data

MDPI

Data Descriptor

A Six-Year, Spatiotemporally Comprehensive Dataset and Data Retrieval Tool for Analyzing Chlorophyll-a, Turbidity, and Temperature in Utah Lake Using Sentinel and MODIS Imagery

Kaylee B. Tanner, Anna C. Cardall and Gustavious P. Williams *

Department of Civil and Construction Engineering, Brigham Young University, Provo, UT 84602, USA; kbt36@byu.edu (K.B.T.); cardalla@student.byu.edu (A.C.C.)
* Correspondence: gus.williams@byu.edu

Abstract

Data from earth observation satellites provide unique and valuable information about water quality conditions in freshwater lakes but require significant processing before they can be used, even with the use of tools like Google Earth Engine. We use imagery from Sentinel 2 and MODIS and in situ data from the State of Utah Ambient Water Quality Management System (AQWMS) database to develop models and to generate a highly accessible, easy-to-use CSV file of chlorophyll-a (which is an indicator of algal biomass), turbidity, and water temperature measurements on Utah Lake. From a collection of 937 Sentinel 2 images spanning the period from January 2019 to May 2025, we generated 262,081 estimates each of chlorophyll-a and turbidity, with an additional 1,140,777 data points interpolated from those estimates to provide a dataset with a consistent time step. From a collection of 2333 MODIS images spanning the same time period, we extracted 1,390,800 measurements each of daytime water surface temperature and nighttime water surface temperature and interpolated or imputed an additional 12,058 data points from those estimates. We interpolated the data using piecewise cubic Hermite interpolation polynomials to preserve the original distribution of the data and provide the most accurate estimates of measurements between observations. We demonstrate the processing steps required to extract usable, accurate estimates of these three water quality parameters from satellite imagery and format them for analysis. We include summary statistics and charts for the resulting dataset, which show the usefulness of this data for informing Utah Lake management issues. We include the Jupyter Notebook with the implemented processing steps and the formatted CSV file of data as supplemental materials. The Jupyter Notebook can be used to update the Utah Lake data or can be easily modified to generate similar data for other waterbodies. We provide this method, tool set, and data to make remotely sensed water quality data more accessible to researchers, water managers, and others interested in Utah Lake and to facilitate the use of satellite data for those interested in applying remote sensing techniques to other waterbodies.

Dataset: doi.org/10.5281/zenodo.15677448

Dataset License: CC0

Keywords: Utah Lake; water quality; remote sensing; chlorophyll-a; turbidity; MODIS; Sentinel

Academic Editors: Vladimir Sreckovic and Zoran Mijic

Received: 13 June 2025
Revised: 7 August 2025
Accepted: 11 August 2025
Published: 13 August 2025

Citation: Tanner, K.B.; Cardall, A.C.; Williams, G.P. A Six-Year, Spatiotemporally Comprehensive Dataset and Data Retrieval Tool for Analyzing Chlorophyll-a, Turbidity, and Temperature in Utah Lake Using Sentinel and MODIS Imagery. *Data* **2025**, *10*, 128. https://doi.org/10.3390/data10080128

1. Summary

Earth observation satellites provide unique and valuable data on environmental conditions all over the globe and have been used since the 1970s to study surface freshwater bodies. The key advantage of remotely sensed data provided by these satellites is spatial comprehensiveness—a single satellite image can provide data that covers an entire water body, as opposed to the single points provided by in situ samples. These data are useful for analyzing large-scale spatial trends and patterns in water quality and for accurately characterizing conditions on the entire waterbody, both of which are difficult to achieve with in situ samples. Remote sensing methods are particularly useful for analyzing long-term and spatial patterns in large, ecologically complex lakes where comprehensive in situ sampling is infeasible and water quality conditions are highly variable and difficult to predict.

Utah Lake, a basin-bottom lake in Utah Valley, USA, is uniquely suited to and can benefit from such remote sensing studies for several reasons and was the focus of one of the first aquatic remote sensing papers in 1974 [1]. The lake is large and spatially heterogenous, which means in situ samples fail to adequately characterize general conditions. It is highly turbid, which has been found to help reduce bias in spectral measurements over water [2,3], and it has a comprehensive history of in situ samples that can be used to develop and validate models for transforming spectral imagery into estimates of water quality parameters. Utah Lake is also located in an arid region with generally clear skies, making cloud-free optical imagery more consistently available.

Interest in Utah Lake's water quality dynamics is high due to concern over the lake's ecological condition and dissensus over the best methods for protecting it and mitigating issues like harmful algal blooms (HABs). The lake faces numerous challenges, including a rapidly expanding human population in its watershed, invasive species, and drought [4]. Predicting and managing the effects of such challenges requires detailed, comprehensive data on the lake's past and current state. Remote sensing is a powerful tool for studying large-scale trends, patterns, and relationships in water quality, and remotely sensed water quality data for Utah Lake facilitate the kind of spatially and temporally comprehensive analysis of water quality conditions that is required to better understand this highly unique ecosystem and to predict the potential success or failure of various proposed management strategies and the effects of future changes to its watershed [5].

Recent advances have greatly improved access to usable earth observation data, expanding its utility for large-scale water quality research [6–11]. Google Earth Engine (GEE) has made large remote sensing datasets much more accessible [12], but transforming even the processed satellite imagery in the GEE catalog into usable data is still a complex and time-consuming process. In addition, there are significant challenges associated with measuring water quality parameters spectrally, including interference from cloud cover, water's high absorption of red and near-infrared wavelengths [13], the difficulty of isolating the spectral signatures of materials of interest from other materials in the water with overlapping spectra [5,14], and the fixed but intermittent (and sometimes irregular due to cloud cover) revisit time of satellite imagery. Using earth observation data of Utah Lake presents further challenges due to the extreme optical complexity of its water column caused by high dissolved and suspended solids, biological activity, and wave action [15]. Utah Lake's extremely high turbidity and high rates of calcite precipitation can interfere with the accuracy of commonly used chl-a estimation algorithms due to spectral interference with the bands used in those algorithms [16]. Because of this, obtaining accurate estimates of chl-a for Utah Lake requires a more specialized approach.

In order to make remote sensing data on Utah Lake more accessible to researchers and water quality managers, we used a combination of image processing and data cleaning tech-

niques to produce a clean dataset of spatially and temporally comprehensive chlorophyll-a (chl-a), turbidity, and water temperature measurements derived from Sentinel 2 and MODIS imagery of Utah Lake in a convenient CSV format. We provide the processing methods in the form of a Python code notebook to generate the dataset as an example tool and to update the dataset. While the code is specific to Utah Lake, users can easily adapt it to different locations and waterbodies, making these tools general purpose.

We used images from the Sentinel-2 Multispectral Instrument, launched by the European Space Agency in 2015, due to its high spatial resolution (10 m), frequent revisit time (2–3 days), and variety of spectral bands useful for studying water quality [3,17]. We extended methods from Cardall, et al. [18] to generate the data using the GEE Python API. While tools and methods exist for processing all pixels contained in a lake or waterbody, these datasets are massive—millions of pixels per image and large numbers of images per year. For most analyses, a subset of pixels, randomly selected to represent lake processes, can be used to characterize lake processes and spatial patterns. We generated three different sets of 200 random points: one that covers the whole lake, one with 50 points each in four boxes placed in areas of the lake where we expected water conditions to vary, and one with 100 points in 'near-shore' areas of the lake and 100 points in 'open-water' areas. At each point we extracted the associated band values from every usable image in the collection. The subsample point method allowed us to obtain data even from partially clouded images, because we extracted values from individual pixels only if they were unaffected by clouds or cloud shadow. We used methods adapted from Cardall, et al. [18] to create empirical models fitted using in situ data and near-coincident satellite imagery to estimate chl-a and turbidity values. We acquired daily daytime and nighttime water temperature measurements at each point from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) imagery. We performed quality assurance checks on the data and interpolated the data to a daily timescale for use in statistical time series analyses.

We created this comprehensive water quality dataset for Utah Lake for use by researchers and water quality managers interested in Utah Lake. We provide both the dataset and tools used to generate it as an example of how to generate and use this type of data from satellite imagery and to allow others to easily adapt these tools to different locations. These data and the summary statistics represent highly accessible and comprehensive information about Utah Lake water quality and demonstrate an efficient method to access these types of remotely sensed image products and to generate reasonably sized data sets for analysis that capture both spatial and temporal patterns.

We used these data to support two research projects focused on HABs on the lake. The first analyzes relationships between chl-a, turbidity, and temperature on the lake to better understand how temperature and turbidity affect algal growth to inform strategies to reduce the blooms. The second analyzes temporal bloom 'dynamics'—assessing when and where on the lake algal blooms most often occur, how fast they grow and decline, how long they last, and their spatial behavior, as preliminary research indicates that most blooms initiate in either Provo or Goshen Bay, then move into the lake [15]. This information will help inform bloom mitigation strategies and cannot be generated with in situ water samples, which are spatially and temporally limited, and would be impractical to acquire using complete satellite images due to the very large data storage and processing requirements. Both studies will be submitted for publication and will reference this manuscript for their data.

The dataset presented here provides accessible, high-quality data on Utah Lake's water quality and serves as a resource for researchers and water managers investigating Utah Lake. It also serves as a template for similar studies on other waterbodies, demonstrating methods

for addressing challenges in using satellite imagery for water research and providing efficient data processing steps.

2. Data Description

The three datasets are comprised of daily values for chl-a (μg/L), turbidity (NTU), and daytime and nighttime water temperature (°C) either derived from band values from a satellite image or interpolated from surrounding data. The data sets are available as both supplementary materials with this manuscript (https://www.mdpi.com/article/doi/s1) and in a Zenodo repository (https://doi.org/10.5281/zenodo.15677448).

The three sets of sampling data are formatted as a single CSV file with the following columns:

- Date: Daily timestep from 1 January 2019 to 20 March 2025, formatted as mm/dd/yyyy.
- Point_id: A unique integer identifier for each sample point location.
- Latitude: In the WGS 84 projection.
- Longitude: In the WGS 84 projection.
- Dataset: A label indicating the set of sampling points that the data point belongs to: whole-lake, boxes, or clusters (see Section 3.3).
- Category: For the clusters and boxes datasets, a label that indicates which sub-category the data point belongs to. For clusters, either Open Water or Near Shore; for boxes, one of Provo Bay, North Lake, Center Lake, or Goshen Bay.
- In_PB: FALSE if the data point is outside Provo Bay, TRUE if inside.
- Int_flag: FALSE if the data point is from a satellite image, TRUE if the data point is interpolated from surrounding data (see Section 3.4). Note that this flag only applies to chl-a and turbidity data, not MODIS-derived data.
- Parameter: Chl-a, turbidity, dayTemp, or nightTemp. Chl-a and turbidity are derived from Sentinel 2 imagery; dayTemp and nightTemp are derived from MODIS imagery.
- value: computed parameter value for the image pixel in the location specified by the coordinates on the specified date, or the interpolated value if satellite data for that pixel was not available on that date.
    - The units for chl-a, turbidity, and temperature are μg/L, NTU, and °C, respectively.
- There are 5,611,432 rows in total, with 468,400 values for each of the three datasets and four parameters.

We published a second CSV file that includes the red, green, blue, and red edge band values extracted from GEE, which we used to compute the chl-a and turbidity values. We chose to publish this as separate data to make the main dataset as user-friendly as possible, but the values can be easily joined with the main dataset using the date and point_id fields.

- Date: Daily timestep from 1 January 2019 to 20 March 2025, formatted as mm/dd/yyyy.
- Point_id: A unique integer identifier for each sample point location.
- Red: Value of the red band.
- Green: Value of the green band.
- Blue: Value of the blue band.
- RE1: Value of the red edge 1 band.

