Mapping Water Clarity in Small Oligotrophic Lakes Using Sentinel-2 Imagery and Machine Learning Methods: A Case Study of Canandaigua Lake in Finger Lakes, New York

Rabia Munsaf Khan, *Graduate Student Member, IEEE*, Bahram Salehi, *Senior Member, IEEE*, Milad Niroumand-Jadidi, *Member, IEEE*, and Masoud Mahdianpari, *Senior Member, IEEE*

***Abstract*—Optical remote sensing of water quality poses challenges in small oligotrophic lakes due to the diminishing contribution of constituents to the water-leaving radiance as water clarity increases. For monitoring water clarity over such lakes, this study utilizes machine learning models and data from citizen science to develop effective models for retrieving Secchi disk depth (SDD) in Canandaigua Lake, USA. Using Sentinel-2 band ratios as input, we trained random forest (RF), adaptive boosting, extreme gradient boosting, and support vector regression models using spatiotemporally distributed in situ data within 7 days of Senitnel-2 overpass. Each model was optimized using hyperparameter tuning, and cross-validation was used for accuracy assessment to compare the models' effectiveness in retrieving SDD. The results indicate the superior performance of RF with an R^2 of ∼0.74 and a root mean squared error of ∼0.72 m. A feature importance analysis for RF indicated the high relevance of the blue and green bands ratio in the estimation of SDD. The RF model was subsequently employed to generate temporal maps for Canandaigua Lake, indicating that water clarity tends to be higher during the early summer months (May and June) but declines during late summer and fall (September and October). This pattern can be closely associated with the increased algal presence in the lake. The spatial variability of the SDD indicated the possibility of greater sediments entering from the southern part of the lake. This study can be expanded to encompass other Finger Lakes, offering a comprehensive understanding of water clarity in these lake systems.**

***Index Terms*—Freshwater, machine learning (ML), oligotrophic lakes, random forest (RF), remote sensing, Secchi disk depth (SDD), Sentinel-2, water quality.**

Manuscript received 25 August 2023; revised 4 December 2023; accepted 13 January 2024. Date of publication 29 January 2024; date of current version 16 February 2024. The work of Rabia Munsaf Khan was supported by the Fulbright Program funded by the U.S. Department of State and Higher Education Commission, Pakistan. *(Corresponding author: Rabia Munsaf Khan.)*

Rabia Munsaf Khan and Bahram Salehi are with the Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210 USA (e-mail: rabiamunsaf@gmail.com; bsalehi@esf.edu).

Milad Niroumand-Jadidi is with Digital Society Center, Fondazione Bruno Kessler, 18 I-38123 Trento, Italy (e-mail: mniroumand@fbk.eu).

Masoud Mahdianpari is with C-CORE, St. John's, NL A1B 3X5, Canada, and also with the Department of Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada (e-mail: m.mahdianpari@mun.ca).

Digital Object Identifier 10.1109/JSTARS.2024.3359648

I. Introduction

LAKES are an important source of freshwater and are a substantial component of the hydrologic ecosystem. They hold pivotal socioeconomic value by providing ecosystem services like drinking and agricultural water, recreational activities, aquatic habitats, and biodiversity [1]. However, inland waters, including lakes, are endangered due to numerous natural and anthropogenic factors, such as point and nonpoint source pollution, deforestation, agricultural runoff, and climate change [2], [3], [4]. Increased nutrient concentration, along with environmental conditions, can lead to the development of harmful algal blooms (HABs) that render freshwater unfit for human recreation and consumption [5], [6]. The harmful effects are not limited to drinking water only but can affect the entire watershed and its linked biodiversity [7]. Therefore, it is important to efficiently manage freshwater resources by timely monitoring of water quality.

Water quality can be assessed by various parameters, among which water clarity is often used [8], [9]. The first instrument used to quantify water clarity was a Secchi disk, which is a round disk of black and white color (generally ∼20 to ∼30 cm), used to measure the depth of the water column when the disk is no longer visible [10]. The resulting depth is known as Secchi disk depth (SDD) and is affected by the presence of various dissolved and suspended matter in the water column, also referred to as Optically Active Constituents (OAC), that changes the underwater light field as the concentration of OACs change [11]. These OACs can be plankton or suspended sediments and are responsible for regulating many biophysical processes, such as primary productivity and thermal stratification; processes which can lead to the development of HABs [12]. This measure of water clarity is widely used by volunteer programs and citizen scientists, owing to its simplicity and no requirement for prior knowledge of the field [13], [14], [15]. However, continuous monitoring in terms of both spatial and temporal domains is required to gain a comprehensive understanding of trends in SDD for inland water bodies, which is hindered by the limited in situ data.

Remote sensing plays a crucial role in providing spatially and temporally explicit information on large scales as opposed to in

situ measurements, which are limited both in space and time. Satellite-based remote sensing data for water quality monitoring often uses ocean color sensors, such as Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer, and Ocean and Land Color Instrument [5], [16]. However, due to their coarse (∼250 to 110 m) spatial resolution, their applicability to small inland water bodies is limited. Hence, sensors with finer spatial resolutions (e.g., Sentinel-2) are required for the smaller lakes (area < 50 km^2). The Multispectral Instrument (MSI) onboard Sentinel-2 provides freely available data at varying spatial resolutions (10 m, 20 m, and 60 m) across 13 spectral bands with 12 bits of radiometric resolution and a revisit time of ∼5 days. Sentinel-2 has been used for water quality monitoring globally [1], [17], [18], [19] specifically for the estimation of water clarity [17], Chl-a [20], colored dissolved organic matter [21], and suspended particulate matter [22].

While these studies have demonstrated promising results in utilizing Sentinel-2 for SDD mapping, certain limitations arise when monitoring clear water bodies due to the low level of the water-leaving signal due to the high absorption of pure water. Clear waters being low in reflectance, pronounces the contribution of artifacts like atmospheric effects. In aquatic remote sensing, the sensor receives the water-leaving radiance, radiance from the atmosphere, and specular reflections (sun glints) from the water surface. The contribution of water-leaving radiance is often less than 20% of the at-sensor radiance due to the low reflectance of pure water in the visible and near-infrared spectrum. In oligotrophic lakes, where the concentration of constituents is lower, the absorption of water increases, resulting in a decrease in the water-leaving signal. Thus, atmospheric artifacts can account for ∼90% of the total reflectance posing severe challenges for retrieving water quality parameters [23]. Therefore, correction for atmospheric contributions is necessary to accurately estimate the remote sensing reflectance R_{rs} over water bodies, particularly clear lakes [24], [25]. The atmospheric correction methods vary for different environmental conditions, geographical locations, water types, and the type of remote sensing sensor [25], [26]. Although various atmospheric correction methods are used in literature, the use of Case-2 Regional Coast Color (C2RCC) has proved to be promising for clear waters [27], [28], [29].

The methods for water clarity estimation can be divided into two broad categories: physics-based and empirical-based methods. The physics-based methods include model parameterization based on the inherent optical properties of water and the atmosphere [30]. The advantage of using these models is that they can be generalized outside the range of a given study area. However, the model application requires extensive knowledge of the underlying physical properties of the water bodies, which can be challenging to obtain [31]. On the other hand, empirical methods fit a regression model between spectral features and coincident water quality parameter values. The empirical approaches are relatively straightforward to implement, however, the simple empirical algorithms lack the ability to accurately model the complex relations between R_{rs} and water quality parameters [32]. A solution to this problem is provided by machine learning (ML) algorithms as they are based on nonlinear regression models while being inherently empirical [33]. Literature also suggests the superior performance of ML methods over other empirical models [5], [34]. In particular, ML methods such as random forest (RF), adaptive boosting (AB), support vector regression (SVR), and neural networks have been used for estimating water clarity for inland water bodies [35], [36], [37], [38]. Among the various ML methods, RF and AB exhibited superior results as compared to other empirical and ML algorithms [37], [38], [39], [40], [41].

