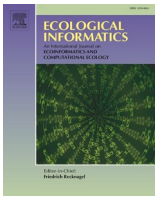
Ecological Informatics 75 (2023) 102033

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Integrated workflow for interpretation of satellite imageries using machine learning to assess and monitor algal blooms in Utah Lake, USA

Robert Davis [a], Palash Panja [a,b,*], John McLennan [a,b]

[a] Department of Chemical Engineering, University of Utah, 50 S. Central Campus Dr., Room 3290 MEB, Salt Lake City, UT 84112, USA
[b] Energy & Geoscience Institute, University of Utah, 423 Wakara Way, Suite 300, Salt Lake City, UT 84108, USA

ARTICLE INFO

Keywords:
Harmful algal blooms (HABs)
Satellite imagery
Atmospheric correction
Feature importance
Machine learning algorithms
Utah lake

ABSTRACT

An integrated workflow is developed to estimate the spatial distribution of harmful algal blooms, especially cyanobacteria concentrations in inland water bodies. The methodology comprises satellite data extraction and preprocessing for atmospheric and water surface corrections, identifying feature importance, in-situ sample collection, training and testing of machine learning algorithms, and prediction. Six input bands are selected using feature importance algorithms from 12 original bands of Sentinel-2 satellite imagery. In-situ sample data that are synchronous with Sentinel-2 image capture time were obtained from a public database. These models are evaluated and compared using spider plots of different error calculations. The workflow developed in this study and the predicted spatial concentration of harmful algal blooms across the lake can be used to improve warning and advisory systems for the public and avoid exposure. The incorporation of other parameters such as water temperature, nutrient concentrations, and surface wind speed could improve the machine-learning models.

1. Introduction

Harmful algal blooms (HABs) can produce toxins that pose a risk to aquatic and terrestrial wildlife, domesticated animals, as well as human health. These blooms can negatively affect local economies by limiting recreational or commercial use of affected waterbodies, and can severely damage the local ecosystem due to the depletion of dissolved oxygen (Brooks et al., 2016). At Utah Lake, in north-central Utah, annual blooms are common. The Utah Department of Environmental Quality regularly assesses the lake for hazardous conditions, often resulting in warnings to the public or even temporary closure (Utah Department of Natural Resources, 2016). In-situ sampling and data collection are critical for monitoring the condition of this inland water body. Monitoring cyanobacteria typically involves sampling water and analyzing it for cyanobacteria and their toxins. It is both labor and time-intensive to collect and analyze bloom samples and even aggressive sampling can only provide results at a select number of point locations. Considering the size of a lake, such as Utah Lake, samples from a restricted number of locations may not be representative of the conditions of the entire water body. There is an opportunity to leverage publicly available remotely sensed data that is collected over a region at regular time intervals. In recent years, there has been an increasing use of remote sensing technologies, such as satellites and drones, to monitor cyanobacteria. These technologies can provide a more cost-effective and efficient way to detect and monitor HABs in larger bodies of water. Machine learning methods can use these data to develop a larger spatiotemporal map of bloom events and potentially aid in predicting their occurrence.

Satellites currently producing remote sensing data (relevant to Utah Lake) include Landsat 8, Sentinel 2A/B, and Sentinel 3A/B. These satellites have differences in the spatial and spectral resolution of onboard instrumentation as well as differences in global coverage and revisit times. The Landsat 8 and Sentinel 2 satellites are focused on land and coastal regions and acquire data in 9- and 13- spectral bands in the visible, near-infrared, and short-wave infrared spectrums. They capture imagery at medium spatial resolutions of 30 m and 10–60 m and have a revisit time of 16 days and 5 days, respectively (NASA Landsat Science, 2013; The European Space Agency, 2015). Sentinel 3 was optimized for detection over the open ocean and coastal zones and captures data over 21 spectral bands in the visible to near-infrared region. This satellite has a revisit time of less than 2 days but collects imagery at a significantly

Abbreviation: ANN, Artificial Neural Networks; CatBoost, Category Boost; DEQ, Department Of Environmental Quality; HAB, Harmful Algal Bloom; MAE, Mean Absolute Error; MAPE, Mean Absolute Percent Error; ML, Machine Learning; MVR, Multivariate Linear Regression; NRMSE, Normalized Root Mean Squared Error; ReLU, Rectified Linear Unit; RF, Random Forest; SVR, Support Vector Regression; XGBoost, Extreme Gradient Boost.

* Corresponding author at: Energy & Geoscience Institute, University of Utah, 423 Wakara Way, Suite 300, Salt Lake City, UT 84108, USA.
E-mail addresses: u1312801@utah.edu (R. Davis), ppanja@egi.utah.edu (P. Panja), jmclennan@egi.utah.edu (J. McLennan).

https://doi.org/10.1016/j.ecoinf.2023.102033
Received 22 July 2022; Received in revised form 18 February 2023; Accepted 19 February 2023
Available online 24 February 2023

coarser spatial resolution of 300 m over land and coastal regions and 1.2 km over the open ocean (Earth Observation Portal, 2021). Remotely sensed products are typically available to users as a 'top of atmosphere' data package and need to be corrected for atmospheric effects to better represent conditions in the upper water column. Up to 90–98% of the signal obtained by the instrument is due to contributions of the water surface and atmosphere. Only the remaining 2–10% of the acquired data is the portion associated with the optically active water constituents of interest. Therefore, applying appropriate atmospheric correction measures is paramount (Dörnhöfer and Oppelt, 2016). Several atmospheric correction algorithms are sensor and/or site-specific. Many studies resolve uncertainty in atmospheric correction by collecting ground-based spectral readings and comparing them with remotely sensed data (Bresciani et al., 2018; Dörnhöfer and Oppelt, 2016; Watanabe et al., 2015).

There are additional optical complexities in Case II water bodies such as shallow coastal or inland waters. These complications can be due to potentially significant effects of suspended solids, dissolved organic matter, and bottom reflectance on measured water-leaving reflectance. This can greatly interfere with deriving a signal representative of the constituent(s) of concern (Pettersson and Pozdnyakov, 2013). In these optically complex waters, machine learning algorithms may be employed to derive complicated relationships in the underlying data without needing to explicitly state those relationships.

To develop a model for quantifying the bloom extent and concentrations most studies have either followed an empirical or a semi-analytical approach. Empirical methods relate remotely-sensed spectral information to coincident or near-coincident in-situ measurements through multivariate regression, often with visible and near-infrared bands and/or band ratios as input variables. The regression is then applied to all pixels across the body of water to develop a spatial depiction of the concentrations of interest (Potes et al., 2018; Vargas-Lopez et al., 2021). Semi-analytical methods require detailed spectral signatures of the sensor-affecting water constituents and are based on solving the radiative transfer equation using either look-up tables of spectral databases or through inversion methods (Carder et al., 1999; Garver and Siegel, 1997).

In this study, we develop a workflow for predicting cyanobacteria concentrations across Utah Lake using an empirical approach with in-situ composite cyanobacteria concentration data and Sentinel-2 remote sensing data. The *sen2cor* atmospheric correction algorithm is used to obtain surface reflectances and feature importance analysis is conducted to reduce the number of input features. Results using artificial neural networks (ANN), random forest regression (RF), multivariate regression (MVR), and support vector regression (SVR) models are compared and suggestions are made for model improvements; critical factors for this effort are also identified.