3. Methods

3.1. Sentinel Image Processing

We extended methods developed by Cardall, et al. [18] to download and process imagery from the Sentinel 2 mission using GEE [12]. We used the harmonized Sentinel 2 MSI Level-2A orthorectified and atmospherically corrected surface reflectance image

collection available on the Google Earth Engine catalog. The Sentinel 2A and Sentinel 2B satellites provide images every 2–6 days starting in 2015, but the first usable images of Utah Lake are from 2019. This results in a collection of 937 usable images of Utah Lake collected roughly every 2–3 days (with larger gaps occurring mainly during the winter when cloud cover is more frequent) from 2019 to 2024. We accessed these images using tools built on the GEE Python API using Python (v 3.12) code implemented in a Jupyter Notebook (v7).

We used the Sentinel Scene Classification Layer (SCL) to filter pixels contaminated with clouds, cloud shadow, sensor failures, or other issues, and pixels that contained land, snow, or ice. We kept all pixels with an SCL value of 4, 5, 6, or 7, which represent vegetation, bare soils, water, and low-probability clouds, respectively. We chose to include the soils and vegetation categories because visual inspection of images showed that extremely turbid areas of Utah Lake were sometimes classified as soils in the SCL layer, and very intense algal blooms were sometimes classified as vegetation. Because we obtained data from the satellite imagery by extracting pixel values at known sample points, and our sample points were all in locations on Utah Lake that are known to be covered by water (see Section 3.3 for details), this method allowed us to preserve the most data possible and not erroneously exclude data from extremely turbid water or intense algal blooms. In addition, we masked each image using a 50% occurrence threshold from the JRC Global Surface Water Mapping Layers dataset [19] to further ensure that only pixels we could reasonably expect to be land-free were included in our dataset.

3.2. Remote Sensing Models

Remote sensing models provide information about water column conditions by measuring light reflection from optically active water column constituents such as chl-a (a plant pigment commonly used as an index for algal biomass) and suspended solids. Each of these has a unique "spectral signature"—absorbance and reflectance peaks at specific wavelengths. Multispectral satellite imagery measures the intensity of earth-leaving radiance in a set of spectral bands selected specifically to characterize various materials, such as vegetation and soils (which can be used for turbidity). We use the intensity of the radiance in the bands that correspond with a material's spectral signature to generate models that estimate the amount or concentration of that material in a pixel [5].

For some applications, physics-based models built on known spectral characteristics of a material provide sufficiently accurate results; however, for optically complex waters, such as Utah Lake, these models often fail due to the presence of other materials that have overlapping spectral signatures or cause unpredictable scattering of reflectance [15]. Turbid, productive lakes, like Utah Lake, contain optically active constituents with absorption features that overlap those of chl-a. For example, colored dissolved organic matter (CDOM) and detritus reflect strongly in the same blue–green area of the electromagnetic spectrum as chl-a, so empirical models that rely solely on blue and green spectral channels cannot provide accurate estimates of chl-a when CDOM and detritus are present [20]. The spectral features of algae blooms also vary with the species dominating the bloom and other water column chemistry conditions, so models based on in situ data from the study location provide more accurate results than generalized physics-based models [16].

In these cases, empirical regression models based on data pairs of in situ measurements and coincident or near-coincident satellite imagery provide better results. These models are fit using data from the image bands. The regression analysis quantifies relationships not apparent from physical characteristics alone [21]. This means a model based on observed data is better able to identify unique, specific spectral characteristics of a material in a certain area than a physics-based model.

Empirical models based on observed data are well-suited to Utah Lake for two reasons: first, it is extremely optically complex due to the high turbidity levels and high diversity of algal species, which vary throughout the season and year [22], making it difficult for a physics-based model to adequately characterize chl-a and turbidity. Second, because of its environmental and economic importance, there is a long-running and comprehensive history of water samples collected on Utah Lake, providing sufficient data to generate accurate empirical models.

3.2.1. In Situ Data

To build our models, we acquired in situ Utah Lake water quality measurements from the Utah Department of Environmental Quality's Ambient Water Quality Monitoring System (AWQMS). We obtained 752 measurements of turbidity and 1937 measurements of chl-a taken at various locations on the lake from 1978 to 2022. We then used GEE to identify and process Sentinel images of Utah Lake taken within 12 h of an in situ sample collected on Utah Lake. We identified the pixel associated with the in situ sample location and extracted the image band values to generate "matchups"—sets of in situ measurements and near-coincident satellite image data. We found 154 pixel matchups for chl-a and 113 pixel matchups for turbidity. The locations and number of matchups at each location are shown in Figure 1. The lake is fairly well-represented in the data, although there are more data points close to the shoreline than in the open lake.

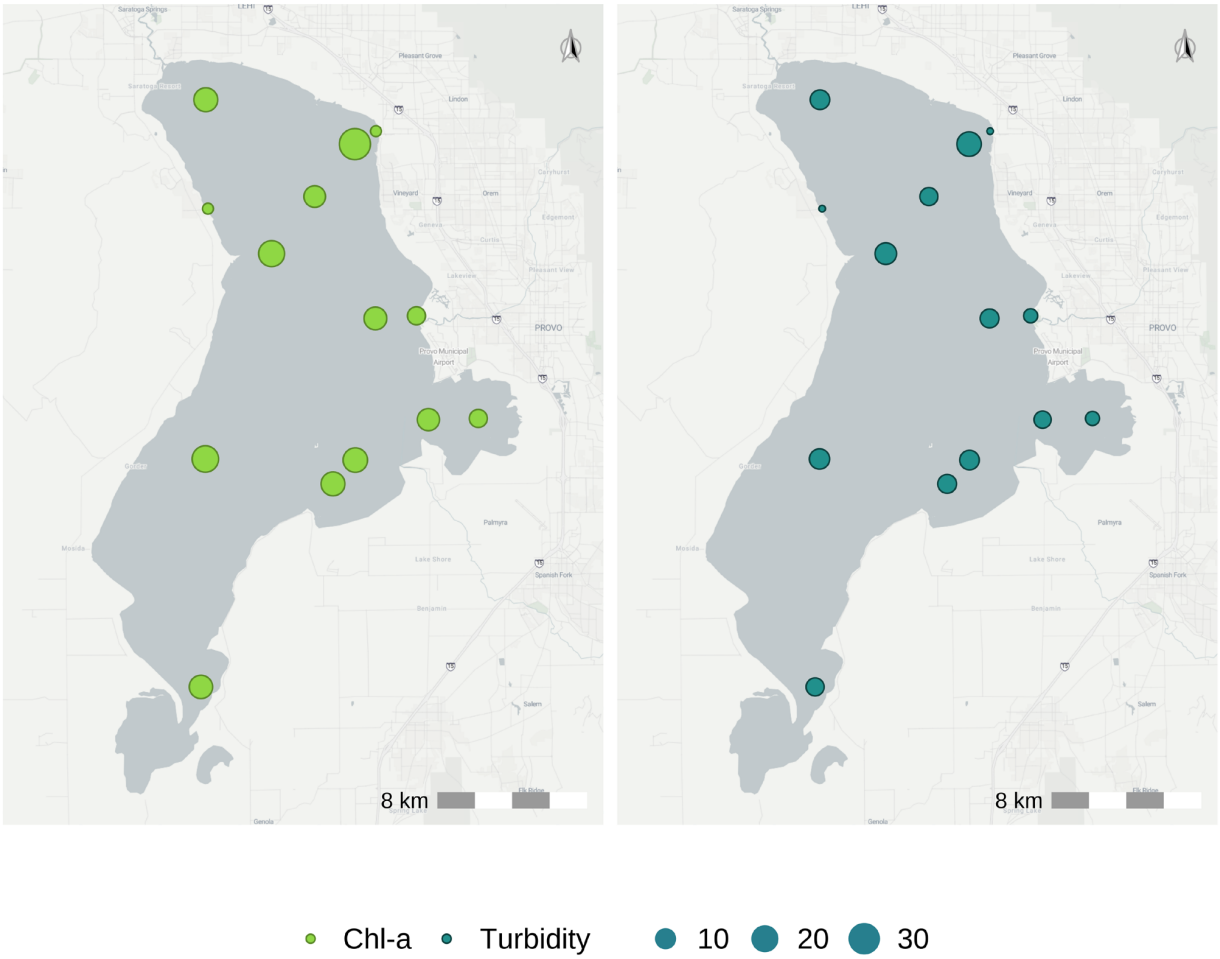

Figure 1. Map of locations represented by pairs of near-coincident Sentinel-2 satellite and in situ water quality measurements between 2019 and 2022. In situ data were collected by the Utah Department of Environmental Quality. Dot size indicates the number of data points at that location for which there was a near-coincident Sentinel 2 image.

We evaluated various methods for developing empirical models of chl-a and turbidity for Utah Lake. One method used machine learning to generate models using a near-exhaustive set of bands and band combination values as potential model features [21]. Another method used in situ data to improve a physics-based normalized difference index model (the difference of a pair of bands divided by the sum of those bands) by fitting a site-specific coefficient to the index value [20].

3.2.2. Combining Physics-Based and Empirical Models

Ultimately, a combination of these methods produced the most accurate models for both chl-a and turbidity on Utah Lake.

We tested regression models created using least absolute shrinkage and selection operator (LASSO) regularization, which selects model terms from a large feature space of all possible band combinations and modifications (such as the product of two bands, or the inverse or square of a band) [21]. The LASSO models were prone to overfitting and included an excessive number of terms with very large or very small coefficients and non-random errors. They were helpful, however, in identifying potential model features.

The normalized index models we tested generated good estimates when fitted with coefficients based on the in situ data but introduced an artificial limit on predicted values at the maximum of the function, which was lower than the values we expected to see based on in situ data. To address this, we added image band values identified by the LASSO models as being well-correlated with the target parameter as additional model features or terms, which increased the accuracy of the models and allowed them to predict values over the full expected data range.

We fit the models using the LinearRegression function from the scikit-learn Python library [23]. We evaluated the accuracy of the models using the coefficient of determination (R^2) and root mean square error (RMSE) values. For aquatic remote sensing models, R^2 values between 0.6 and 0.8 are commonly accepted [24], while evaluation of RMSE values depends on the range of the value being modeled. In [24], the authors achieved an R^2 of 0.98 with an empirical model for TSS.

For chl-a, the range of measured in situ values was 254 µg/L, with a standard deviation of 35 µg/L and a mean of 24 µg/L. Since the error of the in situ chl-a measurements was likely ~10% or greater [25], we considered models with RMSEs below 2.5 (i.e., within the average error) to be useful. For turbidity, the range of measured values was 158 NTU, with a standard deviation of 36 NTU and a mean of 25 NTU. The error for the turbidity measurements is much lower than for chl-a—around 0.3 FNU—but we considered models with RMSEs below 5 NTU, which is a small value relative to the range and average of the data, to be useful.

3.2.3. Chlorophyll-a

Although season-specific chl-a models have been shown to be more accurate on some lakes [26], the limited amount of data available and the extreme inter-annual variability of Utah Lake as a basin-bottom lake and a managed reservoir made this approach impractical. Season-specific models also require a significant amount of in situ data to characterize behavior in the different seasons. We were unable to generate useful models if we sub-setted the data into seasonal groups because of the limited amount of data for each season, so we used a whole-season chl-a model based on data collected between April and September.

We used methods similar to those described by Cardall, et al. [21] to analyze a number of potential model features and develop a sparse, interpretable chl-a model. We found that models based on the normalized difference chlorophyll index (NDCI) developed by Mishra

and Mishra [20] performed better than models based on other indices or on plain band values. The NDCI is calculated as follows:

$$\text{NDCI} = \frac{\rho_{red} - \rho_{RE}}{\rho_{red} + \rho_{RE}} \quad (1)$$

where ρ_{red} and ρ_{RE} represent the red and red-edge band values, respectively.

We fit the linear regression coefficients, including terms for the NDCI value and the square of the NDCI value. This model provided acceptable results (R-squared values around 0.8); however, it introduced an artificial limit on predicted chl-a values at the maximum of the equation. The maximum from this initial equation was approximately 110, which is much lower than the highest chl-a values we expected based on in situ measurements, some of which were as high as 300 µg/L. To address this, we added two band value terms to the linear regression equation; this allowed the model to predict chl-a values over the full expected range.