Previous works based on Sentinel-2 images focus mostly on mesotrophic or eutrophic lakes [15], [39]. It is important to monitor water clarity over small oligotrophic lakes, especially when they are the primary source of drinking water for the neighboring community [14]. One such case is Canandaigua Lake, which is part of the Finger Lakes in New York State, United States. These lakes have long been studied in terms of history, ecology [42], and water quality [43]. To build on these previous studies, the New York State Department of Environmental Conservation (NYSDEC) has conducted studies on these lakes with the help of citizen science. The published reports [13], [14], [44] indicate the need to closely monitor the quality of these lakes due to their integral socioeconomic value. However, continuous monitoring of these lakes using remote sensing data has been limited since small inland water bodies add to the existing complexities associated with aquatic remote sensing. Thus, this study evaluates the suitability of Sentinel-2 processed data for the estimation and mapping of SDD in Canandaigua Lake using different bagging and boosting ML methods and SVR methods. The specific research objectives are defined as follows: 1) compare multiple ML regression algorithms for mapping SDD in the oligotrophic water system of Canandaigua Lake using Sentinel-2 imagery and 2) apply the best regression algorithm to perform spatiotemporal trend analysis for SDD over Canandaigua Lake.

II. Materials and Methods

A. Study Area

This study is carried out on Canandaigua Lake (meaning "the chosen place") with a length of 24.9 km and a shoreline extending 66 km. This lake has a maximum depth of 83.5 m with a mean depth of 38.8 m. Among the Finger Lakes, Canandaigua Lake has a greater watershed area to surface area ratio (i.e., 11.3), with a watershed area being 482 km^2 and a surface area of 42.6 km^2. Located in Ontario and Yates Counties, it is a class AA water body according to water body classification [14] that renders it the safest for drinking and recreational purposes. With a total volume of 1600 million m^3, this lake provides drinking water to the city of Canandaigua and neighboring communities within its watershed.

B. In Situ Data

For this study, SDD is used as a water quality indicator mainly because of the readily available in situ SDD data. in situ data over Canandaigua Lake were collected from 2020–2022 for the

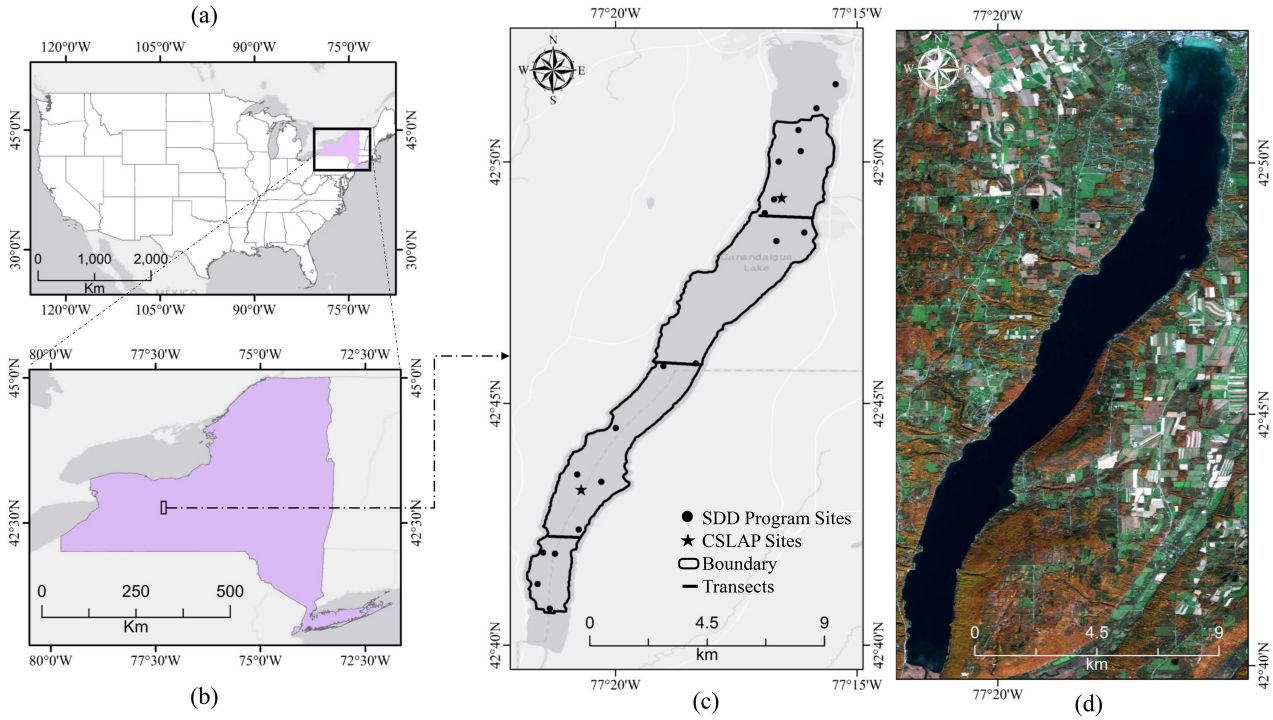

Fig. 1. (a) Study area showing the relative location of Canandaigua Lake within the conterminous USA. (b) Within the New York state. (c) Sampling sites over Canandaigua Lake. (d) Sentinel-2 imagery over Canandaigua Lake.

months of May to November, using the Citizen Science program conducted by the Canandaigua Lake Watershed Association (CLWA). To measure SDD, the Secchi disk is lowered into the water body until it disappears. The corresponding water depth is noted, then the disk is lowered further and raised again; the depth at which the disk reappears is noted again. The average of these two values is considered to be SDD. These consist of 18 sampling sites which majorly occupy the northern and southern parts of the lake (see Fig. 1).

In addition, data from the Citizens Statewide Lake Assessment Program (CSLAP) is used for 2018 and 2019. CSLAP is a lake management initiative taken by the New York State carried out through a partnership between the NYSDEC and New York State Federations of Lake Associations. Under this program, there are two sampling sites for Canandaigua Lake monitored every two weeks from June to September (which can extend to October in some cases). The locations of sampling sites are shown in Fig. 1 as two stars, namely North and South sites. As measurements over the shallow part of the lake include contribution from bottom reflectance, these are masked out and the boundary is shown in Fig. 1. The sampling sites from shallow areas of the lake are not included in model calibration and validation. Furthermore, for analysis purposes pixel values against three transects (North, Middle, and South) are compared for all years [see Fig. 1(c)].

C. Satellite Imagery

Sentinel-2 satellites (Sentinel-2A and Sentinel-2B) were launched by the European Space Agency in 2015 and 2017, respectively. These satellites are equipped with MSI, collecting imagery in 13 spectral bands within the blue to shortwave infrared portion of the spectrum at 10–60 m spatial resolutions. When using both satellites, the temporal resolution increases to five days over Canandaigua Lake, which renders them suitable for monitoring temporal dynamics of water quality parameters [39]. For this study, Sentinel-2 level-1 imagery from 2017 to 2022 for May to October was downloaded from Copernicus Open Access Hub. Images with partial cloud cover not hindering the sampling locations were also used to increase the number of match-ups with in situ data. The images were then resampled to 20 m before further processing.

D. Methods

1) Data Preparation: The overall methodological framework is presented in Fig. 2. First level-1 data is atmospherically corrected using C2RCC [45], [46]. C2RCC is suitable for atmospheric corrections of various sensors such as Sentinel-2, Landsat, and MODIS [27]. It takes the top-of-atmosphere image as an input and uses neural network-based inversion of the radiative transfer function to generate atmospherically corrected imagery. The output of atmospheric correction provides eight bands of *Rrs* in the visible and near-infrared regions.