2. Method/procedure

2.1. Study area

Utah Lake is located in north-central Utah in the Orem-Provo metropolitan area with a population of over 500,000. The lake is 39 km long and 21 km wide at its widest points and has a surface area of approximately 375 km^2. It is a shallow lake with relatively flat contours with an average depth of around 2.7 m and its deepest point is approximately 4.5 m. The lake bottom consists of soft sediments with an unclear boundary between water and sediment and has frequent mixing from waves in shallow regions. The lake is eutrophic, turbid, and alkaline, largely due to the naturally saline nature of the soil in the watershed and biological activity (Brimhall and Merritt, 1981). Due to these conditions and high evaporation rates, calcium carbonate precipitation is common and is especially prevalent in the summer months when the lake can take on a milky appearance (PSOMAS, 2007). Major inflows to the lake include the Spanish Fork River, Provo River, American Fork River, wastewater effluents, and agricultural runoff. Groundwater flux through seeps and springs may also be a substantial source to the lake (Zanazzi et al., 2020). The Jordan River is the sole outlet from the lake with the flow controlled by pumps based on downstream agricultural and irrigation demands, terminating at the Great Salt Lake. Beneficial use of the lake includes secondary contact recreation (activities like boating, wading, fishing, etc.), habitat for warm water game fish, waterfowl, shorebirds, and other water-oriented wildlife and associated food chains, and agricultural water supply (Utah Office of Administrative Rules, 2006). Boating, fishing, bird watching, and other water activities are common, with multiple marinas and beaches located across the lake. Approximately 200,000 people visit annually with over half visiting in the summer months of June, July, and August (Utah State Parks Office, 2022). Areas of interest include Utah Lake State Park, which provides marina access for boating and offers 30 RV campsites, areas for picnicking, and an ice rink in the winter; hot springs located near Lincoln Beach and Saratoga Springs; Lincoln beach on the south end of Utah Lake; and American Fork Marina. Key locations across the lake are shown in Fig. 1a.

2.2. Data source, extraction

The Utah Department of Environmental Quality (DEQ) seasonally (May–October) samples Utah Lake at select locations. The sampling targets indicators of harmful algal bloom activity measured by cyanobacteria cell density (cells/ml), anatoxin (μg/l), and microcystin (μg/l) concentrations at the surface (as scum) and/or as an elbow-depth composite sample. Each sample location is recorded as a GPS location. All samples were sent to a third-party independent laboratory for analysis. The analysis included cell concentrations for a wide array of algal species and was further classified by algae division. The total cyanobacteria cell density (cells/ml) was determined by summing the measured cells/ml for all species in the cyanobacteria division of a given sample. The available results from sampling in 2016–2019 were used in the analyses here and focused on cyanobacteria cell density since that data set was the largest and most consistent. The composite results were chosen rather than the surface samples because it was inferred that they would be more representative of the average water conditions at each location. Since the data collected by Sentinel-2 have a resolution of 10, 20, or 60 m, remote sensing imagery would be more representative of spatial average conditions rather than peak data. The surface scum data represents the worst-case condition at each sample location. That information is important in issuing public guidelines, but less useful for comparison with satellite data. Utah Lake is a relatively shallow, turbid, and transient lake, which makes it more optically complex than clear-lake or open ocean imagery (Odermatt et al., 2012). Furthermore, in shallow water, the increased effects of suspended solids and bottom reflectance increase the complexity of segregating background effects from the in-water constituents of interest, as compared to deeper lake regions. To simplify the model, given the very limited amount of usable in-situ data, samples less than roughly one-meter depth according to a published bathymetric map were excluded from the data set. The remaining data set consisted of 25 data points; the locations of in-situ samples used in this study are included in Fig. 1b.

2.3. Preprocessing

Several factors are considered in selecting a satellite for acquiring remote sensing data for this application, and in general. The frequency of data capture or revisit time of the same surface location, the resolution (i.e., the size of a pixel in an image), and the number of bands is three major factors. Landsat 8, Sentinel 2, and Sentinel 3 satellites provide multi-spectral images with varied resolutions and have different revisit times. Landsat 8 has 11 spectral bands and 15 to 100 m resolution, and a 16-day revisit time. Sentinel 2 has 13 spectral bands and 10 to 60 m resolution, with a 5-day revisit time. Sentinel 3 has 21 spectral

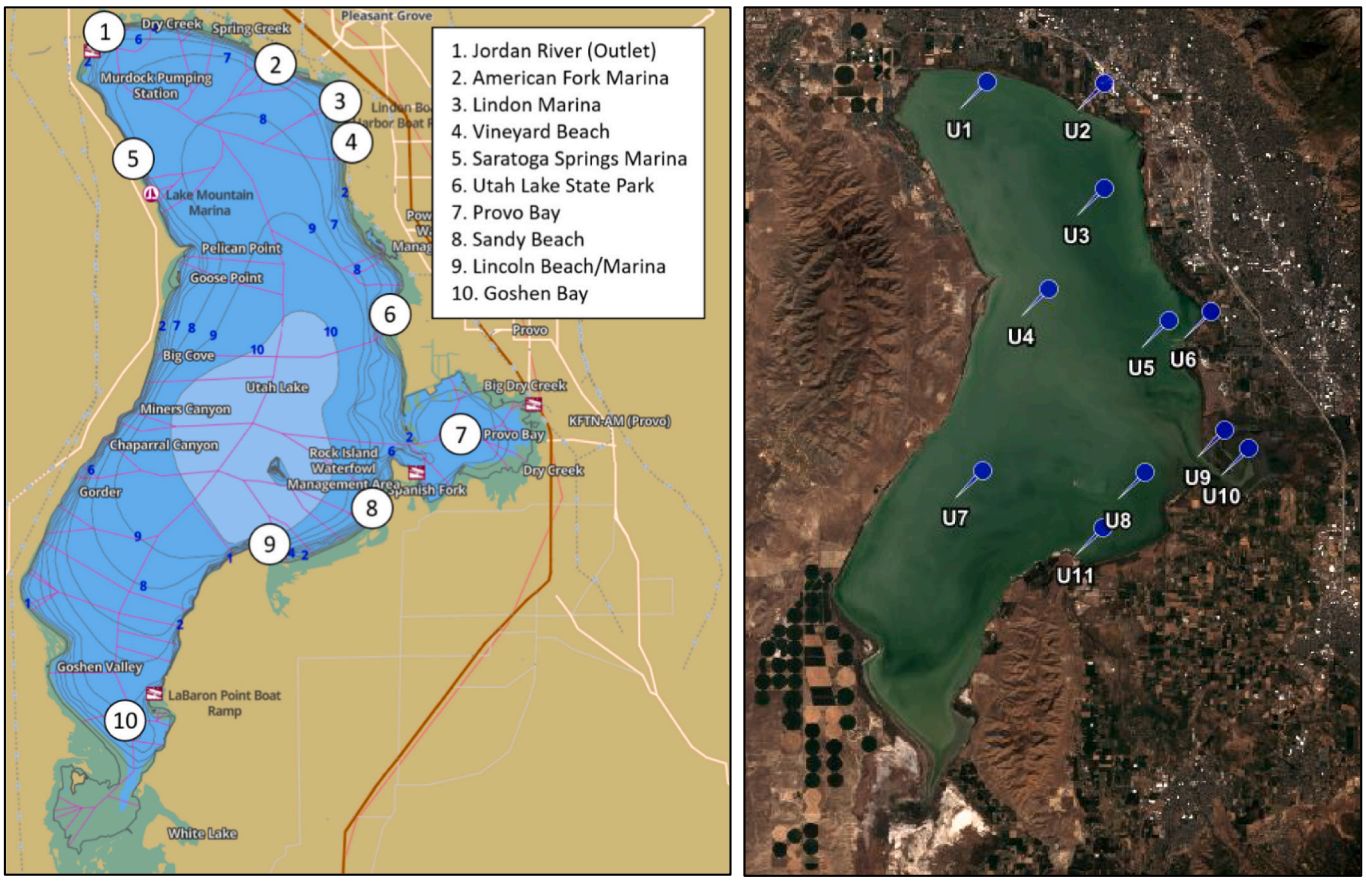


Fig. 1. (a) Key locations around the lake, including recreational points of interest, overlayed on a general bathymetry map, modified from (Utah Lake Fishing Map, 2022), (b) Sample locations used in this study. The number in parenthesis indicates the total samples at each location: U1 (1), U2 (1), U3 (5), U4 (1), U5 (4), U6 (1), U7 (2), U8 (4), U9 (3), U10 (2), U11 (2).