The final adjusted NDCI model is as follows:

$$\ln(chla) = 2.90 - 8.95(\text{NDCI}) - 17.20\left(NDCI^2\right) - 21.96(\rho_{blue}) + 17.11\left(\rho_{green}\right) \quad (2)$$

where ρ_{blue} and ρ_{green} are the red and green band reflectance values, respectively. This model produced an R^2 value of 0.80 and an RMSE of 0.48 (an RMSE of 1.62 when unlogged).

3.2.4. Turbidity

We used a similar modeling approach—a combination of a normalized difference and unmodified band values—to produce the turbidity model. This hybrid model produced better results for turbidity than other types of models. We tested the normalized difference of every pair of bands available from Sentinel 2 and found the blue–red pair had the best correlation with Utah Lake turbidity values. We called this the Utah Lake Normalized Difference Turbidity Index (ULNDTI) and calculated it as follows:

$$\text{ULNDTI} = \frac{\rho_{blue} - \rho_{red}}{\rho_{blue} + \rho_{red}} \quad (3)$$

where ρ_{blue} represents the blue values.

A turbidity model with ULNDTI and squared ULNDTI values as the only terms created an artificial limit on the predicted values (the same issue we found with the chl-a model), so we tested various plain band additions to the model to eliminate this limit. We found that adding a single red-edge band value to the model produced the best results, and the final ULNDTI model was as follows:

$$\ln(turbidity) = 2.47 - 2.73(\text{ULNDTI}) + 0.21\left(\text{ULNDTI}^2\right) + 10.70(\rho_{RE}) \quad (4)$$

which produced an R-squared value of 0.89 and an RMSE of 1.25 (3.50 unlogged).

We generated data based on this model. We excluded model estimates of turbidity with values above 500 NTU because visual inspection of the satellite images of the points associated with these high values indicated that the points with these high values were associated with locations where water levels were much lower than normal and the pixel contained both water and land, or the water was shallow enough to image the lake bottom through the water column. Based on the in situ data, we would not expect to see turbidity values above 500 NTU.

3.2.5. Temperature

We provide temperature values from the lake surface water temperature (LSWT) product, derived from the MODIS Global Daily Surface Temperature and Emissivity 1 km image collection from the Aqua satellite available on the GEE data catalog, without modification. Numerous studies have validated the use of LWST data where in situ lake temperature data are not available [27–32]. Studies of LWST data on Great Salt Lake, which is near Utah Lake, and Lake Taihu in China, both of which share many characteristics with Utah Lake, found a cold bias in LWST estimates for those lakes [33,34], and such a bias may exist for Utah Lake. However, Lazhu, et al. [27] point out that the observed biases may come from a failure to consider the representativeness of in situ samples for the entire area of these large, shallow lakes with slow lateral mixing. In addition, since the use cases for these data are concerned mainly with trends, a consistent bias in one direction (as would likely be the case on Utah Lake) should not affect the validity of these data for analyzing and characterizing trends or changes in lake conditions. Nevertheless, the results of analyses based on MODIS temperature data should be checked against in situ data whenever possible, and the possibility of the cold bias should be considered.

We observed that a portion of the MODIS-derived temperature measurements on Utah Lake were outside the reasonable range of water temperatures, which is to be expected with temperature measurements based on surface emissivity, but the overall distribution of MODIS temperature data matched the distribution of in situ temperature measurements which are presented in detail later in the text. We chose not to exclude or adjust the out-of-range temperature measurements since they are still useful for correlation and trend analyses.

Because MODIS pixels are large, pixels near the shoreline of the lake include both land and water, which means the temperature values in these pixels are not representative of the water temperature. To eliminate the effects of mixed pixels in our data while including temperature for the near-shore areas, we used a nearest-neighbor spatial interpolation technique to replace values from mixed pixels with values from the closest 'pure' pixel. Section 3.4 provides details and justification for this approach.

3.3. Image Sampling and Model Application

Remote sensing studies performed on large areas and over longer time scales use data aggregation or subsampling to extract meaningful insights from large amounts of spectral image data. The method and scale of aggregation or subsampling depend on the characteristics of the study location—if it is largely homogenous, then an aggregate value, such as the mean or median, for the entire area may suffice, but for areas with high heterogeneity, it is necessary to aggregate over smaller areas or use subsamples. We assumed that Utah Lake is spatially heterogenous relative to most other waterbodies because it is not well-mixed laterally—there is significant variation in water column conditions throughout different areas of the lake, especially when algal blooms are present [35]. To characterize the spatial heterogeneity of the lake while also maintaining the dataset at a manageable size, we subsampled the image data by extracting the pixel values at pre-defined sets of points from each image rather than calculating aggregated statistics. This subsampling technique had the following additional benefits:

1. Making the data suitable for statistical analyses that assume a random sample.
2. Allowing us to extract more usable data from images with partial cloud cover.
3. Eliminating the need for complex and imprecise water masking procedures, because we located sample points only in areas where we knew there would be water.
4. Providing a smaller, more accessible dataset relative to the extremely large and complex data set of satellite imagery over extended time periods.

We created three data sets for the lake with the following subsample approaches: one subsample dataset represents the entire lake area, one dataset represents four different areas of the lake we expected to exhibit different behavior based on previous research [15], and one dataset represents the lake divided into near-shore and open-water regions. We identified these near-shore and open-water regions with a machine learning clustering algorithm. The published data includes all three datasets and a flag identifying the dataset to which each measurement belongs.

3.3.1. Whole-Lake Samples

The shoreline of Utah Lake fluctuates significantly with water level, so we could not use a geographically defined shoreline or the shoreline from a single image as the boundary for the set of whole-lake sample points. Instead, we created a shoreline boundary by building a composite image of all Sentinel 2 images of Utah Lake taken during the growing season (April–October) and then applying the modified normalized difference water index (MNDWI) to the composite image and defining pixels with an MNDWI value above 0.25 as water. This provided a minimal lake boundary for the time period covered by the Sentinel 2 images used in this data set. We then generated 200 data points with the locations randomly selected from within this lake area.

Figure 2 shows the 200 randomly generated sample points in the whole-lake collection. There are no sample points at the very south end of the lake in the shallow part of Goshen Bay. Although it is defined as part of Utah Lake on most maps, during the time period covered by the Sentinel 2 data set, this area was dry, so it did not register as water in the MNDWI-thresholded composite image. Therefore, our Sentinel 2-specific lake boundary, which otherwise matches basemap boundaries quite closely, does not include the southernmost area of Goshen Bay, which is correct for our study period.

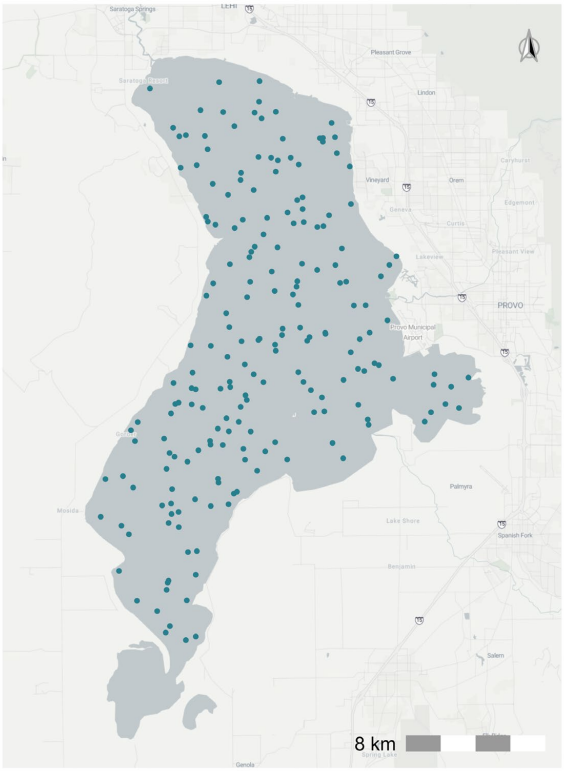

Figure 2. Whole-lake sampling point collection with 200 randomly placed sample points. There are no sampling points at the southern tip of the lake in the shallow part of Goshen Bay because this area was dry during the period covered by the Sentinel 2 collection.

3.3.2. Boxes Sample

We defined four 1.5 km (0.93 mile) boxes (150 hectares) in areas of the lake where we expected different water column conditions based on prior research and knowledge of the lake (Figure 3): one at the north end of the lake near the Jordan River Outflow (North Lake); one in the center of the lake (Center Lake); one at the south end of the lake just above Goshen Bay (Goshen Bay); and one in the center of Provo Bay on the east side of the lake (Provo Bay). Provo Bay is shallow (less than a meter deep on average), hydrologically isolated from the rest of the lake, receives a large portion of the anthropogenically impacted inflow to the lake, is heavily vegetated, and experiences intense and frequent algal blooms. Goshen Bay is completely open to the main lake but is also extremely shallow and surrounded by agricultural land. The Goshen Bay box is just outside the bay itself in the part of the lake nearest the bay outlet (Figure 3). Both Provo and Goshen Bays tend to experience more frequent and intense algal blooms than the main body of the lake [15]. The Center Lake location represents a region in the main lake with deeper water (typically about 3 m) that experiences greater wind and wave action than the bays. The North Lake location represents a region similar to Center Lake but slightly less deep and the closest to heavily developed land as well as the only outlet for the lake. Also, the prevailing wind is to the north, so algal blooms can be driven towards this area.

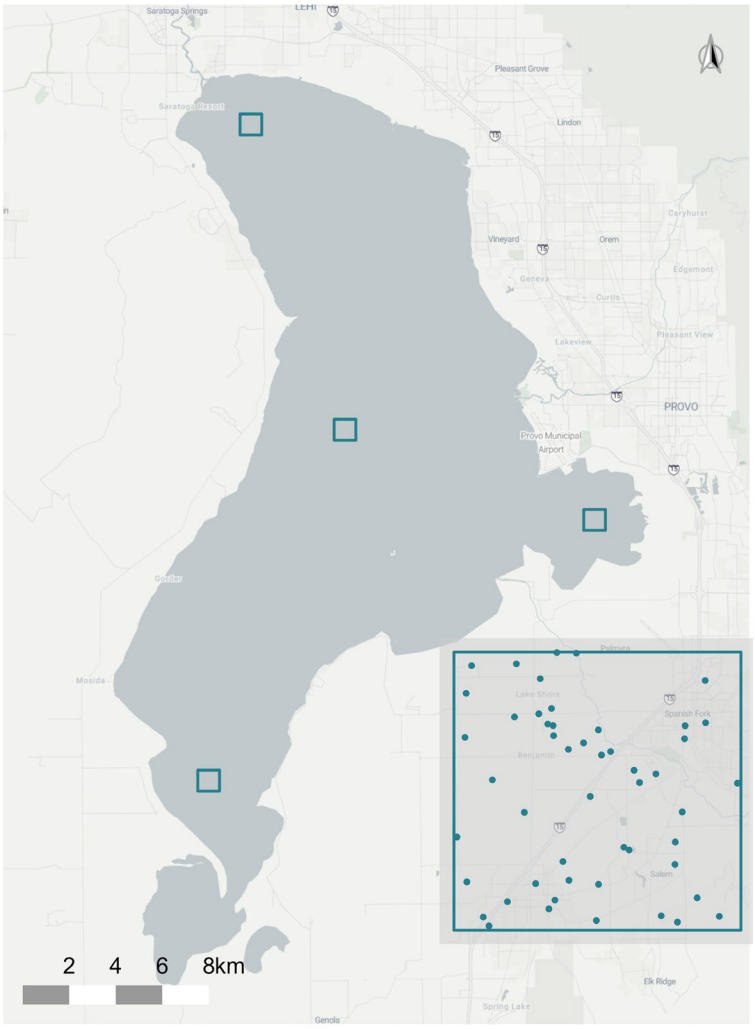

Figure 3. Locations of the four boxes, 1.5 km (0.93 mile) on a side (150 hectares); sample boxes with inset showing the random sampling pattern used to extract 50 data points from each box. We used the same spatial distribution of sample points in each of the four boxes.