To colocate Sentinel-2 *Rrs* pixels with in situ sampling points, the median value of a 3×3 pixel window centered on each sampling point was selected as the corresponding *Rrs* value (match-up) to the measured SDD. To increase the number of match-ups, a temporal window of seven days before and after the in situ sampling measurement was applied to Sentinel-2

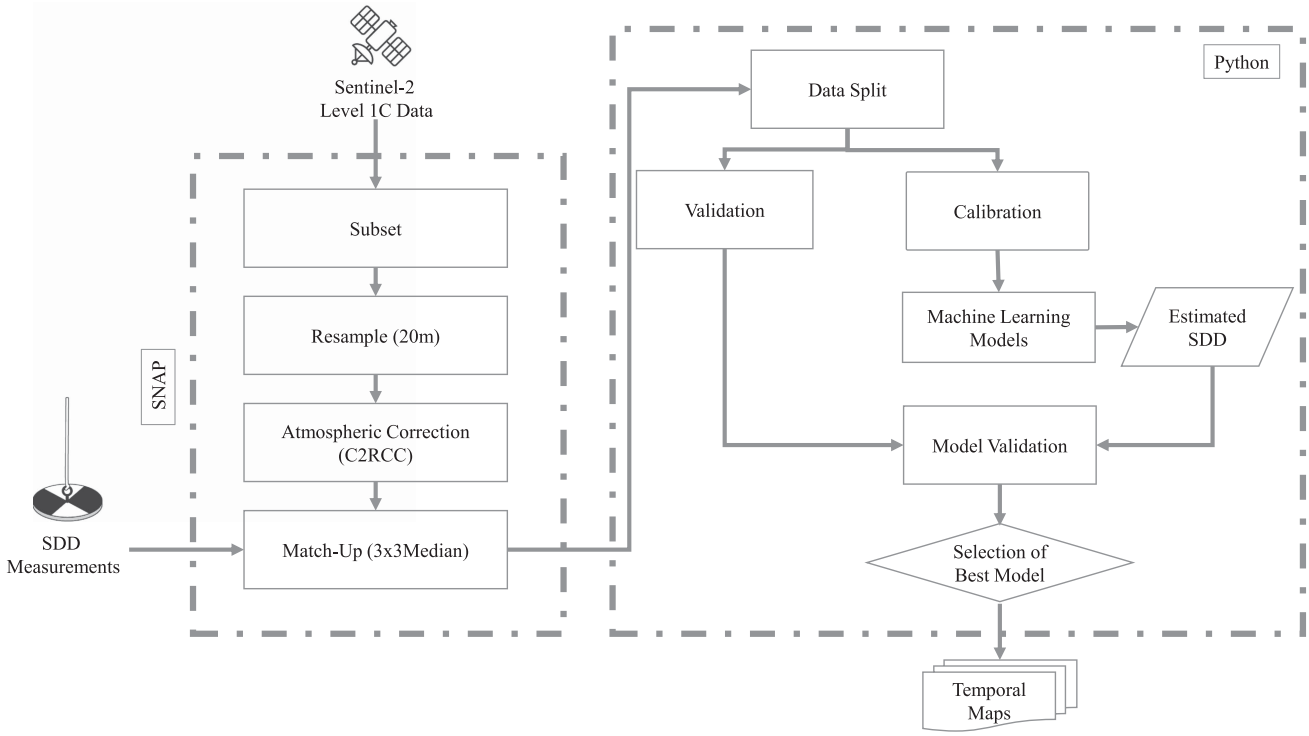

Fig. 2. Methodological workflow for estimating SDD over Canandaigua Lake.

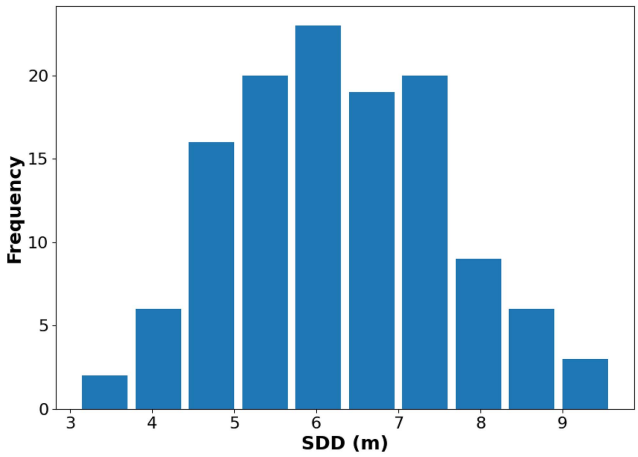

Fig. 3. Histogram showing the distribution for observed in situ SDD based on matchup data (N = 125).

Rrs data which is recommended in the literature for such cases [15], [46]. The total number of match-ups was 125, including 110 from the CLWA Citizen Science program and 15 from the CSLAP program. A histogram showing the distribution of the SDD values at these points is presented in Fig. 3. The SDD generally varies between 3.5 and 9.6 m. These 125 observations have an average (mean) of 6.2 m with a standard deviation of 1.3 m.

2) Machine Learning Models: This study compares the performance of various ML algorithms, including RF, AB, SVR, and extreme gradient boosting (XGB). The performance comparison of ML algorithms is also carried out against a simple empirical algorithm such as multilinear regression (MLR). Among the ML methods, RF is an ensemble model consisting of multiple decision trees which use the bagging method [47]. The model generates predictions against each tree based on a random sample generated from the training data. The final value is representative of outputs from all decision trees. Although each tree uses a subset of the whole training data, due to less correlation between the trees, the output as a whole is more reliable. As opposed to bagging, there are two models based on the boosting method that are used in this study: AB and XGB. Boosting generally uses an iterative training method, reducing the overall error [48]. AB, in particular, is built so that the models "adapt" to minimize errors from the previous trees. The advantage of this method is that the results are based on the output of all trees and not solely on the final tree [49]. XGB was originally developed by Chen and Guestrin [50] and is more powerful than regular gradient-boosting machines due to some additional features. These include tree pruning and regularization, which can lead to better computation [51]. Finally, SVR is used which is based on the generation of a hyperplane with the help of support vectors to differentiate between various values based on the extreme points in the dataset [52]. The usage of hyperplane makes it advantageous for multidimensional data. Furthermore, SVR can work well when the training sample is limited as it transforms the data to a higher dimension to overcome linearity [53].

The above-mentioned ML models were implemented using scikit-learn [54] with Sentinel-2 band ratios as input features as

compared to single spectral bands as their correlation to the in situ SDD is comparatively higher and are more robust features with respect to the atmospheric artifacts [1]. Before model implementation, hyperparameter tuning was carried out to optimize the model. The model parameters were iterated within ranges suggested in the literature to find the optimal ones for each model. In the case of RF, the number of trees (n_estimators) varied from 10 to 1000 (10, 50, 100, 500, and 1000) and the optimized value was 500. The criterion of "absolute error" was used and all other parameters were set at default values. The same range for the number of trees (n_estimators) was used for the boosting methods and the optimal value was 1000 for AB and 100 for XGB. Moreover, for AB, the learning rate was set to 0.1, and for both boosting models, all other parameters were used as default. In the case of SVR, two parameters were optimized: "*C*" which is the regularization factor and "gamma" which is the factor of the kernel used, which in this case was the radial basis function. The range of values tested for "*C*" were 10 to 1000 among which 100 is the optimal value and the gamma value of 0.0001 gives the best result between the values 0.0001 and 10. In the case of MLR, no optimization was needed for the model. Once the hyperparameters were selected for all the models, they were trained using in situ SDD data and validated based on a *k*-fold cross-validation. The number of folds (i.e., *k*) was set to 5, which meant that for every iteration, the percentage ratio of training and validation data split is 80:20. In the case of RF, the feature importance using Gini importance [55] was also calculated to visualize which input parameters contribute more toward the prediction of SDD.

3) Accuracy Assessment Metrics: The statistical accuracy measures used to assess the performance of ML models were the coefficient of determination R^2, root mean squared error (RMSE), mean absolute error (MAE), and bias (1)–(4). A value of R^2 closer to one indicates higher prediction power and the value close to zero indicates a lower prediction power of the model. Conversely, higher RMSE indicates a greater difference between the estimated and actual value which is why a lower RMSE is preferred. It is important to note that MAE and bias were calculated in log space and hence are dimensionless [56]. For interpretation purposes, a bias of 1.2 indicates an overestimation of 20%, and an MAE of 0.8 represents a mean relative error of 20%.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{1}$$

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$\text{MAE} = 10 \wedge \left(\frac{1}{n} \sum_{i=1}^{n} |\log_{10}(y_i) - \log_{10}(\hat{y}_i)| \right) \tag{3}$$

$$\text{Bias} = 10 \wedge \left(\frac{1}{n} \sum_{i=1}^{n} \log_{10}(y_i) - \log_{10}(\hat{y}_i) \right) \tag{4}$$

TABLE I
COMPARISON OF ACCURACY ASSESSMENT METRICS IN TERMS OF R^2, RMSE, MAE, AND BIAS FOR MLR, SVR, AB, XGB, AND RF

	MLR	SVR	AB	XGB	RF
R^2	0.64	0.44	0.72	0.71	**0.74**
RMSE (m)	0.909	1.047	0.746	0.807	**0.725**
MAE	1.130	1.141	1.109	1.119	**1.107**
Bias	**1.00**	0.99	0.99	0.99	0.98

The best performing model in terms of each accuracy metric is presented in bold.

where y_i represents the observed values, $\hat{y}_i$ represents the estimated values, $\hat{y}_i$ represents the mean of observed values, and n represents the number of observations used.