bands and 300 to 1000 m resolution, with a 1- to 2-day revisit time. Additionally, data accessibility is a major factor. Sentinel 2 was selected as the source for remotely sensed data because it has a similar scale of spectral and spatial resolution to Landsat 8, but with three times greater revisit frequency (5-day revisit time versus 16-day). Although it has the shortest revisit time of the satellites considered, the Sentinel 3 satellite was not used due to its coarser spatial resolution. Sentinel 2 Level-1C imagery was obtained from the Copernicus Open Access Hub for dates coincident with composite cyanobacteria sampling by the DEQ. The lake is well-mixed vertically, but may be poorly mixed horizontally and has a residence time of approximately 6 months depending upon agricultural water demands (Zanazzi et al., 2020). Windy conditions are common on the lake, which induces vertical mixing and spatial movement of blooms across the lake, limiting the number of usable images to dates coincident with in-situ sampling (Bresciani et al., 2018; Stumpf et al., 2012). Data with even a one-day difference between sampling and image capture were shown to have a significantly lower correlation and were not a good representation of water conditions at the time of sampling (Hansen et al., 2017). The coincident imagery was manually screened for cloud and cloud shadow coverage, sun glint, and milky conditions indicative of high levels of calcium carbonate and other precipitates. The 10 remaining images with 25 sample points were atmospherically corrected using the *sen2cor* algorithm, which has been shown to provide reliable estimates for eutrophic waters (Pereira-Sandoval et al., 2019). Atmospheric correction is used to remove aerosol or other molecular scattering in the atmosphere such as from thin-cirrus clouds or haze, which is common in Utah Valley, which contains Utah Lake (Warren et al., 2019). The Level 2 images were resampled from 10, 20, and 60 m resolutions to 60 m for spatial uniformity when applying the regression algorithms. The images were then clipped to a shapefile of Utah Lake to eliminate potential noise and scaling issues arising from pixels outside of the lake. Data for 12 bands - including B1-B12 (not including B10, which is eliminated during atmospheric correction) at the sample points for each image were tabulated with the 25 corresponding composite cyanobacteria sampling results.

2.4. Feature importance

With a limited data set and a high number of input bands for each model (12), we looked at the results of four feature importance algorithms to reduce the number of input features to 6. The methods used were the mean decrease in impurity with squared error criteria from a random forest regression model, and permutation importance with squared error objective function and 20 repeats for random forest regression, an XGBoost regression, and CatBoost regression models. The feature importance results are shown in Fig. 2.

The input bands were reduced from 12 to 6 to include bands B1, B2, B3, B4, B5, and B9. Data for these bands were the inputs for each of the four machine-learning algorithms in producing cyanobacteria concentration mappings across the lake.

2.5. Machine learning algorithms

The machine learning algorithms used in this study were artificial neural networks (ANN), random forest (RF) regression, multivariate linear regression (MVR), and support vector regression (SVR). ANN is a

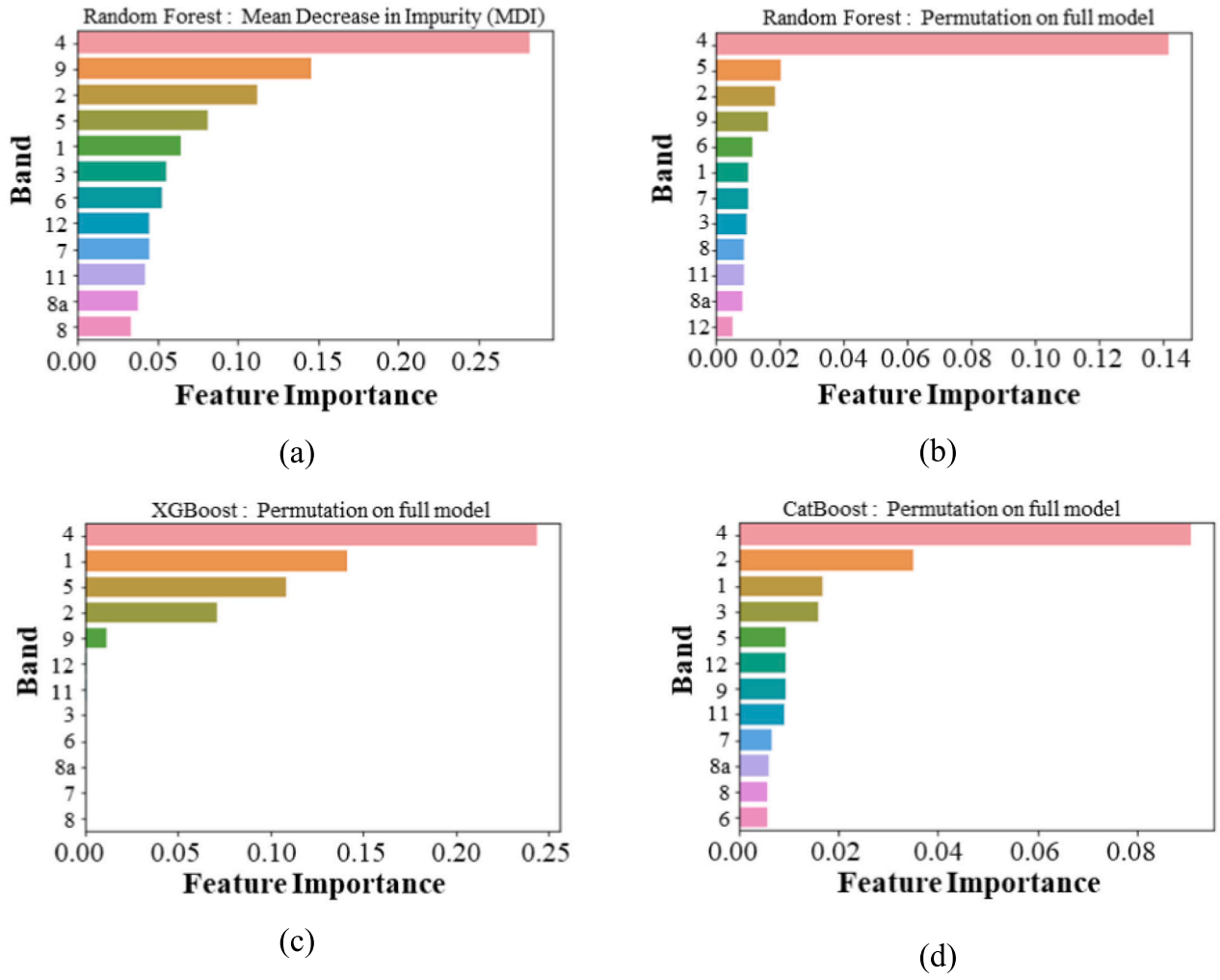


Fig. 2. Results from feature importance analysis based on (a) Mean decrease in the impurity of a random forest regression model; (b) Permutation of a random forest regression model; (c) Permutation of XGBoost regression model; (d) Permutation of CatBoost regression model.

non-linear machine learning model used to uncover complex relationships between inputs and outputs. It is well known that deciding on parameters and functions appropriate for machine learning (ML) algorithms is a challenging task. For example, the number of hidden layers and neurons in artificial neural networks (ANN) is important for proper training of the model to avoid underfitting (use of too few neurons) and overfitting (use of too many neurons). The amount of training time also increases significantly with the number of neurons and hidden layers in ANN (Dinkelbach et al., 2012). Despite the absence of a definitive rule-of-thumb for determining the optimum number of neurons in hidden layers, some formulas and guidance can be acknowledged (Heaton, 2008). Choosing the correct activation function is also critical for data transformation during training. Similarly, the kernel function plays an important role in support vector machine (SVM) modeling. In the case of non-parametric models such as random forest (RF), the size of the dataset, the maximum depth of the individual trees, and the number of random features are relevant. In this regard, sensitivity analysis can be conducted. In this study, the ML parameters and functions are determined by trial and error and experience from previous studies (Panja et al., 2016; Panja et al., 2018; Panja et al., 2022).