We generated a sampling pattern of 50 randomly located points using the GEE randomPoints function within a 1.5 km box. We used the same spatial distribution of sample points in each of the four boxes. The locations of the boxes and the random sampling pattern are shown in Figure 3.

3.3.3. Machine Learning Near-Shore, Open-Water Clusters

Based on prior research and experience with Utah Lake, we knew that near-shore water column conditions often differ greatly from conditions on the open lake. We wanted to provide data for analyses that accounted for this phenomenon, but defining what exactly is near-shore versus open-water on Utah Lake is difficult due to the lake's shallow bathymetry and dynamic shoreline. Rather than trying to define an arbitrary boundary or buffer around the shore, we used a machine learning algorithm to identify areas where pixels tend to share characteristics over multiple images. With this clustering algorithm we delineated near-shore and open-water areas of the lake based on prevailing patterns in spectral characteristics for those areas.

We used the WekaXMeans clustering algorithm [36] in GEE because it determines the correct number of clusters itself rather than requiring the user to set the number, and we were unsure of how many distinct clusters would be present in the lake. WekaXMeans is a K-means-based algorithm that incorporates model selection to more efficiently estimate the number and location of clusters using a Euclidean distance function. To reduce the computational requirements for the algorithm, we filtered the image collection to only include images with less than 50% cloud cover (which resulted in 598 images) and collapsed the image collection into a single multi-band image by calculating the 25th, 50th, and 75th percentile band values across the entire image collection for every pixel. The resulting image included the three values for each percentile as a band in the image for each of the original image bands—one for the 25th percentile values, one for the 50th, and one for the 75th. Thus, for the Sentinel images with 10 optical bands, this results in a 30-band image. We applied WekaXMeans to these band data to identify clusters.

In multiple trials, the algorithm consistently identified three clusters within the lake (and a fourth cluster representing land), as shown in panel A of Figure 4. The algorithm does not assign physical meaning to these clusters, but by inspection we classified them as follows: Cluster 2 represents pixels that contained only land in every image in the collection. Cluster 1 represents 'mixed' pixels, which likely contain both land and water in a single image or contain water in some images and land in others. We chose to exclude data from Clusters 1 and 2 because estimates of chl-a and turbidity based on our models are only valid for pure water pixels. The two clusters of interest are Clusters 3 and 4, which represent open-water and near-shore pixels, respectively. Panel B shows the random sampling pattern, with 100 points per cluster, that we used to extract data from each of these clusters. The clustering algorithm classified the majority of Provo Bay and the area near the mouth of Goshen Bay as near-shore, likely because of the shallow nature of these bay areas and water characteristics such as suspended sediments.

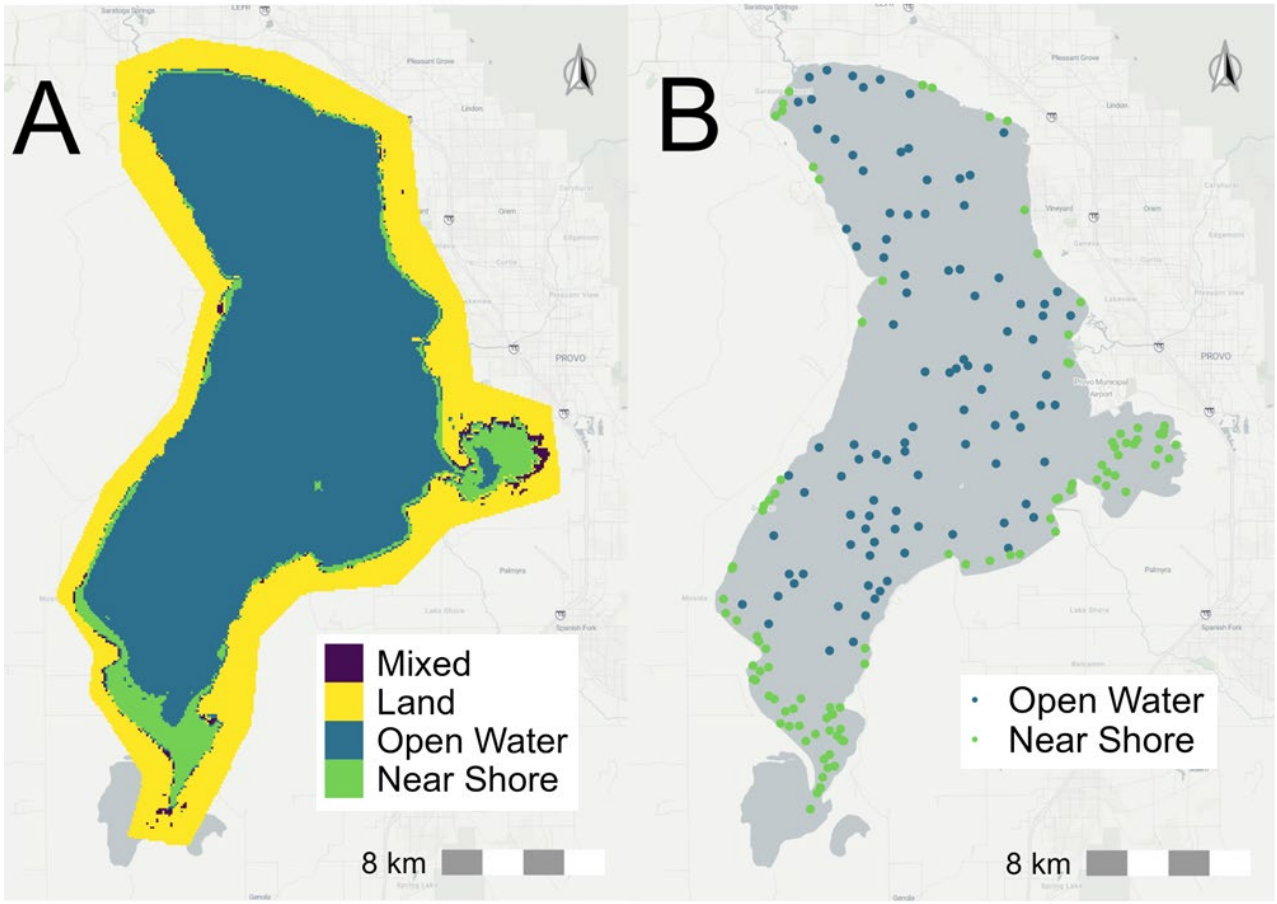

Figure 4. Map of sample clusters. Panel (**A**) shows the cluster areas, and panel (**B**) shows the 200 sample points (100 per cluster) used to extract data in the open-water and near-shore clusters.

3.3.4. Model Application

For each of the three datasets (the whole-lake sample, sampling boxes, and the clusters), we used GEE to export a CSV of band values and MODIS temperature data for every sample point in each image of the image collection. We then used Python to apply the models described in Section 3.2 to the band values for each data point, generating an estimate of chl-a and turbidity for each point. This was more efficient for computation and storage requirements than applying the models to the entire images in GEE and allowed us to perform some additional quality checks on our data.

3.4. Imputation of Missing and Anomalous Values

To facilitate time-series analysis on the irregularly spaced satellite data, we created daily data with no missing values. We used a combination of imputation with mean values (for temperature data) and piecewise cubic Hermitean interpolation polynomials (PCHIP) to interpolate temporally between data points. We used PCHIP interpolation, as it honors local data limits and constructs smooth curves between data points while preserving the shape and monotonicity of the data. Unlike standard cubic splines, PCHIP avoids overshooting and produces more realistic interpolations for datasets with sharp changes or non-uniform behavior. In Section 5, Summary statistics we compare the distributions of the original chl-a, turbidity, and temperature values from each dataset with the distributions of the interpolated values, shows that the distributions match well. This means that PCHIP generated reasonable values between observed data points that do not skew the observations higher or lower or introduce outliers. Unless otherwise specified, each of the three datasets underwent the same processing steps.

3.4.1. Adjusting Anomalous Near-Shore Temperature Values

We observed that MODIS temperature values within ~1 km of the Utah Lake shoreline were often unrealistically high because they came from 'mixed' pixels that included both land and water. We determined that this effect was entirely due to the mixed pixel effect rather than any real pattern of warmer water near-shore by analyzing data from the AWQMS database, which showed that, although there are often spatial differences in temperature, near-shore water is not consistently warmer than water more than 1 km from shore. Because there was not sufficient evidence for a consistent relationship between open-water and near-shore temperatures and we could not quantify the influence of mixed pixels on temperature data, we instead replaced MODIS temperature values collected within 1 km of the Utah Lake shoreline with the value of the nearest MODIS pixel that was further than 1 km from the shoreline. We identified nearest neighbors with the NearestNeighbors function from the scikit-learn package in Python [23] and checked the results after replacement by visually inspecting the data for several images and examining the distributions to ensure there were no extreme shifts as a result of the replacements.

The 1 km buffer excluded the entirety of Provo Bay, but there is one location in the center of the bay with a single unmixed pixel from MODIS where it is possible to collect temperature data. We extracted temperature values for Provo Bay from that point and used these data as temperature values for Provo Bay in all three datasets. We found this to be the best way to exclude unusable data while still preserving usable information about Provo Bay, because it is hydrologically isolated and does not always move with the rest of the lake, so values taken from the open lake would not have been representative of the bay.

3.4.2. Imputing and Interpolating Missing Temperature Data

MODIS products have built-in cloud masks. MODIS pixels are much larger than Sentinel pixels, 1 km compared to 10 m, so the MODIS cloud masks usually cover a larger area than the Sentinel cloud masks, since the higher resolution of Sentinel allows more pixels to be preserved. This resulted in some data points with usable values derived from the Sentinel data having no associated values for temperature. MODIS does, however, provide daily data, while Sentinel data are only collected every 2 to 3 days. Since temperature values do not vary significantly across the lake (the average difference in the minimum and maximum temperature measured on the lake over the study period is 2 °C, with a maximum difference of 15 °C that occurred on two days), we used the median temperature value for the lake to impute missing data rather than rely on temporal interpolation. So for any MODIS image that had at least one usable temperature measurement but with locations missing measurements due to clouds or cloud shadow, we imputed the values for those locations with the median of the locations with usable data for that image. When no data were available (i.e., in the case of total cloud cover), we interpolated the temperature data through time using PCHIP to generate the missing values.

Table 1 shows the number of imputed and interpolated values for the MODIS data for the three sample data sets. Each dataset has 200 points for each time step for a total of 434,800 data points. This reflects the 2023 days from 1 January 2019, when usable Sentinel data starts, to 15 July 2024.

Table 1. Number of values imputed and interpolated for MODIS temperature datasets.

Dataset	Parameter	Values Imputed with Mean	Values Temporally Interpolated with PCHIP
Whole-lake	Day Temp	102,895 (22%)	623 (0.14%)
Whole-lake	Night Temp	97,590 (21%)	640 (0.15%)
Boxes	Day Temp	102,485 (22%)	722 (0.04%)
Boxes	Night Temp	102,522 (22%)	711(0.04%)
Clusters	Day Temp	104,462 (23%)	614 (0.04%)
Clusters	Night Temp	100,103 (22%)	426 (0.02%)

3.4.3. Interpolating Missing Chl-a and Turbidity Values

The Sentinel revisit time alternates between 2 and 3 days. Occasionally, due to partial or complete cloud cover or the presence of ice on the lake, the time gap was longer, but the vast majority of gaps in the data were either 2 or 3 days, as shown in Figure 5, which we created using the set of 200 whole-lake sampling points. The few instances of longer gaps between data points nearly always occurred during the winter months, when cloud cover and ice were more frequent. We used PCHIP to temporally interpolate between Sentinel data to provide data on a daily timescale.