To have a robust selection measure, all accuracy measures were converted in terms of error and an uncertainty index (UI) was calculated for each model. Similar work has been carried out in terms of ranking [28] and normalizing [57] the individual metrics. Specifically, the coefficient of determination is subtracted from 1, the RMSE is converted to a percentage, and the absolute deviation of MAE and bias from 1 is calculated. For instance, a bias of 1.02 shows a 2% overestimation, hence only "0.2" will be included for error calculation. Based on these dimensionless error values denoted as E_i, the UI is calculated as follows:

$$\text{UI} = \frac{\sum_i^n E_i}{n} \tag{5}$$

where E_i is the normalized error for each of the accuracy measures, n is the number of accuracy measures, and UI is the average of all errors. Based on the UI, the best-performing model is selected to generate temporal maps of SDD over Canandaigua Lake.

III. RESULTS

A. Relative Performance of Machine Learning Models

The performance comparison of ML methods is summarized in Table I. The RF model outperformed other models in almost every accuracy metric. It is interesting to note a large difference in terms of R^2 between the methods, as compared to other metrics. RF showed the highest R^2 of 0.74 closely followed by 0.72 for AB, 0.71 for XGB, and 0.64 for MLR. SVM, however, showed poor performance in terms of R^2 (0.44) and RMSE (1.05 m).

In terms of bias, all models performed similarly with bias ranging from 0.98 to 1. As this is calculated in log space, hence, 0.98 indicates an underestimation of 2%. The MAE varies from the lowest value with RF (1.107) to the highest value for SVR (1.192), which indicates an error of 10.7% and 19.2%, respectively, whereas AB and XBG are similar to RF and MLR in terms of all accuracy measures. The results presented in Table II show the lowest error (best performance) by RF, followed by AB (∼3.2% increase), XGB (∼11.25% increase), MLR (∼26.5% increase), and SVR (∼74.78% increase). The performance comparison is illustrated in Fig. 4 in terms of a scatter plot showing the cross-validation results for each algorithm.

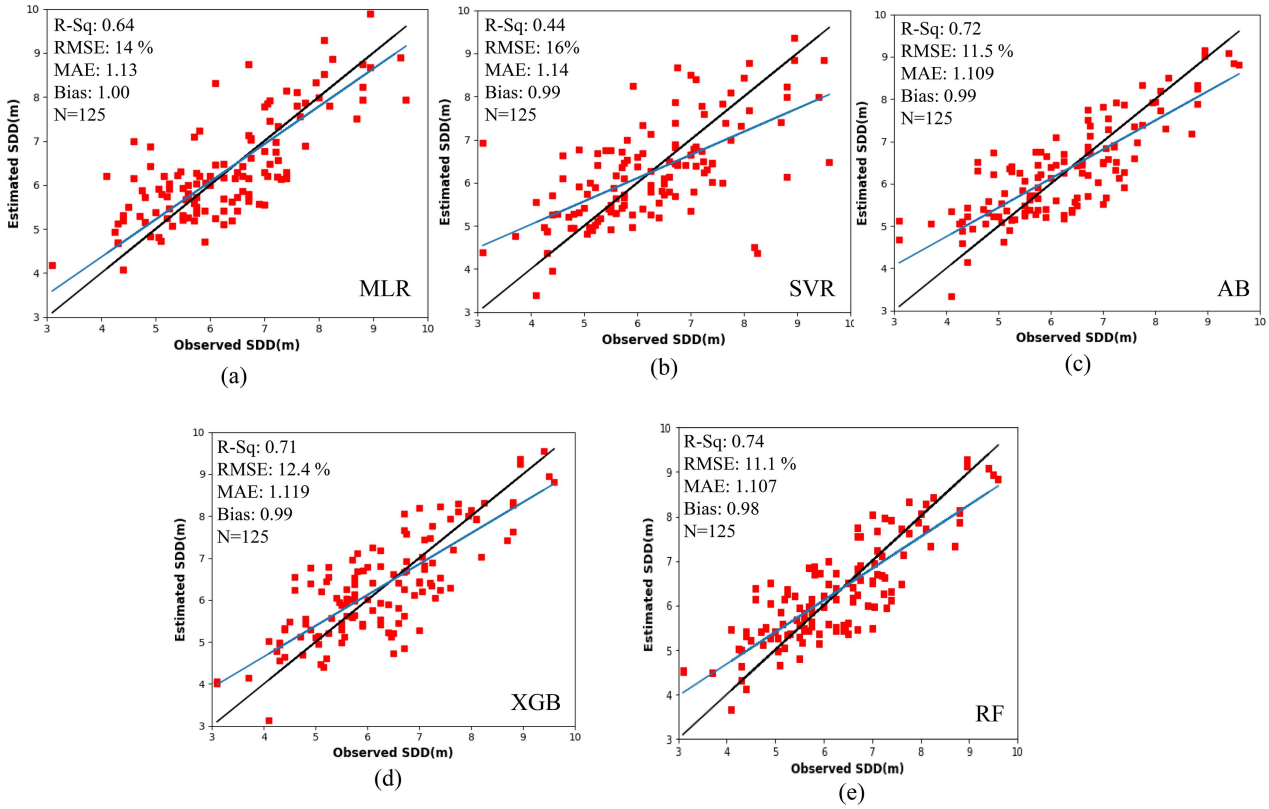

Fig. 4. Cross-validation scatter plot for ML algorithms. (a) MLR. (b) SVR. (c) AB. (d) XGB. (e) RF. The blue line indicates the regression line between observed SDD (m) and estimated SDD (m), whereas the black line indicates a 1:1 line.

TABLE II
COMPARISON OF INDIVIDUAL AND TOTAL ERRORS FOR THE SELECTION OF BEST MODEL AMONG MLR, SVR, AB, XGB, AND RF

	MLR	SVR	AB	XGB	RF
$(1\text{-}R^2)$	0.36	0.56	0.28	0.29	0.26
RMSE (%)	0.14	0.16	0.115	0.124	0.111
|1-MAE|	0.13	0.141	0.109	0.119	0.107
|1-Bias|	0	0.01	0.01	0.01	0.02
UI	0.1575	0.2176	0.1285	0.1358	**0.1245**

The model with least error is presented in bold.

B. Feature Importance for Random Forest Model

The feature importance for RF based on Gini importance is illustrated in Fig. 5 with a band ratio of B2/B3 (490 nm/560 nm) with the highest importance, followed by B2/B4 (490 nm/665 nm) and B1/B2 (443 nm/490 nm). It is important to note that the highest contributing parameters are based on the visible bands which are representative of the plankton and suspended sediments in the lake.

C. Application of Random Forest Model to Canandaigua Lake

The second objective of this research is to test the applicability of ML models to generate multi-temporal maps of the lake's SDD and understand its spatiotemporal variability. To do this,

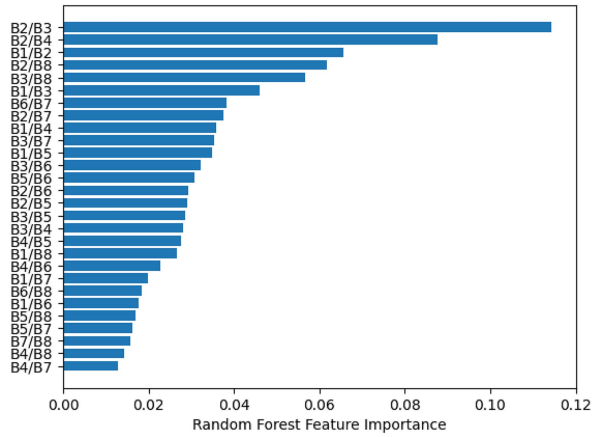

Fig. 5. Feature importance for each input feature for the RF model.