The ANN model used in this study consisted of 3 hidden layers of 50, 20, and 5 nodes, respectively, and the rectified linear unit (ReLU) activation function. The model was compiled with the Adam optimizer and mean squared error loss function over 2000 epochs. The RF algorithm is built on decision tree learning with randomly selected splitting features and data subsets also chosen at random; the output predicted value of the regression is the average of the outputs from the trees. RF regression is unable to extrapolate predictions outside the range of its training data, therefore, for this study, the training set always contained the lowest and highest output points. Our RF regression was based on a model with 10 trees and split quality of squared error. The multivariate linear regression model is based on ordinary least squares linear regression and serves as a baseline for comparison with other methods. SVR is another model that can derive non-linear relationships between variables and allows model adjustment by tuning its hyperparameters. The SVR model in this study used the radial-basis function kernel to transform a non-linear problem into a higher-dimensional linear problem and had a regularization parameter of 1.0 and an epsilon value of 0.1.

2.6. Training and testing

The training and test data sets followed a randomized 80/20 split, with the data points of highest and lowest cyanobacteria concentration always included in the training set to minimize extrapolation, a notable limitation in RF regression.

2.7. Workflow

A workflow of the entire methodology is shown in Fig. 3.
Although all steps in the workflow are discussed in the previous

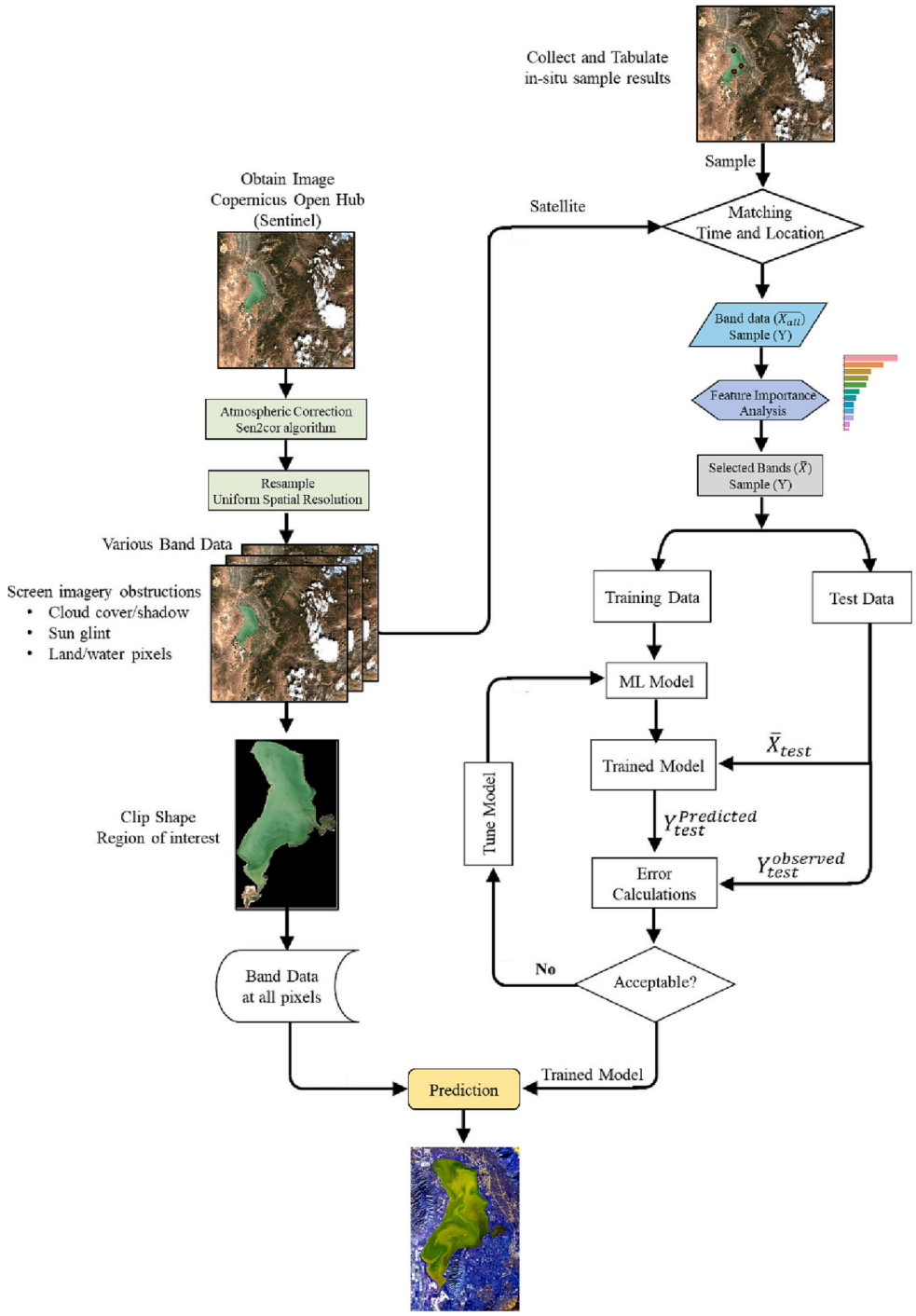


Fig. 3. Workflow followed to produce spatial regression mapping across the lake.

section, they are summarized here. The starting satellite data is typically at top-of-atmosphere conditions; the user needs to identify and apply appropriate atmospheric correction to the image. If the image contains input bands with different resolutions, it may be useful to resample to a uniform spatial resolution for easier regression calculation in python. Otherwise, since the band data will be stored in arrays of different sizes, the user will need to modify the python code to ensure data lines up at each point across the image. One major step is to screen the image to remove pixels with aberrations such as cloud coverage, pixels on the land/water boundary, sun glint, etc. followed by clipping the resulting image to reduce scaling issues that could come from regression results over land pixels. A python coding language library namely *rasterio* is applied to read all input band data.

In-situ sample data are chosen based on acceptable satellite image window and at locations with viable satellite data (i.e. not on land/water boundary, not covered by clouds, etc.). Band data are identified and retrieved at each sample data location. Feature importance analysis is employed to identify the most impactful input features. The acceptable data are split into training and test sets followed by a selection of various models. The models are trained and tuned until achieving acceptable error. The trained models are used to predict output values across the satellite image from the band input data.

3. Results

3.1. Model evaluation

The four regression models were trained and run to predict cyanobacteria concentrations across Utah Lake on an image from September 5, 2018. This was chosen as a representative image since it was a clear day with minimal aberrations needing to be removed from image preprocessing. The results of the predicted versus actual cyanobacteria concentrations for the training and test data sets as well as the relative error for the data within each model are shown in Fig. 4. The test data point numbering corresponds to its physical location as shown in the spatial regression results in Fig. 6.

