huang, Qian Yu & Peng Gong (2008) Modelling spatial-temporal change of Poyang Lake using multitemporal Landsat imagery, International Journal of Remote Sensing, 29:20, 5767-5784, DOI: 10.1080/01431160802060912

[2] Li, J., Wang, J., Yang, L. et al. Spatiotemporal change analysis of long time series inland water in Sri Lanka based on remote sensing cloud computing. Sci Rep 12, 766 (2022). https://doi.org/10.1038/s41598-021-04754-y

[3] Özelkan, E. (2020). Water Body Detection Analysis Using NDWI Indices Derived from Landsat-8 OLI. Polish Journal of Environmental Studies, 29(2), 1759-1769. https://doi.org/10.15244/pjoes/110447

[4] Pekel, JF., Cottam, A., Gorelick, N. et al. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422 (2016). https://doi.org/10.1038/nature20584

[5] Wu, Q., (2020). geemap: A Python package for interactive mapping with Google Earth Engine. The Journal of Open Source Software, 5(51), 2305. https://doi.org/10.21105/joss.02305