RF regression, as the best regression model, was applied to the cloud-free image of Sentinel-2 for each month (May to October) and corresponding SDD maps of the lake for the years 2020, 2021, and 2022 were generated (see Fig. 6). Note that some months are missing due to unavailability of clear (cloud-free) Sentinel-2 imagery over the lake. It is important to note that these estimates of SDD are only valid for the optically deep parts of the lake (i.e., negligible bottom-reflected radiance). For that purpose, the near-infrared band and visual interpretation of the

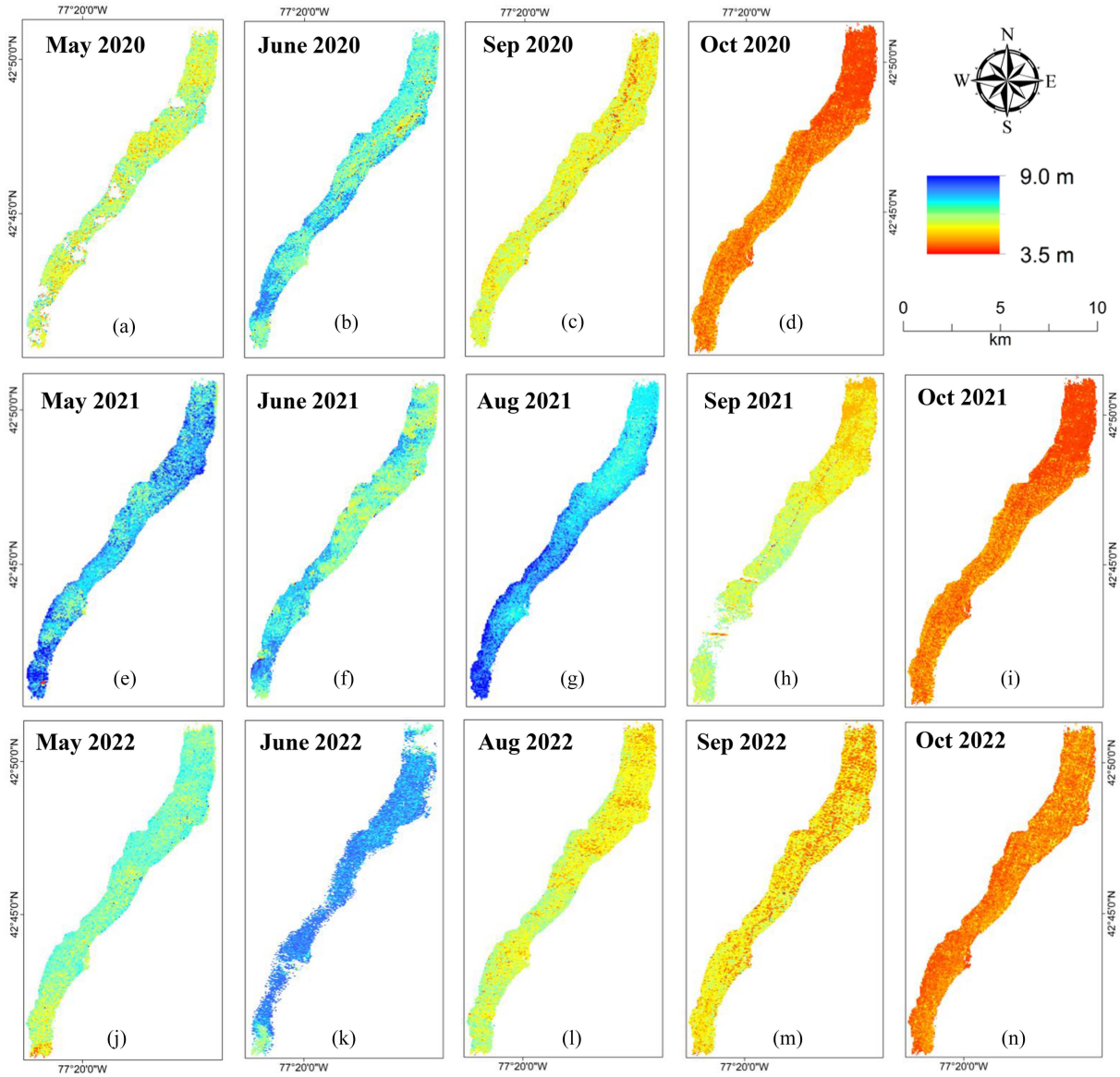

Fig. 6. Temporal maps showing the estimated SDD for Canandaigua Lake using one clear imagery for each of the months from June through October for years 2020 (a)–(d), 2021 (e)–(i) and 2022 (j)–(n). The red color indicates the lower values of SDD indicating turbid water and the blue value indicates high values of SDD indicating clearer waters.

lake, based on bathymetric data, were used to remove the shallow water parts from the lake. As the implemented model is based on the deeper parts of the lake only, hence any estimation provided by RF for shallow waters would not be reliable. Therefore, the maps shown in Fig. 6 are masked from the boundaries and only deeper parts of the lake are visualized. Furthermore, a buffer of 120 m is used to remove the mixed pixels from the boundary of the lake.

The SDD variability over the lake clearly shows a pattern that is similar for all years; the water clarity is high in early summer and low in late summer/fall months. A line of low water clarity pixels running up the lake is apparent each September, most obviously in 2021. The line is usually mid-lake but veers close to the shoreline in places. It tends to thicken and become more distinct further north. In May and June 2022 patches of low water clarity are seen at the south end of the lake, and in general, lower clarity is found along the eastern side of the lake in June of 2020 and 2021. Lower clarity is seen in September and October each year at the northern end of the lake.

D. Spatiotemporal Trend of SDD for Canandaigua Lake

A comparison between the predicted values from RF against in situ sampling sites was carried out for a total eight images, two from 2020 and three from each of 2021 and 2022.

Fig. 7 shows the estimated (by RF) and measured (in situ) means and standard deviations of SDD for sample points in each image. The number of sampling points varies for each image as mentioned in Table III. There is no standard deviation bar for August 2021 as there was only one matchup site against this

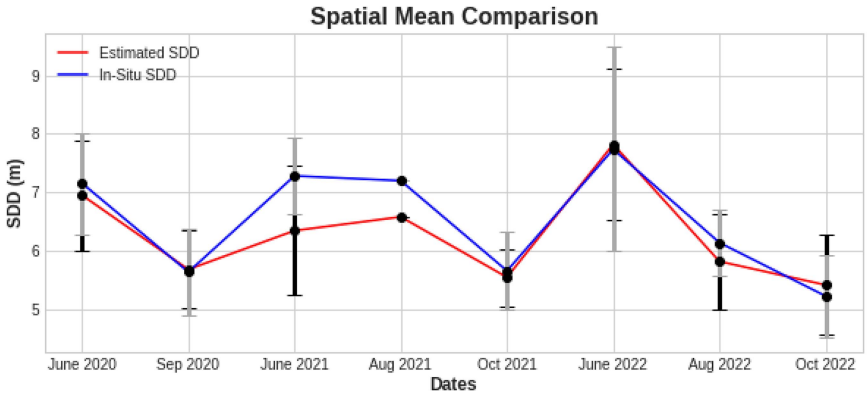

Fig. 7. Comparison of mean (dots) and standard deviation (vertical bars) between estimated and in situ SDD for selected images based on the sampling sites. The black bars indicate the standard deviation for estimated SDD for each image and the grey bars indicate the standard deviation for in situ SDD for each image.