All models had a very high relative error at test point 1, which may be an artifact of low cyanobacteria concentration (actual concentration 141 cells/ml), which makes even modest overestimates result in a grossly high relative error (e.g. the MVR model estimated 830 cells/ml which gives a relative error of 485%, even though the absolute error was 689 cells/ml). The opposite is true for predicting concentrations on the high end. For example, at test data point 4 with an actual cyanobacteria concentration of 30,294 cells/ml, the RF model estimated 6350 cells/ml, which is a − 79% relative error. If strictly comparing the relative errors, one can conclude that the relative error at point 4 (−79%) is a better result than the relative error at point 1 (485%), even though the absolute error is almost 35 times higher. To understand some of the inherent biases in these error calculation methods, it is beneficial to compare both relative error (or MAPE) and absolute error for the models. This is illustrated in Fig. 5. Spider plots were produced to show error calculations for the training and test data sets as a mean absolute error (MAE), normalized root mean squared error (NRMSE), and mean absolute percent error (MAPE) during the spatial regression from each model, shown in Fig. 5.

The values were normalized between 0 and 1 using the highest value for each error analysis method. As seen in the Spider plots (Fig. 5), the SVR model performed the worst with the highest MAPE, MAE, and NRMSE on the test data set, and had the near-worst performance in the training data set. The SVR model also notably has the narrowest range of predicted values from the data used for training and testing with a range between 10^3 and 10^5 cells/ml, as shown in Fig. 4 g. The RF model performed the second-worst with the second-highest MAPE and MAE values for the test data set. The MVR model has the second-highest values for NRMSE and MAE (nearly identical to RF), but a significantly lower MAPE than the other models. The low MAPE result may be more of an indication that the MVR model did not overestimate the smaller concentration point as much as the other models. This is supported by the relative error results of test point 1 (i.e., MVR: 485% compared to 4679% with the ANN model and ~ 8–9000% for RF and SVR models; see right side of Fig. 4). A similar observation is made for MVR training data set results. The model had relatively low MAPE, but the highest MAE and NRMSE. The higher MAE and NRMSE values indicate that in absolute terms, the MVR model had a higher error in both the test and training data sets than the ANN model. With the lowest MAE and NRMSE results, second-lowest MAPE test data set results, and overall best results in the training data set, it could be argued that the ANN model produced the best regression of the four algorithms. Additionally, a combined error can be assessed more quantitatively by calculating the area of the polygon (triangle in this study) in the spider plot. The areas of ANN, RF, MVR, and SVR during training are 0.16, 0.57, 0.87, and 1.17 respectively. The SVR model has the highest area and the ANN model has the lowest area of coverage on the spider plot during training confirming the remarks discussed earlier on the comparison of different models. Similarly, the areas of ANN, RF, MVR, and SVR during testing are 0.58, 0.89, 0.36, and 1.30 respectively. Therefore, MVR has the best performance and SVR has the worst during testing. This indicates that although ANN had the best performance during training, it didn't perform better than MVR during testing.

It is expected that the results from different models would have discrepancies. In general, every machine learning mode has underlying and different mathematical principles, structures, and assumptions, therefore they have some inductive bias associated with them. Consequently, each model has certain attributes and detractions. The models can also be divided into parametric (ANN, SVM, MVR) and non-parametric (RF) styles. Although the performance of a model varies with data sets, all of the models evaluated were trained with the same dataset in this study. There is no defined way to determine the best model a priori. Understanding the dataset and also the expected outcome is crucial in choosing the best model. It is also debated whether the same model can generate different results using the same dataset for selecting different hyperparameters for ANN and kernel function for SVM and the number of trees for RF (Levy et al., 2015; Lucic et al., 2018; Uddin et al., 2019).

3.2. Prediction

The cyanobacteria mapping from each model is shown in Fig. 6. A histogram of the regression results across the lake was reviewed, and the color scale range for each model was appropriately adjusted to show better contrast across the lake. This helps in visualizing the spatial distributions of algal bloom irrespective of the absolute value. This is also helpful to understand the localized growth area which could be a potential danger zone for water activities.

As observed in the test and training set results, the SVR model had the narrowest concentration range across the lake (0–100,000 cells/ml) with a maximum of 117,859 and 0.08% of lake pixels over the 100,000 range. The RF model had the second narrowest range of results (0–150,000 cells/ml) with a predicted maximum of 241,385 cells/ml and 0.69% of pixels over the 150,000 thresholds. The MVR and ANN models had similar ranges (0–500,000 cells/ml), but the MVR model had a higher maximum at 2,818,419 cells/ml compared to 951,283 cells/ml, and more pixels over the 500,000 upper limit at 1.35% compared to 0.69% in the ANN model.

All models predicted high concentrations in the northwest area of Goshen Bay and the near-shore region just north of State Park Marina (refer to Fig. 1 for location designations). The majority of the above-range data points predicted by the ANN and MVR models resided in these two areas. All models have a relatively moderate increase in concentrations along a belt spanning just west of Provo Bay to the northern portion of Goshen Bay and an area of higher concentration N-NW out of Provo Bay up to State Park Marina. The areas of lowest concentration were in the central, western, and northwest portions of the lake leading to the outlet to the Jordan River. Although the models similarly predicted many general areas of high and low concentrations, the magnitudes of the predicted values varied greatly between models. These differences could make a significant impact on regulatory action or mitigation activities. For example, in the near-shore region just north of State Park Marina, the SVR model would suggest no action is required, while the RF model is bordering the 100,000 cells/ml warning advisory. The ANN and MVR models are both well above the warning advisory, which would recommend the public avoid primary contact and suggest a minimum of weekly sampling by the State (Utah-Department-of-Environmental-Quality, 2020).

3.3. Model uncertainty

The robustness of a model depends on its predictability after training with different sub-sets of the same data set. To assess the uncertainty in the models, they are generally trained several times for statistical analysis of the errors. The four models were run 100 times while varying the data points in the training and test sets; the resulting variation in mean absolute percent error for the training and test data sets is shown in Fig. 7.