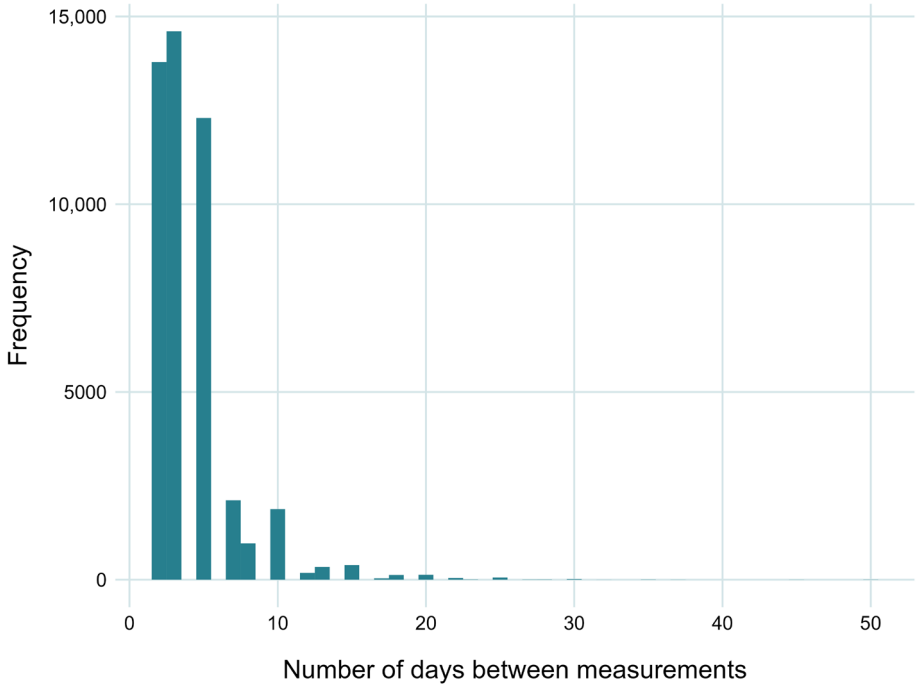

Figure 5. Histogram of data gaps for the sample points in the whole-lake dataset. These data include winter months when cloud cover and ice are more frequent.

In addition to filling in data between the satellite revisit times, we also used PCHIP to fill in data from partially clouded images. Since turbidity and chl-a can vary significantly spatially throughout the lake, we found it was not accurate to spatially impute chl-a and turbidity values missing due to cloud cover with the mean value for un-clouded pixels for that day, as we did with MODIS data. Instead, we imputed missing chl-a and turbidity values using temporal interpolation based on values of that same pixel from previous and future images.

We interpolated daily values for chl-a, temperature, and turbidity for each sample point using the PchipInterpolator function from the SciPy package in Python [37]. Figure 6 shows PCHIP interpolation results and demonstrates how it honors the existing

data without overshooting the local points. We did not use the interpolated data for computing statistical properties such as the means/medians, because PCHIP assumes a smooth transition between two points when there could have been a dip or spike in actual values. We intend the interpolated data to be used mainly for time series and trend analysis where a consistent time step is required.

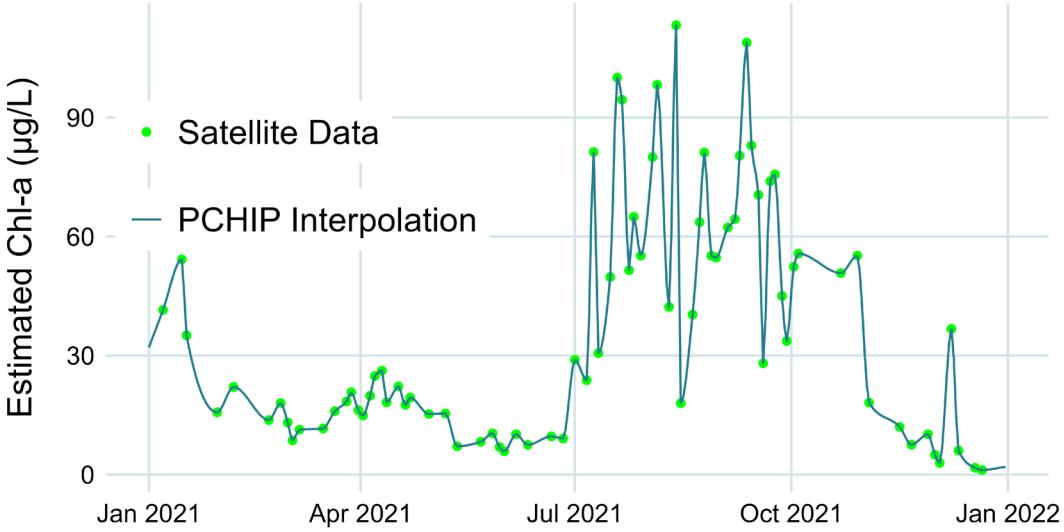

Figure 6. Example of PCHIP interpolation for 2021 data for the sample point at coordinates 40.14755–111.8626. This figure contains 79 measurements and 286 interpolated points.

For the whole-lake, boxes, and clusters datasets, we interpolated 376,805 data points from 91,595 observations, 385,892 data points from 82,508 observations, and 378,080 data points from 87,978 observations, respectively. There are more interpolated points than measurements because we generated daily data from data collected only every 2 or 3 days that also had gaps due to clouds, ice, and other issues. We recommend the interpolated data be used primarily for time series analysis, while actual measurements be used for other statistical characterizations.

4. Jupyter Notebook Implementation

4.1. Notebook Description

We have included a Jupyter Notebook that demonstrates the workflow we used to generate the datasets described in this paper. This Notebook can be easily modified for other locations, or to change sampling locations on Utah Lake. The following is an overview of the code that we used to implement the methods described above.

4.2. Notebook Outline

1. Setup
    a. Load packages and set up GEE API;
    b. Define the rough lake outline with hand-selected coordinates (to analyze a different water body, the user can supply different coordinates).
2. Retrieve and process satellite image collection
    a. Load Sentinel 2 data;
    b. Apply processing functions that scale bands, perform initial quality assurance, and rename bands for future processing.
3. Create sample point collection
    a. Whole-lake collection

i. Generate a whole-lake boundary by creating a composite Sentinel 2 image and applying the Modified Normalized Difference Water Index;
ii. Export boundary as a GEE asset;
iii. Generate sample points inside the lake boundary, add metadata features, and export the feature collection as a GEE asset.

b. Cluster collection
i. Generate cluster boundaries by creating a composite Sentinel 2 image and applying a clustering algorithm to computed band percentiles;
ii. Export clusters as polygons to GEE asset;
iii. Generate sample points inside cluster boundaries, add metadata features, and export feature collection as GEE asset.

c. Boxes collection
i. Generate box boundaries with user-selected coordinates and export as a GEE asset (to analyze a different water body, the user can supply different coordinates);
ii. Generate sample points inside box boundaries, add metadata features, and export feature collection as GEE asset.

d. Combine the three feature collections into one, add a point ID for future merging, and export as a GEE asset and as a shapefile for visualizations.

4. Obtain Sentinel 2 band values from sampling points
a. Load the combined points collection and extract pixel values from Sentinel 2 images at the specified points;
b. Export pixel data with date, location, and metadata to Google Drive (cannot export to asset due to GEE's memory limits even with this reduced dataset).

5. Obtain MODIS temperature values from sampling points
a. Retrieve and process MODIS imagery collection
i. Apply processing functions that scale bands and set metadata properties;
ii. Extract pixel values from images at the specified points;
iii. Export pixel data with date, location, and metadata to Google Drive (cannot export to asset due to GEE's memory limits);
iv. Extract temperature values from the single usable MODIS pixel in Provo Bay and export to Google Drive (not necessary for other waterbodies unless there is a similar issue with a small area entirely excluded by the 1 km buffer).

b. Process extracted MODIS data
i. Replace values for pixels located in Provo Bay with the value of a single unmixed pixel (not necessary for other waterbodies unless there is a similar issue with a small area where only a single MODIS pixel is valid);
ii. Replace values of pixels within 1 km of shore with nearest neighbor values
iii. Impute missing data from partially clouded images with daily lake temperature median;
iv. Impute missing data from fully clouded images with PCHIP temporal interpolation.

6. Apply pre-defined chl-a and turbidity models to Sentinel 2 data
a. Load the exported dataset of band values;
b. Apply band models and filter for valid values;

c. Impute missing data with PCHIP temporal interpolation.

7. Combine and export the final dataset
   a. Merge processed Sentinel 2 and MODIS datasets and perform additional data cleaning and formatting.

4.3. Key Outputs

- Waterbody boundary polygon shapefile.
- Random sample point dataset with coordinates.
- Chl-a and turbidity estimates for each point from Sentinel 2.
- Day and night surface temperature estimates for each point from MODIS.

This notebook provides example code that implements most of the processes we used to create the dataset published with this paper. It demonstrates data fusion between high-resolution optical (Sentinel-2) and low-resolution thermal (MODIS) imagery and generates daily data for analysis. The notebook is readily adaptable to other lakes where users can define their own sample regions or boxes and provides a clear demonstration of how to generate valid, accessible, and useful water quality data from satellite imagery using the GEE platform.

5. User Notes

We did not exclude data that appeared to be outliers—anomalous values due to sensor error or mixed pixels were already excluded by the QA process. We expect occasional very high and very low values in all three computed parameters because of the high variability intrinsic to environmental data.

There are no null or missing values in the dataset because any missing data (whether due to cloud cover, sensor error, or temporal gaps in the satellite imagery) were either imputed with the median value or interpolated. Imputed and interpolated data are flagged with a TRUE value in the int_flag column.

We generated the data from a spatial random sample, but these data do not meet the assumption of independence (temporally or spatially) because they are next to each other in the same body of water, and these processes have both spatial and temporal correlation.

Provo Bay is treated as a separate water body in Utah State law, so it may be useful to analyze it separately from the main lake using either the Boxes dataset or the 'in_PB' data flag in the data, which identifies samples within Provo Bay.

5.1. Summary Statistics

We present summary statistics for the un-interpolated datasets (Table 2, Figures 7 and 8). Since the purpose of the daily interpolated datasets is for time-series statistics that require regular time steps, and the distributions of those datasets are highly similar to the un-interpolated datasets (see Figure 8), we did not provide summary statistics for the interpolated datasets other than Figure 8, which compares the distributions of the interpolated and un-interpolated datasets.

Table 2. Summary statistics for the three datasets. Note that we combined the sub-categories within the Boxes and Clusters datasets for these descriptive statistics.

Dataset	Parameter	Min	Max	Standard Deviation	Interquartile Range	Skewness	Kurtosis
Whole-Lake	Chl-a	0.0	290	24	21	1.86	6.30
Whole-Lake	Turbidity	1	499	34	30	3.05	19.97
Whole-Lake	Day Temp	−7	45	10	19	0.047	1.68
Whole-Lake	Night Temp	−20	30	34	30	−0.12	1.92
Boxes	Chl-a	0	235	28	38	1.25	4.00
Boxes	Turbidity	1	500	35	30	3.49	25.20
Boxes	Day Temp	−7	37	10	19	0.05	1.70
Boxes	Night Temp	−18	27	9	16	−0.12	1.92
Clusters	Chl-a	0	301	26	29	1.53	5.05
Clusters	Turbidity	1	500	39	35	2.94	17.85
Clusters	Day Temp	−9	44	10	19	0.06	1.71
Clusters	Night Temp	−20	28	10	16	−0.10	1.93

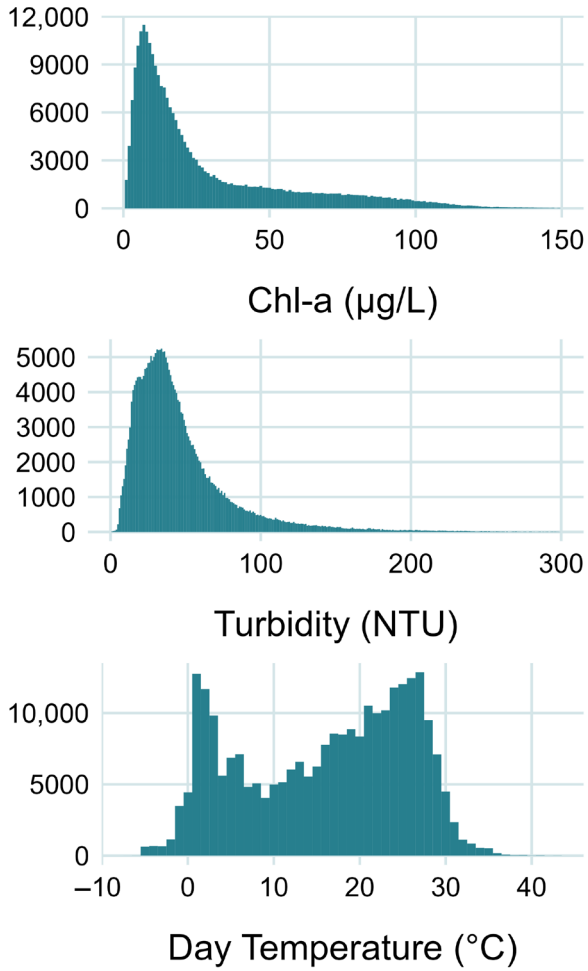

Figure 7. Histograms of the three parameters (combined datasets). Counts for temperature are higher because this is un-interpolated data, and MODIS has a daily revisit time, which more than doubles the amount of data compared to Sentinel 2 with its 2–3 day revisit time.