TABLE III
NUMBER OF IN SITU SAMPLES AND COMPARISON OF ACTUAL SDD RANGE WITH MODEL ESTIMATED SDD RANGE

Imagery Date	No. of Ground Sample Points	Ground SDD Range [m]	Predicted SDD Range [m]
June 2020	8	6.25–8.8	5.3–8.1
September 2020	10	4.4–6.7	4.76–6.39
June 2021	7	6.7–8.7	4.2–7.69
August 2021	1	7.2 (Single value)	6.58
September 2020	6	4.6–6.6	4.75–6.05
June 2022	4	5.5–9.4	5.95–8.89
August 2022	6	5.4–6.9	4.8–6.92
October 2022	7	4.3–6.4	4.17–6.12

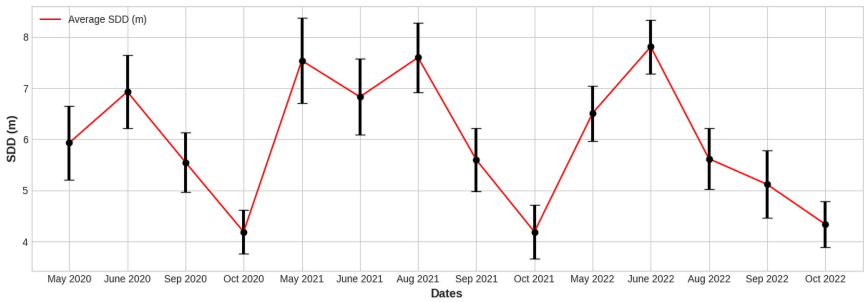

Fig. 8. Temporal variation of the mean (dots) and standard deviation (black vertical bars) of the estimated SDD for the entire lake for the years 2020, 2021, and 2022.

image. The graph shows a strong similarity between the estimated and in situ values for all images with a slight underestimation for June 2021.

The mean value of the estimated SDD is plotted for all the temporal maps along with their standard deviation in Fig. 8. As demonstrated in the temporal maps, the lowest average values are for the month of October. Specifically, the average for October 2020, 2021, and 2022 are 4.19 m, 4.19 m, and 4.34 m, respectively, with an approximate standard deviation between 0.43 and 0.53 m. The highest average SDD is shown in the months of May and June with means varying between 5.9 and 7.8 m with a standard deviation of 0.53 m to 0.83 m.

To further understand the SDD spatial variability in different regions of the lake, values along three transects were extracted from each temporal map. The three transects are in the North, Middle, and South parts of the lake as shown in black lines in Fig. 1. The extracted values against each transect are plotted for the years 2020, 2021, and 2022, as shown in Figs. 9–11.

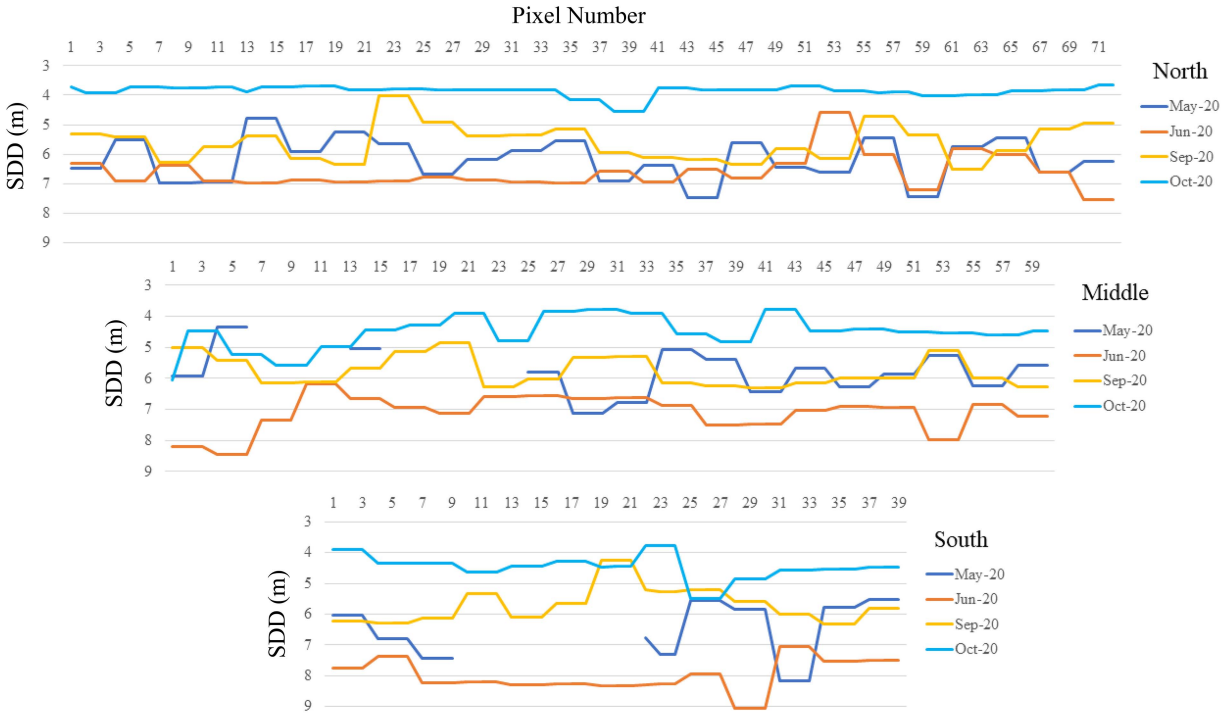

Fig. 9. Spatial variation along three transects for the year 2020 for Canandaigua Lake. Each line on the graph shows the spatial variation in SDD for the months of May, June, September, and October.

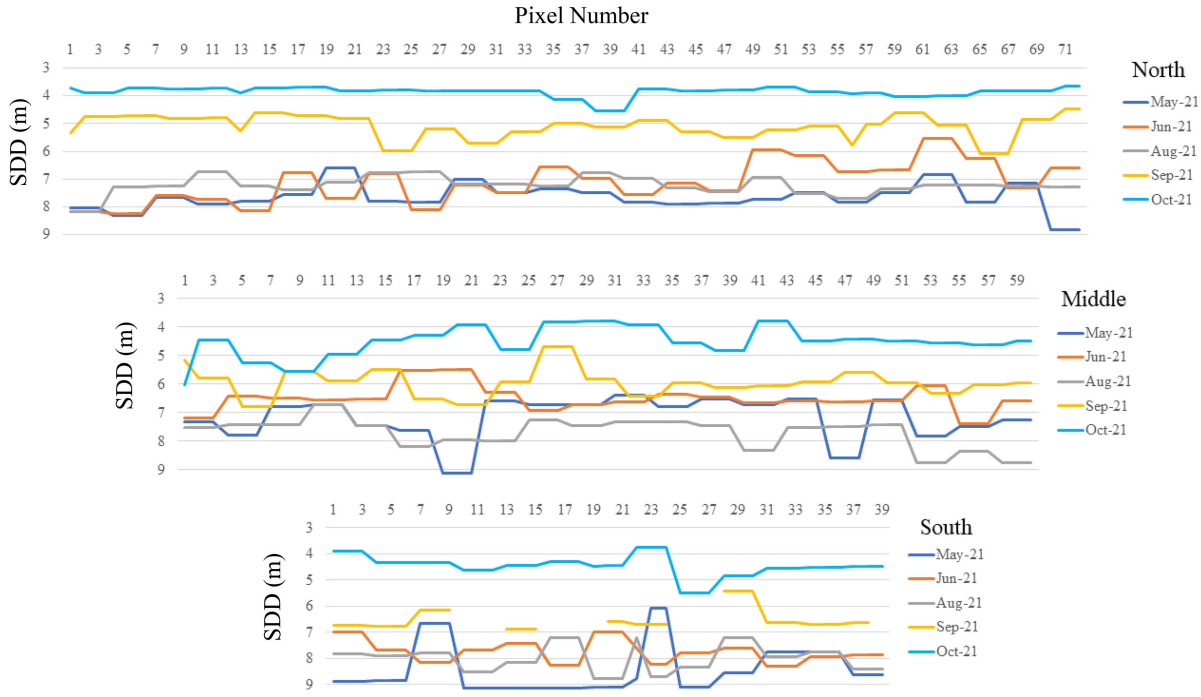

Fig. 10. Spatial variation along three transects for the year 2021 for Canandaigua Lake. Each line on the graph shows the spatial variation in SDD for the months of May, June, September, and October.