Fig. 7 shows the sensitivity of the model based on the data points

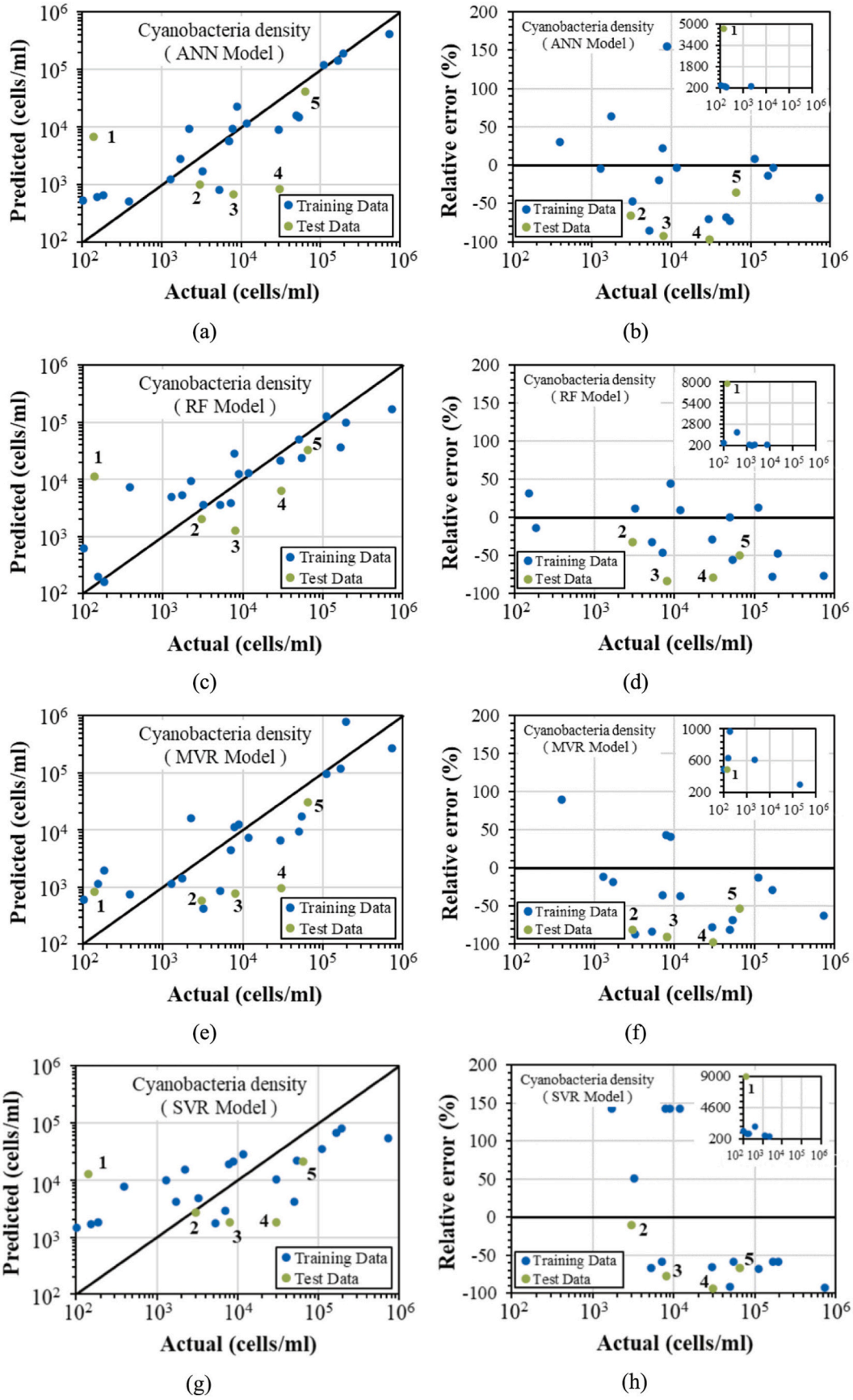


Fig. 4. (left side) Cross plots of actual cyanobacteria density and predicted density of the same from (a) ANN model (c) RF model (e) MVR model (g) SVR model; (right side) relative errors from (b) ANN model (d) RF model (f) MVR model (h) SVR model. In general, the models tended to overestimate cyanobacteria concentrations in lower cell count areas and underestimate them in higher concentration regions.

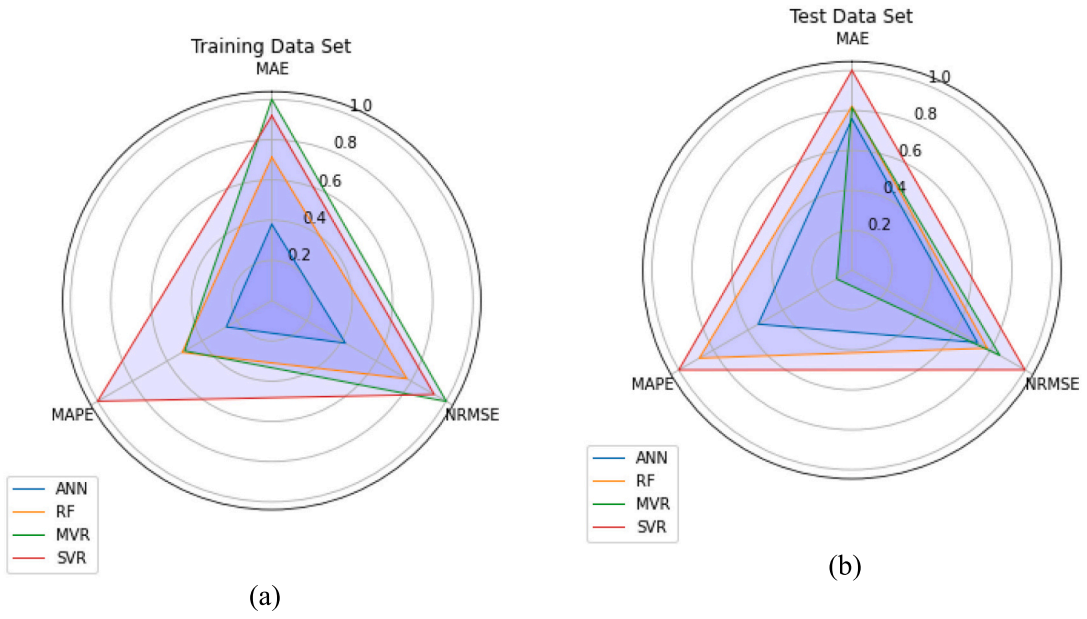


Fig. 5. Spider plot of MAE, MAPE, and NRMSE (a) Training data set (b) Test data set.

selected in the training/test sets. Recall that the highest and lowest values were always included in the training set since random forest regression cannot predict values outside its training data range. The SVR model had the highest median MAPE for both the training and test data sets, as well as the largest first to third quartile spread. The median and spread of the ANN, RF, and MVR models on the test set are similar, although notably the ANN model and RF models had two and six outlier results significantly higher than shown on the graph. Overall, there is a wide-spread in MAPE results on the test data sets across all of the models. This spread shows that the results from each model are highly sensitive to the data points allocated in the training and test data sets. This is not surprising given the small sample size.

4. Discussions

There is a possibility of model overfitting where a model learns the detail and noise in the training data and consequently, it performs poorly on the unseen test set. In most instances, where there are adequate data, this would be avoided by resampling and holding back a validation data set (Deisenroth et al., 2020; James et al., 2013). Techniques such as these improve model accuracy by avoiding overfitting and using the test data set during training. A validation dataset is a subset of training data that is held back to evaluate the performance of a trained model. As a result, the model can be improved by re-training it until an accepted error range is obtained. K-fold cross-validation is a popular resampling technique. This method is similar to the validation dataset except the resampling is conducted k-times by randomly dividing the training data into subsets training and subset testing. Based on the results of k-times training, the performance of the model is estimated before evaluation with the unseen test dataset.

Both methods require that a certain fraction of training data is held back. This means that all training data cannot be used during the training of the model. Considering the small sample size available for this study, these techniques cannot be applied. Alternatively, resampling of the entire dataset to obtain training and test data was performed several times to evaluate the model uncertainty. This provides an idea of the error range in different models.

The primary limitation of this study was the small number of data points available as inputs since the dataset was restricted to coincident points and sampling was not frequent enough to accumulate a sufficient quantity. Additionally, there were days when samples for anatoxin or microcystin concentrations, or surface samples were collected rather than composite cyanobacteria samples. Datapoints were further restricted to areas greater than 1-m. depth to reduce the influence of backscattering and suspended solids in shallow waters from biasing the model. When the models included shallow data they generally produced worse results due to inadequate data to account for the additional optical complexities of those shallow regions. Coincidentally, these more challenging, shallow regions are often the areas of greatest interest as they have the highest potential for human or pet exposure. Sample points were limited to 10 viable locations. To test the robustness of a good model, it would be useful to verify the results with additional sample locations across the lake.