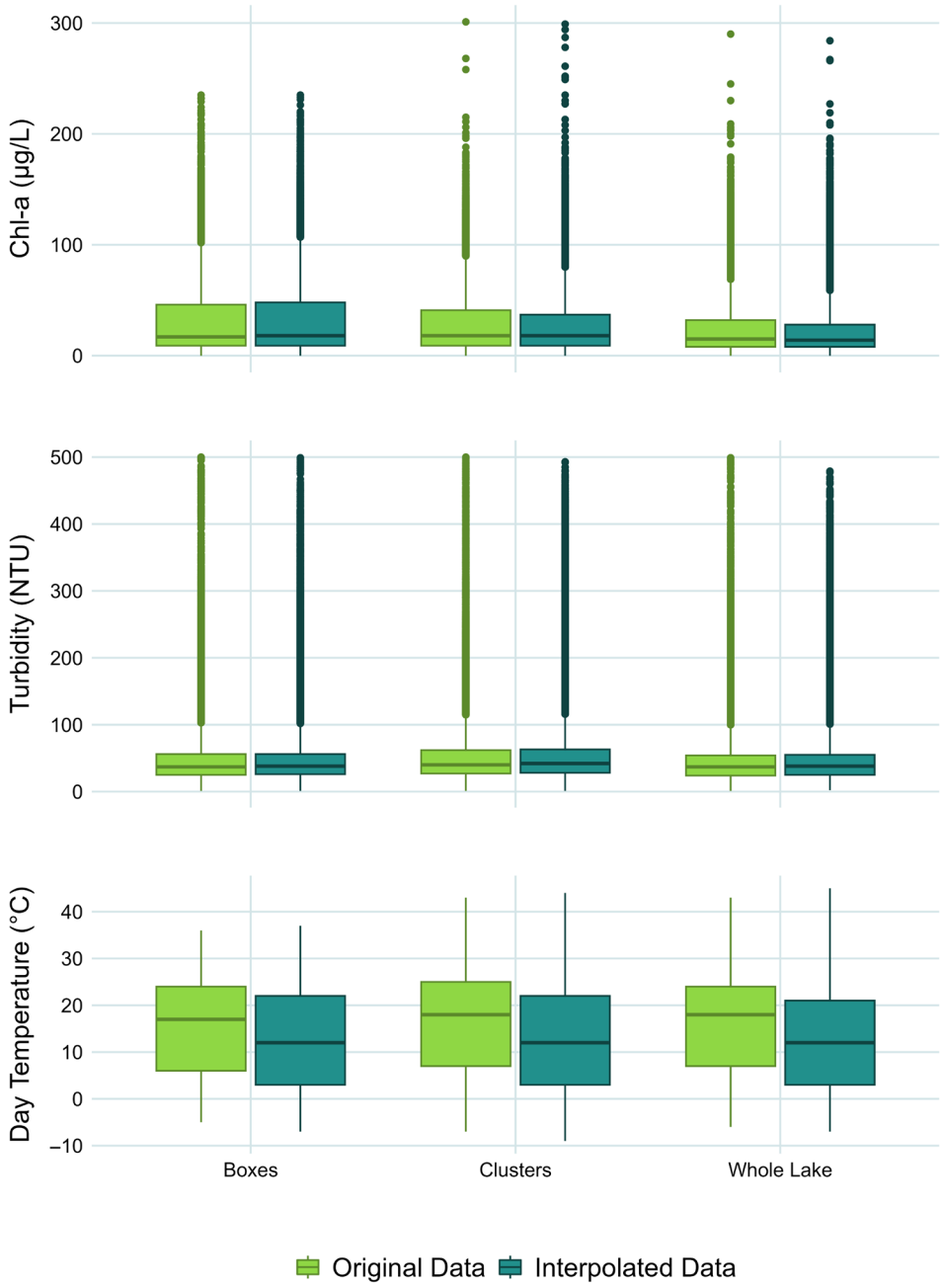

Figure 8. Comparison of distributions of original and interpolated data. The box ends represent the 25th and 75th percentiles, with the center line the 50th percentile or median value. The whiskers represent 1.5 × IQR (interquartile range), with the dots representing the outliers, or values higher or lower than 1.5 × IQR. The interpolated temperature distributions are much smaller because there are so few interpolated data points for temperature, as most missing temperature data points were filled with the median imputation method described in Section 3.4.2.

5.2. Summary Plots

Figure 8 compares the distributions of the interpolated and original data (all datasets) for each parameter. The interpolated temperature data have a much tighter distribution than the measured data because the interpolated data set is small. As discussed in Section 3.4, for Sentinel data, PCHIP-interpolated datasets are 99–120% the size of the measured datasets they were based on due to the longer gaps between Sentinel images, while the interpolated MODIS datasets are between 0.14% and 0.99% the size of the measured datasets, since MODIS has a daily revisit time and there were very few days in the collection where the lake was completely clouded, so many missing data points were filled

with mean imputation rather than interpolation. We provide Figures 9–13 to provide an overview of the data and its distributions. These graphs show both box plots and time series for the data grouped by location.

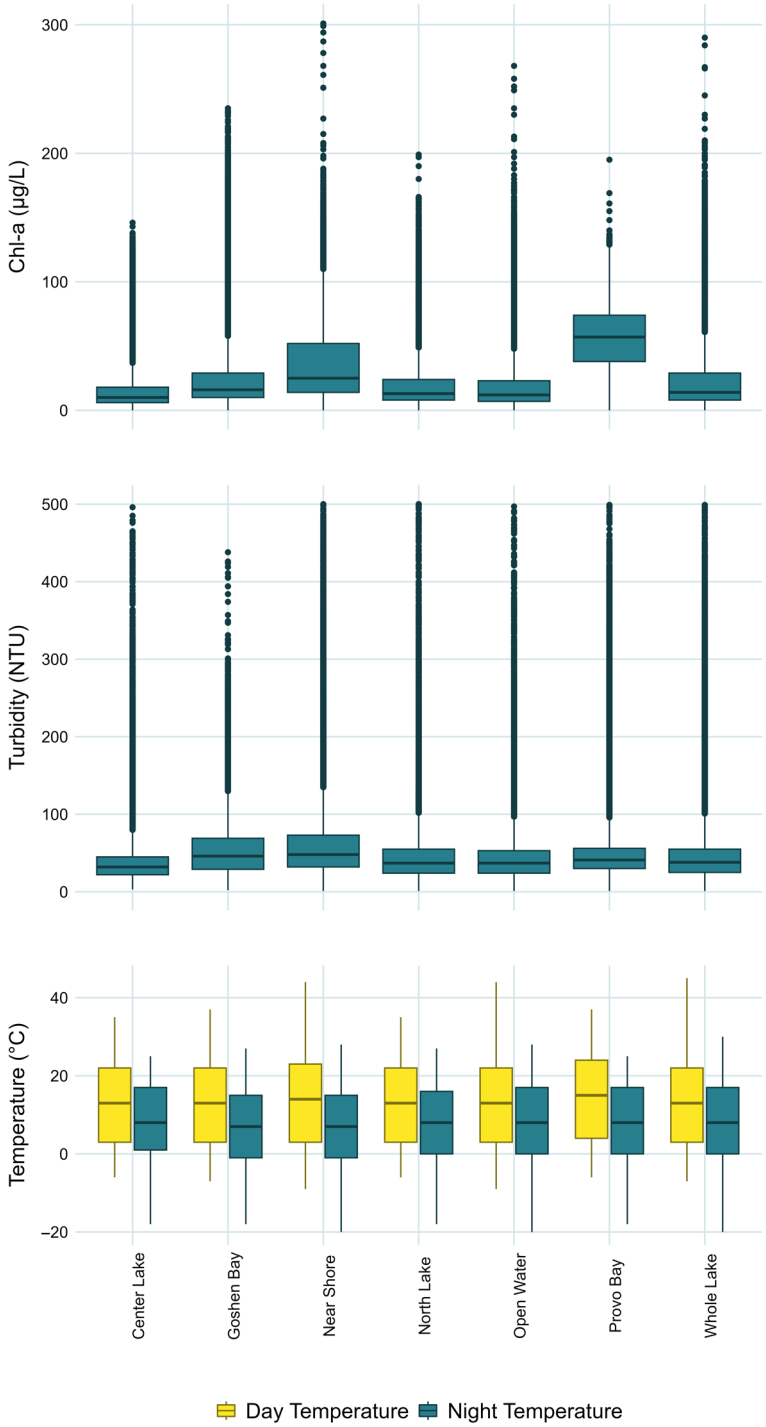

Figure 9. Distribution of chl-a, turbidity, and temperature measurements by sample set. The box ends represent the 25th and 75th percentiles, with the center line the 50th percentile or median value. The whiskers represent 1.5 × IQR (interquartile range), with the dots representing outliers, or values higher or lower than 1.5 × IQR.