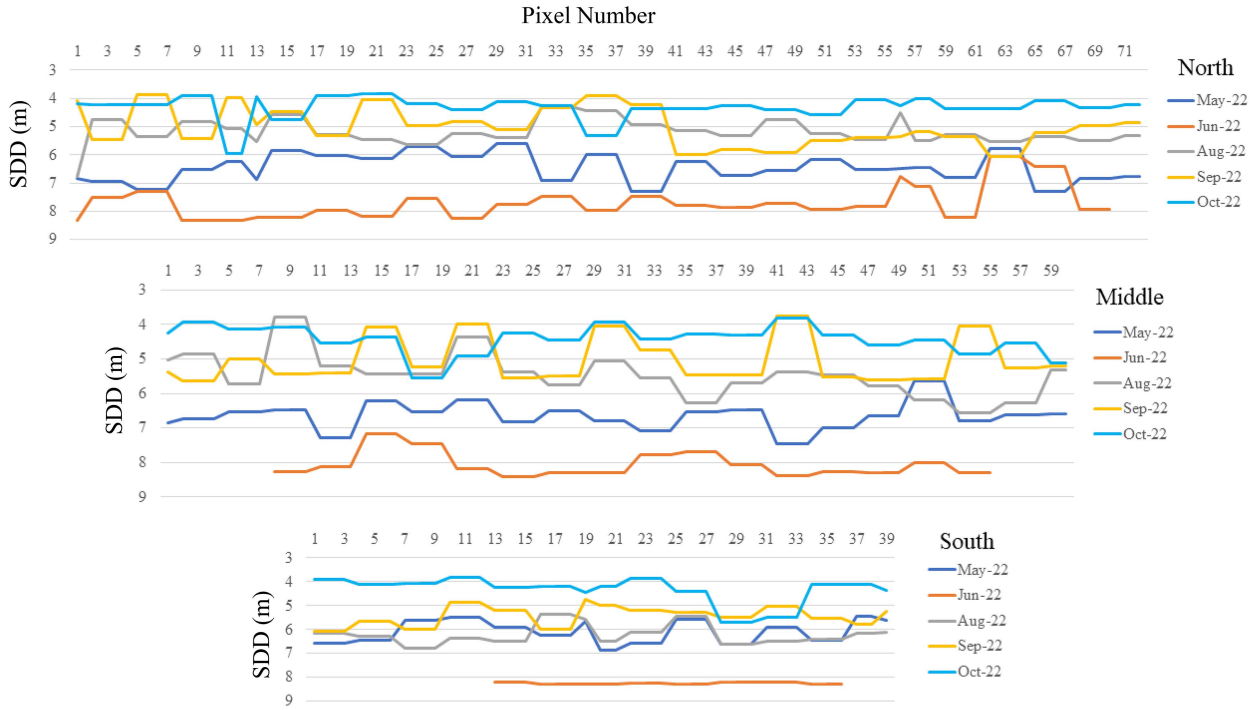

Fig. 11. Spatial variation along three transects for the year 2022 for Canandaigua Lake. Each line on the graph shows the spatial variation in SDD for the months of May, June, September, and October.

Although in terms of monthly variation the variation in SDD follows a similar pattern, i.e., lowest values in October and highest in early summer (May and June), it is interesting to see the variation between the transects. Transect 1 representing the North side of the lake has less variation in SDD values for various months as compared to the other two transects. Transect 2 represents the middle of the lake and SDD values against each month can be separated easily as compared to the North side of the lake. Finally, transect 3 represents the south side of the lake and it is the smallest transect in terms of distance as the lake gets narrower toward the South end. In this transect, the variability of SDD among different months is larger than in the other two transects. In addition, the variation along the transect is relatively smoother for the South side as compared to the North and middle parts of the lake. As there were some masked areas due to cloud cover and shadows, there are corresponding missing values along the transects for those images.

IV. Discussion

A. Model Performance

This study compared the performance of various ML models in mapping SDD and RF achieved the highest accuracy for Canandaigua Lake. To the best of our knowledge, this study is the first one to use remote sensing-based data and ML algorithms to estimate SDD over Canandaigua Lake, hence no direct comparison with previous studies can be made. The superior performance of RF is associated with its ability to compensate for missing data and ease in tuning the model on the dataset. It is important to note that despite the inconsistency in the distribution of sampling points in both space and time, RF-predicted SDD provides a reliable estimate for Canandaigua Lake supported by the validation results. For instance, as the in situ data was collected by citizen scientists, the samples were not coincident with the satellite overpass which resulted in a lower number of match-ups. In addition, due to cloud cover and shadows, the number of match-ups was reduced further and a longer temporal window (7 days) had to be selected to have a sufficient number of samples for training and validation. Furthermore, the data were collected at different times throughout the day which can also induce errors due to varying sun angle and illumination conditions [30].

In comparison with other studies on water clarity, the literature reports the RMSE values in the range of 30%–40% using linear empirical models [58], [59]. The results from ML models such as RF and AB are reported with higher accuracies ($R^2 > 0.65$ and RMSE < 1 m) with Sentinel-2 data [35], [39], [40], [60]. It is important to note that these studies were carried out on mesotrophic to eutrophic lakes which have higher concentrations of OACs hence higher contribution to the reflectance values. Therefore, water quality estimation over clear oligotrophic lakes in itself is challenging due to less reflectance from the water body as compared to turbid water bodies [28]. The collection of in situ reflectance spectra in addition to the water clarity parameters will be helpful to analyze errors due to atmospheric correction methods.

B. Feature Importance

In this study, Gini importance for input features was also calculated to understand the role of each input parameter in the estimation of SDD over Canandaigua Lake. To quantify the

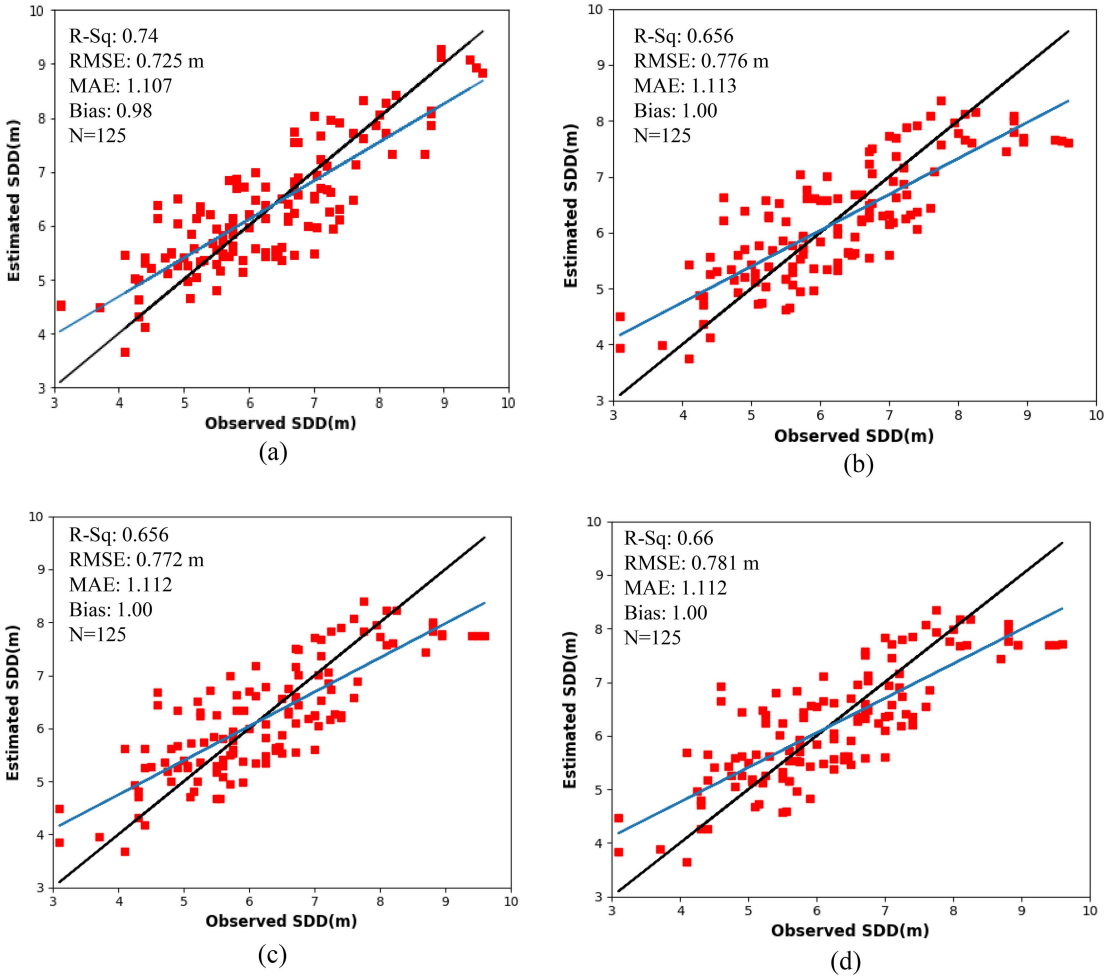

Fig. 12. Model validation results using (a) all input features, (b) top 75% input features, (c) top 50% input features, and (d) top 25% input features for estimation of SDD over Canandaigua Lake using RF.

impacts of input features, model validation was carried out using the top 75%, 50%, and 25% of the input features ranked as shown in Fig. 5. The scatter plots and associated accuracy metrics are illustrated in Fig. 12. Although the performance is greater with all input features involved, this type of analysis is important when dealing with large amounts of data. For this study, all input features were used as the dataset was small ($N = 125$) and the use of all input features was not resource-intensive. In contrast, for studies over larger areas with hundreds of sample points over multiple lakes (such as all Finger Lakes), this type of feature selection will be helpful to reduce the computation cost while maintaining comparative accuracy levels. While there is around a 9% decrease in terms of R^2, other accuracy matrices are not changed significantly for all the proportions of input features.