Largely due to windy conditions, fluctuating nutrient levels, and biological activity, the optical signature of the lake varied greatly between images, even on dates when algal blooms were not present. The availability of adequate data for background or “zero” cyanobacteria concentration across various regions of the lake was a challenge in this study and would be a key component for a more accurate model. Obtaining seasonal background data may also be useful as the biological composition of the lake, including that of cyanobacteria, is known to have seasonality (Hansen and Williams, 2018).

In future work we would suggest regular composite cyanobacteria sampling, ideally to match the frequency of the remote sensing instrument of choice. Data from the Sentinel 3 satellite may be considered, due to a short re-visit time of 1–2 days. Using that satellite would increase the useable dataset, although at the expense of spatial resolution. Sentinel 3 also has the spectral resolution to potentially identify the presence of phycocyanin, which can differentiate cyanobacteria from other algae (Ogashawara, 2019). While this study focused on predicting composite cyanobacteria concentrations, developing a model for anatoxin or microcystins may also be considered. Excess levels of any of those components would trigger protective measures. Collecting other ecological parameters such as water temperature, turbidity or Secchi disc depth, or nutrient concentrations could also be helpful to develop a physical model, which could be used independently or in conjunction with empirical machine learning methods.

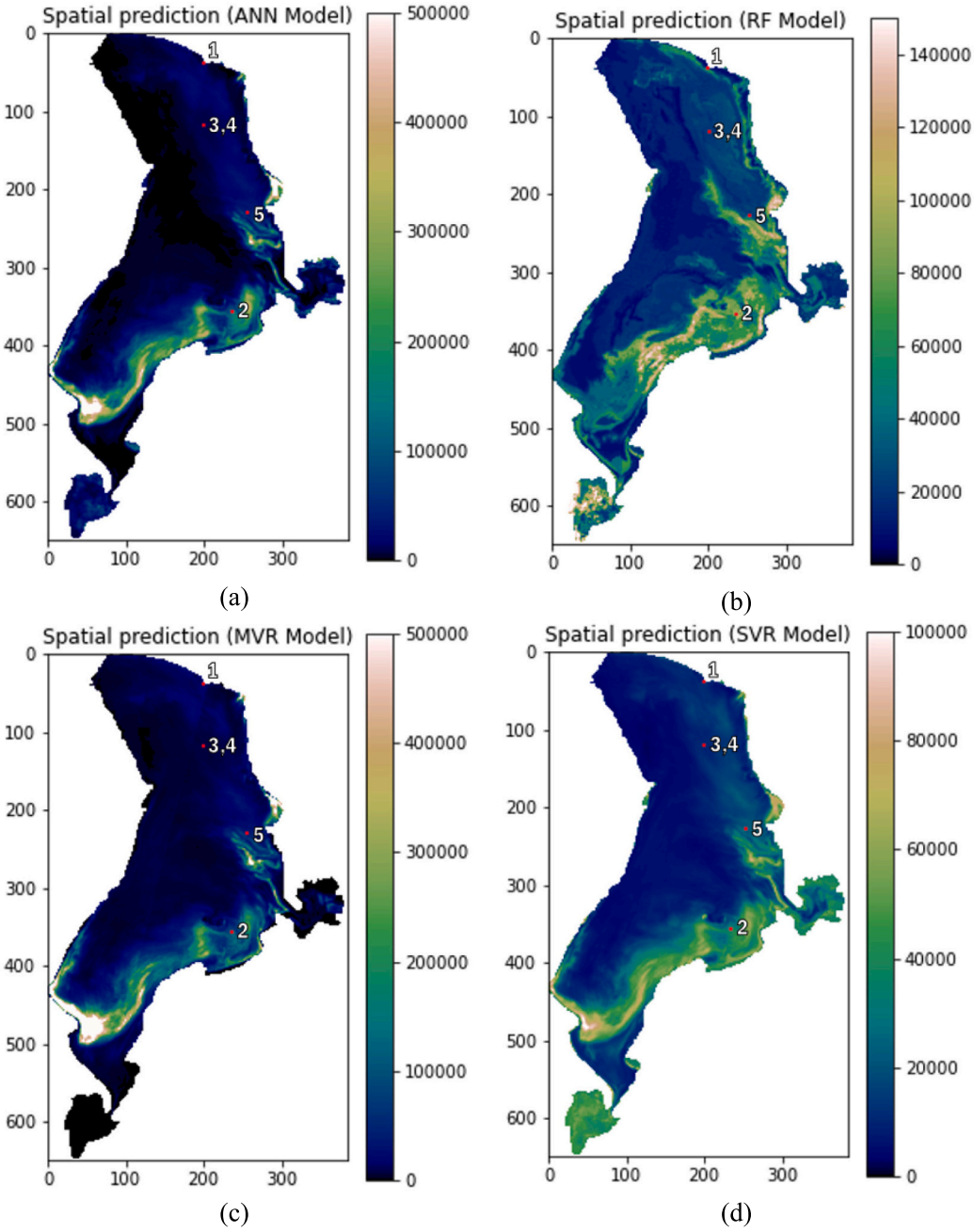


Fig. 6. Mapping cyanobacteria concentrations in the entire lake by interpreting satellite images using trained machine learning models (a) ANN (b) RF (c) MVR (d) SVR. Note that the figures have different scales for better visualization.

5. Conclusions

A workflow enfranchising remote-sensed data acquisition, pre-processing, feature importance assessment, and training machine learning models has been developed to provide predictions of harmful algae concentration in Utah lake and similar vulnerable inland water environments. Atmospheric corrections of the satellite images improved the quality of input features. Four algorithms (three machine learning and one polynomial) were evaluated. These algorithms are artificial neural networks (ANN), random forest (RF) regression, multivariate linear regression (MVR), and support vector regression (SVR). Of these protocols, an artificial neural network (ANN) performed better based on a comparison of the different types of error analysis. The error calculations included the determination of the mean absolute error (MAE), normalized root mean squared error (NRMSE) and mean absolute percent error (MAPE). ML models are superior to traditional modeling for capturing the intricate relationships between the band data from satellite imagery and the algae concentration. Most models were able to effectively identify the locations of high and low concentrations of cyanobacteria concentration across Utah Lake. Spatial cyanobacteria distribution across the lake prescribed by these or similar ML models will be useful for providing advisory warnings of hazardous exposure to humans and pets.

In the future, the integration of supplementary information such as water temperature and nutrient concentrations in the water is anticipated to improve model training and prediction. Frequent collection of in-situ samples from distributed locations of the lake could also enhance the model accuracy, particularly if sampling is synchronous with satellite orbits.

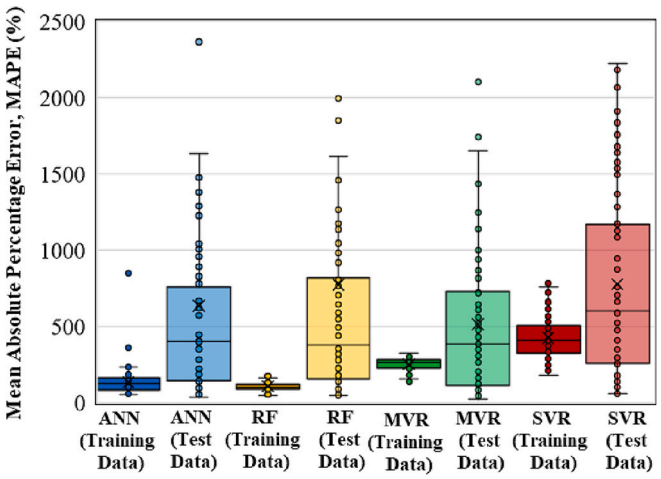


Fig. 7. Uncertainty analysis for different models expressed as Box-Whisker plots of mean absolute percent error.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Funding for this project was provided by the College of Engineering Research Incentive Seed Grant Program 2021, University of Utah, USA. We are grateful for their support. The authors gratefully acknowledge the support of Dr. Kate Fickas from the Utah Department of Environmental Quality (DEQ), Utah, USA for personal consulting and providing data.