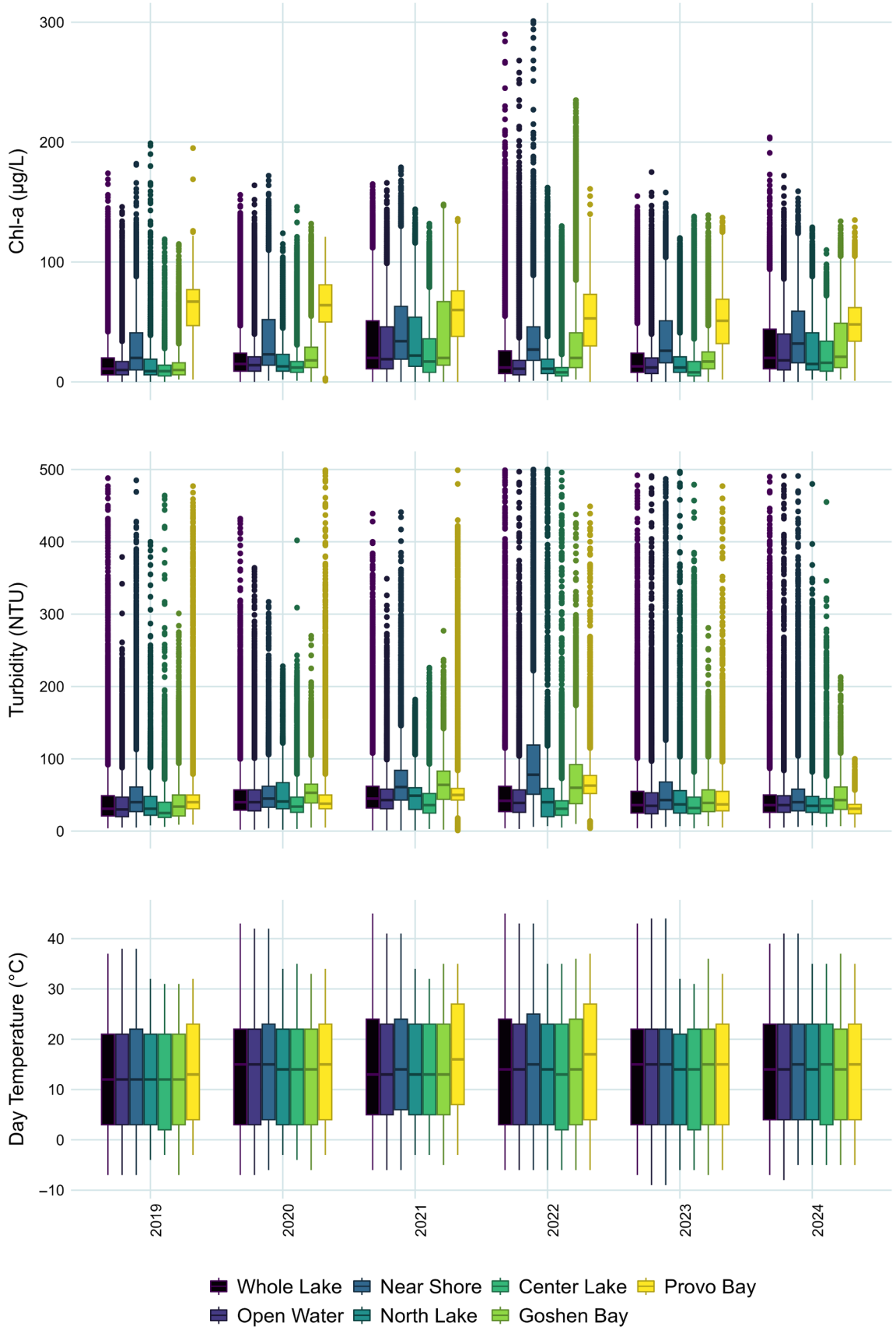

Figure 10. Distributions of the chl-a, turbidity, and temperature values by year and dataset. The box ends represent the 25th and 75th percentiles, with the center line the 50th percentile or median value. The whiskers represent 1.5 × IQR (interquartile range), with the dots representing outliers, or values higher or lower than 1.5 × IQR.

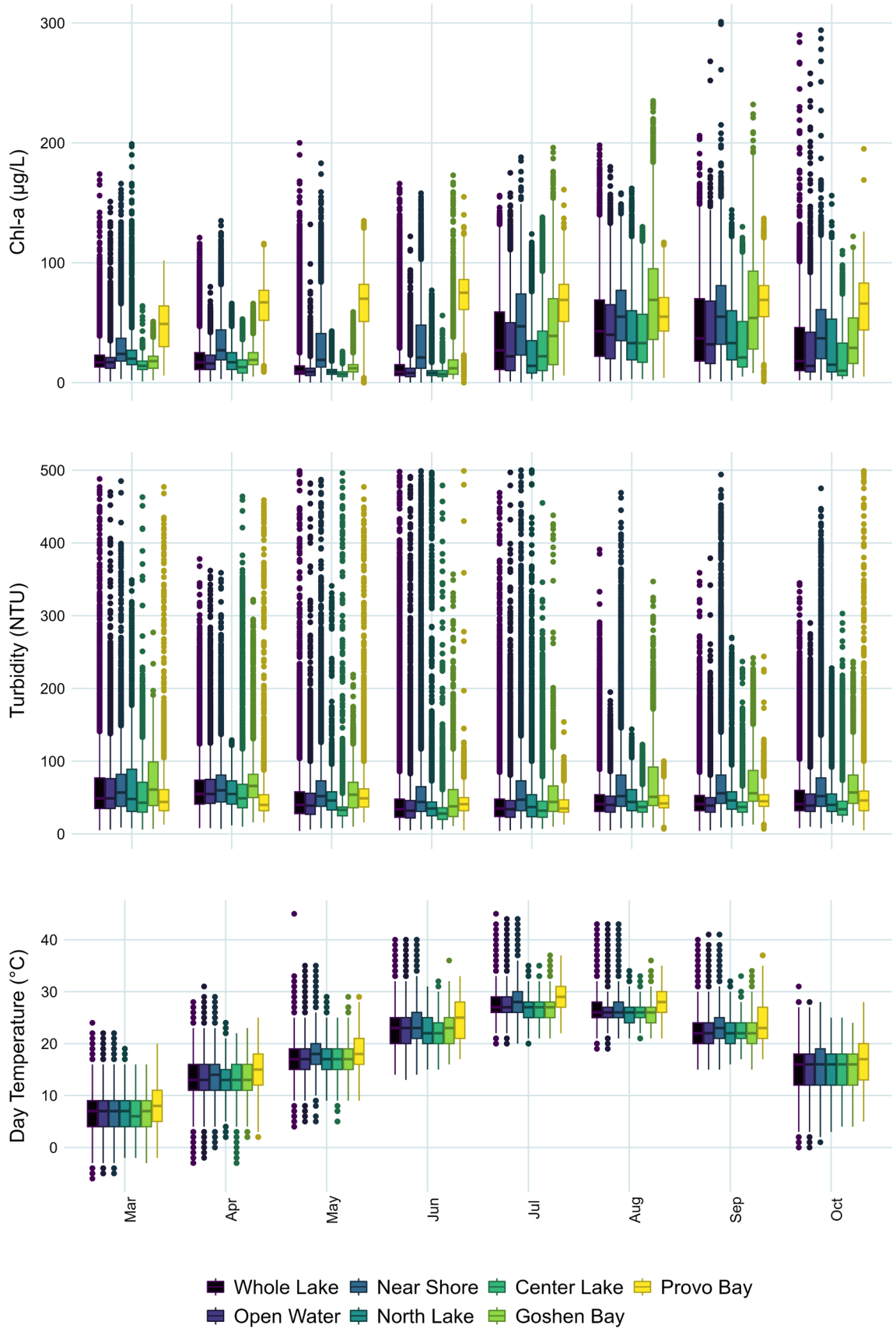

Figure 11. Distributions of the chl-a, turbidity, and temperature values by month and dataset. The box ends represent the 25th and 75th percentiles, with the center line the 50th percentile or median value. The whiskers represent 1.5 × IQR (interquartile range), with the dots representing outliers, or values higher or lower than 1.5 × IQR.

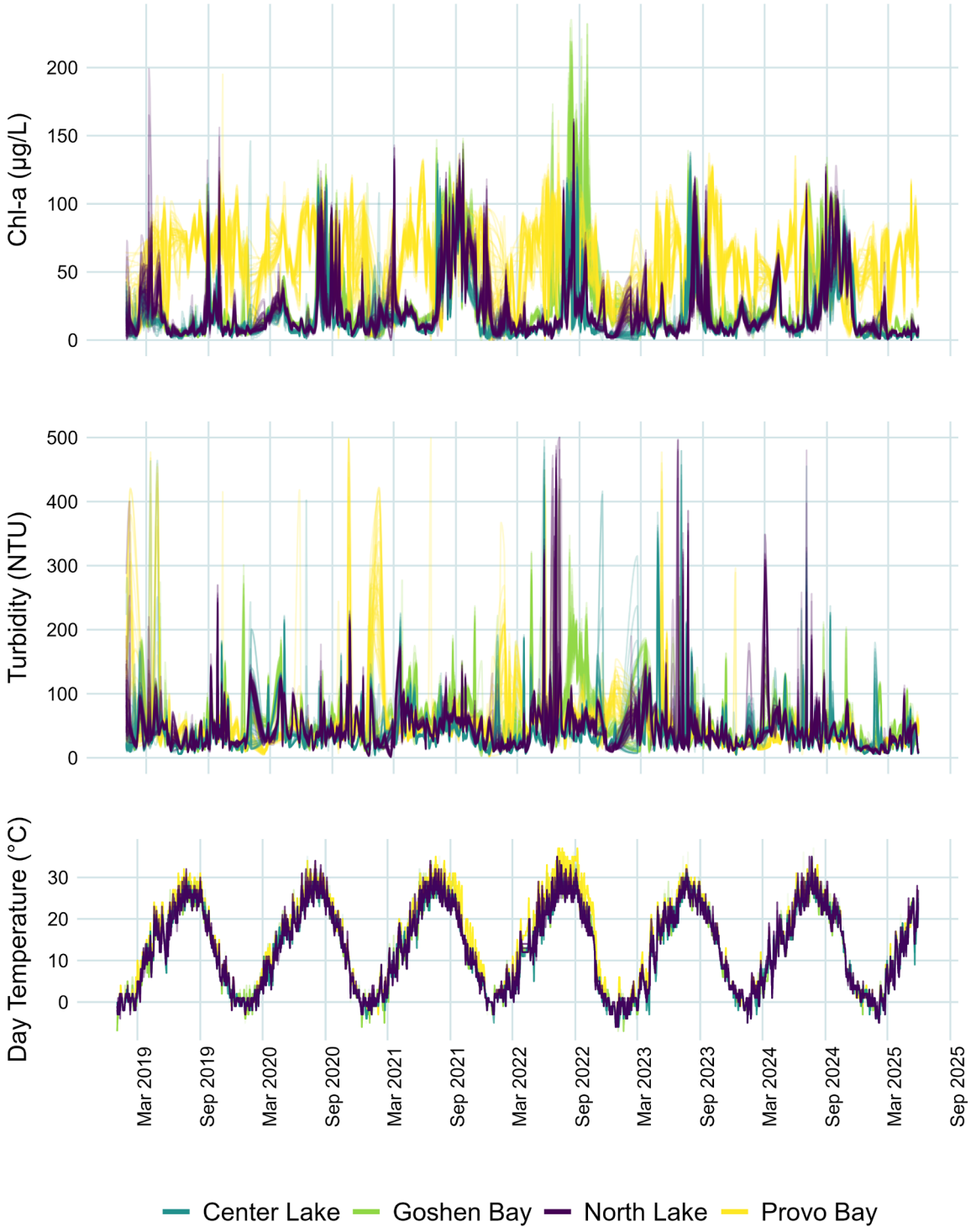

Figure 12. Temporal (time) plots based on the interpolated boxes dataset to show the utility and validity of the daily interpolated data.

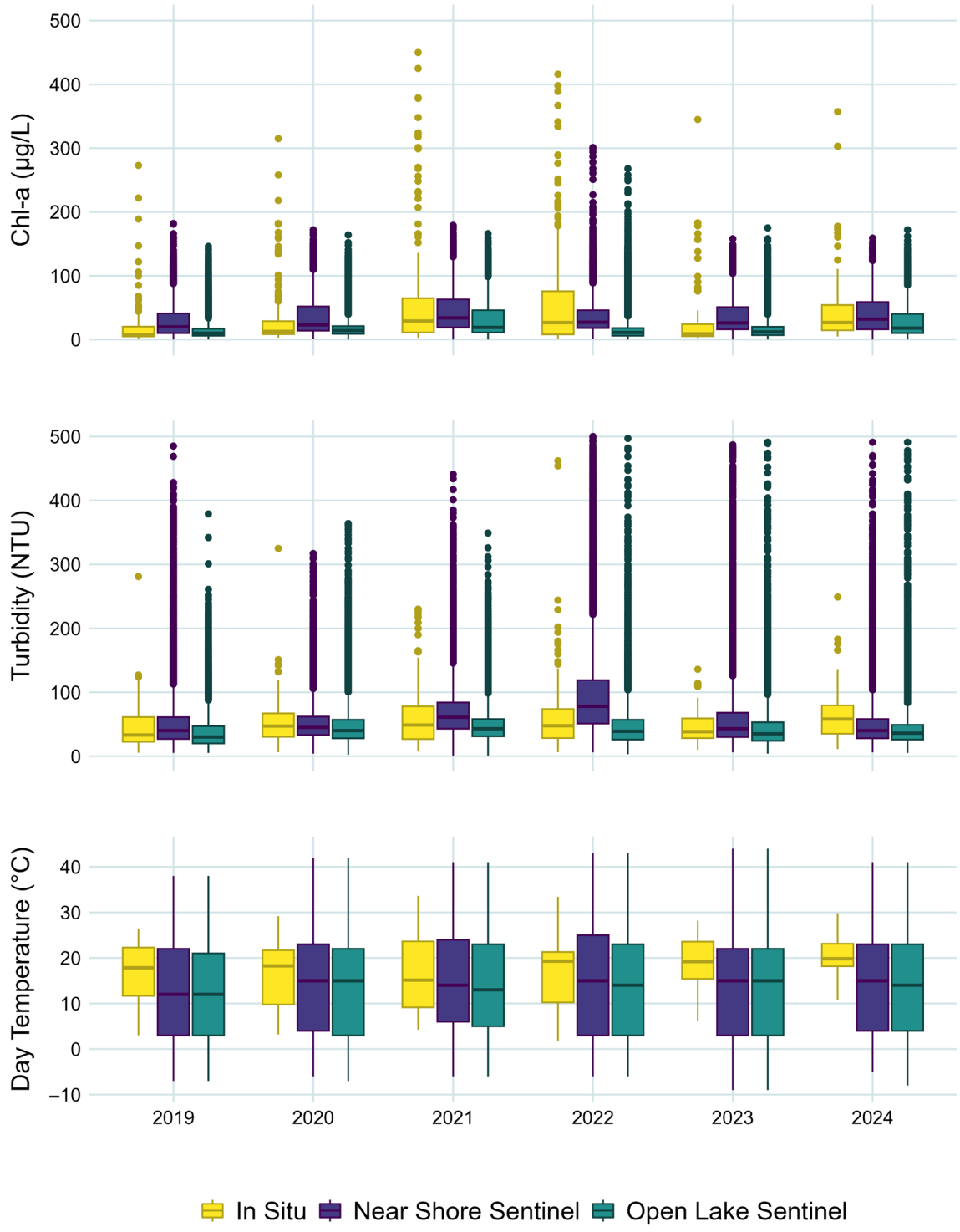

Figure 13. A comparison of in situ data from the Department of Environmental Quality with our estimates of chl-a, turbidity, and temperature. The distributions match quite well but have some differences, showing that the satellite data are valid and provide additional context about Utah Lake water quality missing from the spatially and temporally limited in situ samples. The box ends represent the 25th and 75th percentiles, with the center line the 50th percentile or median value. The whiskers represent 1.5 × IQR (interquartile range), with the dots representing outliers, or values higher or lower than 1.5 × IQR. We excluded four in situ chl-a measurements above 500 µg/L from the plot for readability.

Supplementary Materials: The data described in this manuscript can be downloaded at: https://www.mdpi.com/article/10.3390/data10080128/s1; and on https://doi.org/10.5281/zenodo.15677448, accessed on 6 August 2025.

Author Contributions: Conceptualization, G.P.W. and K.B.T.; methodology, G.P.W., K.B.T., and A.C.C.; formal analysis, G.P.W. and K.B.T.; data curation, K.B.T. and A.C.C.; writing—original draft preparation, K.B.T.; writing—review and editing, G.P.W. and A.C.C.; visualization, K.B.T. and A.C.C.; supervision, G.P.W.; project administration, G.P.W.; funding acquisition, G.P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Utah NASA Space Grant Consortium.

Data Availability Statement: The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Strong, A.E. Remote sensing of algal blooms by aircraft and satellite in Lake Erie and Utah Lake. *Remote Sens. Environ.* **1974**, *3*, 99–107. [CrossRef]
2. Maciel, D.A.; Pahlevan, N.; Barbosa, C.C.F.; de Novo, E.M.L.d.M.; Paulino, R.S.; Martins, V.S.; Vermote, E.; Crawford, C.J. Validity of the Landsat surface reflectance archive for aquatic science: Implications for cloud-based analysis. *Limnol. Oceanogr. Lett.* **2023**, *8*, 850–858. [CrossRef]
3. Toming, K.; Kutser, T.; Laas, A.; Sepp, M.; Paavel, B.; Nõges, T. First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery. *Remote Sens.* **2016**, *8*, 640. [CrossRef]
4. Taggart, J.B.; Ryan, R.L.; Williams, G.P.; Miller, A.W.; Valek, R.A.; Tanner, K.B.; Cardall, A.C. Historical Phosphorus Mass and Concentrations in Utah Lake: A Case Study with Implications for Nutrient Load Management in a Sorption-Dominated Shallow Lake. *Water* **2024**, *16*, 933. [CrossRef]
5. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. *Sensors* **2016**, *16*, 1298. [CrossRef] [PubMed]
6. Nelson, S.A.C.; Soranno, P.A.; Cheruvelil, K.S.; Batzli, S.A.; Skole, D.L. Regional assessment of lake water clarity using satellite remote sensing. *J. Limnol.* **2003**, *62*, 27. [CrossRef]
7. Kutser, T. Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing. *Limnol. Oceanogr.* **2004**, *49*, 2179–2189. [CrossRef]
8. Ogashawara, I. Determination of Phycocyanin from Space—A Bibliometric Analysis. *Remote Sens.* **2020**, *12*, 567. [CrossRef]
9. Olmanson, L.G.; Bauer, M.E.; Brezonik, P.L. A 20-year Landsat water clarity census of Minnesota's 10,000 lakes. *Remote Sens. Environ.* **2008**, *112*, 4086–4097. [CrossRef]
10. Shi, K.; Zhang, Y.; Qin, B.; Zhou, B. Remote sensing of cyanobacterial blooms in inland waters: Present knowledge and future challenges. *Sci. Bull.* **2019**, *64*, 1540–1556. [CrossRef]
11. Hadjimitsis, D.G.; Clayton, C. Assessment of temporal variations of water quality in inland water bodies using atmospheric corrected satellite remotely sensed image data. *Environ. Monit. Assess.* **2009**, *159*, 281–292. [CrossRef]
12. Hansen, C. Google Earth Engine as a Platform for Making Remote Sensing of Water Resources a Reality for Monitoring Inland Waters. In Proceedings of the World Environmental and Water Resources Congress 2015, Austin, TX, USA, 17–21 May 2015.
13. Sogandares, F.M.; Fry, E.S. Absorption spectrum (340–640 nm) of pure water. I. Photothermal measurements. *Appl. Opt.* **1997**, *36*, 8699–8709. [CrossRef]
14. Matthews, M.W. A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *Int. J. Remote Sens.* **2011**, *32*, 6855–6899. [CrossRef]
15. Tanner, K.B.; Cardall, A.C.; Williams, G.P. A Spatial Long-Term Trend Analysis of Estimated Chlorophyll-a Concentrations in Utah Lake Using Earth Observation Data. *Remote Sens.* **2022**, *14*, 3664. [CrossRef]
16. Hansen, C.H.; Williams, G.P. Evaluating Remote Sensing Model Specification Methods for Estimating Water Quality in Optically Diverse Lakes Throughout the Growing Season. *Hydrology* **2018**, *5*, 62. [CrossRef]
17. Pahlevan, N.; Sarkar, S.; Franz, B.A.; Balasubramanian, S.V.; He, J. Sentinel-2 MultiSpectral Instrument (MSI) data processing for aquatic science applications: Demonstrations and validations. *Remote Sens. Environ.* **2017**, *201*, 47–56. [CrossRef]
18. Cardall, A.; Tanner, K.B.; Williams, G.P. Google Earth Engine Tools for Long-Term Spatiotemporal Monitoring of Chlorophyll-a Concentrations. *Open Water J.* **2021**, *7*, 4.

19. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. [CrossRef]
20. Mishra, S.; Mishra, D.R. Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sens. Environ.* **2012**, *117*, 394–406. [CrossRef]
21. Cardall, A.C.; Hales, R.C.; Tanner, K.B.; Williams, G.P.; Markert, K.N. LASSO (L1) Regularization for Development of Sparse Remote-Sensing Models with Applications in Optically Complex Waters Using GEE Tools. *Remote Sens.* **2023**, *15*, 1670. [CrossRef]
22. PSOMAS; SWAC. Utah Lake TMDL: Pollutant Loading Assessment & Designated Beneficial Use Impairment Assessment. In *Prepared for the Utah Division of Water Quality*; PSOMAS: Salt Lake City, UT, USA; SWAC: Salt Lake City, UT, USA, 2007; pp. 1–88.
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
24. Cruz-Retana, A.; Becerril-Piña, R.; Fonseca, C.R.; Gómez-Albores, M.A.; Gaytán-Aguilar, S.; Hernández-Téllez, M.; Mastachi-Loza, C.A. Assessment of Regression Models for Surface Water Quality Modeling via Remote Sensing of a Water Body in the Mexican Highlands. *Water* **2023**, *15*, 3828. [CrossRef]
25. Nahorniak, J.S.; Abbott, M.R.; Letelier, R.M.; Scott Pegau, W. Analysis of a Method to Estimate Chlorophyll-a Concentration from Irradiance Measurements at Varying Depths. *J. Atmos. Ocean. Technol.* **2001**, *18*, 2063–2073. [CrossRef]
26. Hansen, C.H.; Williams, G.P.; Adjei, Z.; Barlow, A.; Nelson, E.J.; Miller, A.W. Reservoir water quality monitoring using remote sensing with seasonal models: Case study of five central-Utah reservoirs. *Lake Reserv. Manag.* **2015**, *31*, 225–240. [CrossRef]
27. Lazhu; Yang, K.; Qin, J.; Hou, J.; Lei, Y.; Wang, J.; Huang, A.; Chen, Y.; Ding, B.; Li, X. A Strict Validation of MODIS Lake Surface Water Temperature on the Tibetan Plateau. *Remote Sens.* **2022**, *14*, 5454. [CrossRef]
28. Pour, H.K.; Duguay, C.R.; Solberg, R.; Rudjord, Ø. Impact of satellite-based lake surface observations on the initial state of HIRLAM. Part I: Evaluation of remotely-sensed lake surface water temperature observations. *Tellus A Dyn. Meteorol. Oceanogr.* **2014**, *66*, 21534. [CrossRef]
29. Tavares, M.H.; Cunha, A.H.F.; Motta-Marques, D.; Ruhoff, A.L.; Cavalcanti, J.R.; Fragoso, C.R., Jr.; Martín Bravo, J.; Munar, A.M.; Fan, F.M.; Rodrigues, L.H.R. Comparison of methods to estimate lake-surface-water temperature using Landsat 7 ETM+ and MODIS imagery: Case study of a large shallow subtropical lake in southern Brazil. *Water* **2019**, *11*, 168. [CrossRef]
30. Chavula, G.; Brezonik, P.; Thenkabail, P.; Johnson, T.; Bauer, M. Estimating the surface temperature of Lake Malawi using AVHRR and MODIS satellite imagery. *Phys. Chem. Earth* **2009**, *34*, 749–754. [CrossRef]
31. Hook, S.J.; Vaughan, R.G.; Tonooka, H.; Schladow, S.G. Absolute radiometric in-flight validation of mid infrared and thermal infrared data from ASTER and MODIS on the Terra spacecraft using the Lake Tahoe, CA/NV, USA, automated validation site. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1798–1807. [CrossRef]
32. Pareeth, S.; Salmaso, N.; Adrian, R.; Neteler, M. Homogenised daily lake surface water temperature data generated from multiple satellite sensors: A long-term case study of a large sub-Alpine lake. *Sci. Rep.* **2016**, *6*, 31251. [CrossRef]
33. Crosman, E.T.; Horel, J.D. MODIS-derived surface temperature of the Great Salt Lake. *Remote Sens. Environ.* **2009**, *113*, 73–81. [CrossRef]
34. Liu, G.; Ou, W.; Zhang, Y.; Wu, T.; Zhu, G.; Shi, K.; Qin, B. Validating and mapping surface water temperatures in Lake Taihu: Results from MODIS land surface temperature products. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1230–1244. [CrossRef]
35. Zanazzi, A.; Wang, W.; Peterson, H.; Emerman, S.H. Using Stable Isotopes to Determine the Water Balance of Utah Lake (Utah, USA). *Hydrology* **2020**, *7*, 88. [CrossRef]
36. Pelleg, D.; Moore, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, USA, 29 June–2 July 2000.
37. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]