References

Bresciani, M., Cazzaniga, I., Austoni, M., Sforzi, T., Buzzi, F., Morabito, G., Giardino, C., 2018. Mapping phytoplankton blooms in deep subalpine lakes from Sentinel-2A and Landsat-8. Hydrobiologia 824 (1), 197–214.

Brimhall, W.H., Merritt, L.B., 1981. Geology of Utah Lake: implications for resource management. Great Basin Natural. Mem. 5, 24–42.

Brooks, B.W., Lazorchak, J.M., Howard, M.D., Johnson, M.V., Morton, S.L., Perkins, D.A., Reavie, E.D., Scott, G.I., Smith, S.A., Steevens, J.A., 2016. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? Environ. Toxicol. Chem. 35 (1), 6–13.

Carder, K.L., Chen, F.R., Lee, Z.P., Hawes, S.K., Kamykowski, D., 1999. Semianalytic Moderate-Resolution Imaging Spectrometer algorithms for chlorophylla and absorption with bio-optical domains based on nitrate-depletion temperatures. J. Geophys. Res. Oceans 104 (C3), 5403–5421.

Deisenroth, M.P., Faisal, A.A., Ong, C.S., 2020. Mathematics for Machine Learning. Cambridge University Press.

Dinkelbach, H.Ü., Vitay, J., Beuth, F., Hamker, F.H., 2012. Comparison of GPU- and CPU-implementations of mean-firing rate neural networks on parallel hardware. Netw. Comput. Neural Syst. 23 (4), 212–236.

Dörnhöfer, K., Oppelt, N., 2016. Remote sensing for lake research and monitoring – Recent advances. Ecol. Indic. 64, 105–122.

Earth Observation Portal, 2021. Copernicus: Sentinel-3 – Global sea/land monitoring mission including altimetry; Sensor complement. https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-3#sensors.

Garver, S.A., Siegel, D.A., 1997. Inherent optical property inversion of ocean color spectra and its biogeochemical interpretation: 1. Time series from the Sargasso Sea. J. Geophys. Res. Oceans 102 (C8), 18607–18625.

Hansen, C., Williams, G., 2018. Evaluating Remote Sensing Model Specification Methods for Estimating Water Quality in Optically Diverse Lakes throughout the Growing Season. Hydrology 5 (4), 62–65.

Hansen, C., Burian, S., Dennison, P., Williams, G., 2017. Spatiotemporal variability of lake water quality in the context of remote sensing models. Remote Sens. 9 (5), 409–413.

Heaton, J., 2008. Introduction to Neural Networks with Java, 2nd ed. Heaton Research, Inc.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer, New York.

Levy, O., Goldberg, Y., Dagan, I., 2015. Improving distributional similarity with lessons learned from word embeddings. Transact. Associat. Computat. Ling. 3, 211–225.

Lucic, M., Kurach, K., Michalski, M., Bousquet, O., Gelly, S., 2018. Are GANs created equal? a large-scale study, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

NASA Landsat Science, 2013. Landsat 8 Bands. https://landsat.gsfc.nasa.gov/satellites/landsat-8/landsat-8-bands/.

Odermatt, D., Gitelson, A., Brando, V.E., Schaepman, M., 2012. Review of constituent retrieval in optically deep and complex waters from satellite imagery. Remote Sens. Environ. 118, 116–126.

Ogashawara, I., 2019. The use of sentinel-3 Imagery to monitor cyanobacterial blooms. Environments 6 (6), 60–65.

Panja, P., Pathak, M., Velasco, R., Deo, M., 2016. Least Square Support Vector Machine: An Emerging Tool for Data Analysis, SPE Low Perm Symposium. Society of Petroleum Engineers, Denver, Colorado, USA, pp. 1–22.

Panja, P., Velasco, R., Pathak, M., Deo, M., 2018. Application of artificial intelligence to forecast hydrocarbon production from shales. Petroleum 4 (1), 75–89.

Panja, P., Jia, W., McPherson, B., 2022. Prediction of well performance in SACROC field using stacked Long Short-Term Memory (LSTM) network, 117670. Expert Syst. Appl. 205, 1–25.

Pereira-Sandoval, M., Ruescas, A., Urrego, P., Ruiz-Verdú, A., Delegido, J., Tenjo, C., Soria-Perpinyà, X., Vicente, E., Soria, J., Moreno, J., 2019. Evaluation of atmospheric correction algorithms over spanish inland waters for Sentinel-2 multi spectral imagery data, 1469. Remote Sens. 11 (12), 1–23.

Pettersson, L.H., Pozdnyakov, D., 2013. Monitoring of Harmful Algal Blooms, 1 ed. Springer, Berlin, Heidelberg.

Potes, M., Rodrigues, G., Penha, A.M., Novais, M.H., Costa, M.J., Salgado, R., Morais, M. M., 2018. Use of Sentinel 2 – MSI for water quality monitoring at Alqueva reservoir, Portugal. Proceed. Int. Associat. Hydrol. Sci. 380, 73–79.

PSOMAS, 2007. Utah Lake TMDL: Pollutant Loading Assessment & Designated Beneficial Use Impairment Assessment.

Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnenstiel, G.L., 2012. Interannual variability of cyanobacterial blooms in Lake Erie, e42444. PLoS One 7 (8), 1–11.

The European Space Agency, 2015. Sentinel-2 MSI. Radiometric Resolutions. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric.

Uddin, S., Khan, A., Hossain, M.E., Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Informat. Dec. Making 19 (281), 1–16.

Utah Department of Environmental Quality, 2020. Utah HAB Guidance Summary; January 2020.

Utah Department of Natural Resources, 2016. Utah Lake closed due to harmful algal bloom. https://stateparks.utah.gov/2016/07/15/utah-lake-closed-due-to-harmful-algae-bloom/.

Utah Lake Fishing Map, 2022. http://www.gpsnauticalcharts.com/main/us_ut_01446894-utah-lake-nautical-chart.html.

Utah Office of Administrative Rules, 2006. Utah Admin Code R317–2–13-12.

Utah-State-Parks-Office, 2022. Utah State Parks Office, 2022. Utah state Parks Visitation Data by Fiscal Year.

Vargas-Lopez, I., Rivera-Monroy, V., Day, J., Whitbeck, J., Maiti, K., Madden, C., Trasviña-Castro, A., 2021. Assessing chlorophyll a spatiotemporal patterns combining in situ continuous fluorometry measurements and landsat 8/OLI Data across the Barataria Basin (Louisiana, USA), 512. Water 13 (4), 1–23.

Warren, M.A., Simis, S.G.H., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C., Ansper, A., 2019. Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters. Remote Sens. Environ. 225, 267–289.

Watanabe, F.S., Alcantara, E., Rodrigues, T.W., Imai, N.N., Barbosa, C.C., Rotta, L.H., 2015. Estimation of chlorophyll-a concentration and the trophic state of the barra bonita hydroelectric reservoir using OLI/Landsat-8 images. Int. J. Environ. Res. Public Health 12 (9), 10391–10417.

Zanazzi, A., Wang, W., Peterson, H., Emerman, S.H., 2020. Using stable isotopes to determine the water balance of Utah Lake (Utah, USA), 88. Hydrology 7 (4), 1–25.