It is interesting to note that the highest contributing features in terms of Gini importance for RF belong to visible and near-infrared bands for Sentinel-2. The highest contributing band ratio for RF was B2/B3 (blue/green) which is also supported by literature for SDD monitoring in inland waters [61]. In comparison optimal band ratio analysis (OBRA) [62] was carried out to understand the linear dependency of SDD over Sentinel-2 band ratios. The results for OBRA are presented in Fig. 13 in which the highest correlation is shown between SDD and B2/B4 ratio. Although some of the band ratios (i.e., B2/B4 and B2/B3) among the top 25% of features for RF (as shown in Fig. 5) also show high correlation linearly, not all of the input features are correlated linearly with SDD.

For future studies, it will be interesting to add other input features such as environmental and meteorological parameters, as well as some indices in addition to the band ratios.

C. Spatiotemporal Variation of SDD

The spatiotemporal maps indicated low water clarity in the fall months (September and October) as compared to early summer (May and June). Although the range of SDD generated by RF and in situ SDD follow a similar pattern temporally, the pixel-by-pixel accuracy for each image is not accessed in this study. For instance, the line of low clarity appearing in the September image of each year can be associated with the presence of foam on that particular day. In elongated lakes like Canandaigua, when the upward and downward current meets, it can create this foam-like separation (see Fig. 14).

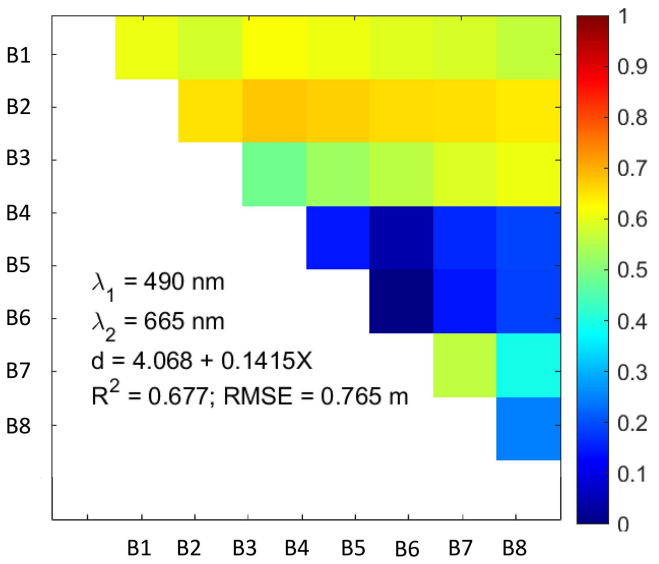

Fig. 13. OBRA for SDD against all Sentinel-2 atmospherically corrected bands consisting of visible to near infra-red bands. Here λ_1 and λ_2 denote the two wavelengths which result in the best correlation with in situ SDD measured in terms of R^2 and RMSE.

Fig. 14. Mid-lake foam patches and streaks on Canandaigua Lake, September 5, 2019. Outboard motor cowl for scale. In addition, the clarity of the lake is impacted by relatively high phytoplankton concentrations, resulting in a green color. Photo credit: A. Prestigiacomo, NYSDEC.

For spatial variation in the North and South sides of the lake, based on the transects, it is observed that the southern side is affected more in case of storm events as compared to the northern side of the lake. This indicates a possibility that the nutrients mostly enter the lake from the Southern side. The southern transect used for this study falls close to the tributary Vine Valley, and it is known to have high concentrations of phosphorous [63]. In addition, minor tributaries along the shoreline have small catchment areas with steep slopes, hence significant erosive power, and may also contribute to minor turbid plumes into the lake.

It is evident from Figs. 6 and 8 that there is a similar pattern of estimated SDD over all the years. However, this does not imply that this pattern will occur every year. For instance, a study recently published on Canandaigua Lake, indicates higher water clarity for October 2019 [64] which contradicts our observations. However, it is important to note that the maps generated in our study are based on a single day and are not representative of the monthly average. Prestigiacomo et al. [64] reported higher water clarity in June which aligns well with our observations, whereas the lower water clarity reported in that study is on August 22 and September 2 for the year 2019. This is also in agreement with our observations of low water clarity for September. Decreased SDD in their study was associated with increased algal greenness and visible cyanobacterial colonies. Prevailing winds in the Finger Lakes are south-westerly during the summer, causing buoyant cyanobacteria to be concentrated in the northern and eastern areas of each lake, in accordance with the known effects of lake fetch and orientation [65]. The New York state has seen an increase in reported cyanobacteria blooms over the last decade [66], with an associated surge in interest and concern from the public and professional water managers alike. The correlation of decreased clarity with increased cyanobacterial colonies and cyanotoxin concentrations could provide a practical application of our findings, by providing water managers with a frequent and spatially distributed perspective of a factor correlated with cyanobacteria and cyanotoxin hazards. However, there is a need to analyze longer temporal maps to establish trends in the SDD variability over Canandaigua Lake, for instance, using historical and ongoing Landsat missions.

V. Conclusion

Water quality monitoring of freshwater resources, especially lakes, is important as they are continuously endangered and thus, timely monitoring is required for conserving the ecosystem. For that purpose, this study assessed the performance of various ML methods, RF, SVR, AB, and XGB, for mapping SDD for Canandaigua Lake using Sentinel-2 imagery. Sentinel-2 imagery was first processed using C2RCC and then 28 band ratios were used as inputs for all the models. A robust scheme for model validation comparison revealed the superior performance of RF as compared to other ML methods. The trained RF model performed in terms of cross-validation with $R^2 \sim 0.74$, RMSE of about 11%, and MAE and bias of 1.107 and 0.98, respectively, which is greater than the accuracy metrics reported in the literature for linear empirical models (RMSE $\sim$ 30%). The feature importance for RF was also calculated and it showed the highest importance for B2/B3 (the ratio of the band ratio of blue to green band). The temporal maps for Canandaigua Lake indicated that the water clarity is generally higher in early summer (May and June) and decreases in late summer and fall months (September and October) which are in accordance with the increased algal presence. The spatial variability of the lake indicates lower water clarity in the Southern part of the lake which may be associated with the greater amount of nutrients entering from the southern end of the lake. However, further research is needed to confirm these trends based on the results from

a longer time period. This can be achieved by utilizing the longer temporal database of Landsat imagery. Furthermore, other ML and deep learning methods based on neural networks can be used in conjunction with additional input features based on environmental data, to potentially improve the accuracy of SDD estimation. As the ML models are data-driven and do not require any prior knowledge about the water bodies, the methodological framework used in this research can also be replicated for other water bodies with similar characteristics and SDD datasets and would provide useful information regarding water quality parameters.

ACKNOWLEDGMENT

The authors would like to thank the Canandaigua Lake Watershed Association, the Skaneateles Lake Association, and the New York State Department of Environmental Conservation (DEC) Citizens Statewide Lake Assessment Program and their citizen scientists for the collection and provision of data used in this study. The authors also extend gratitude to the New York State Center of Excellence in Healthy Water Solutions for their support of this research. Special thanks are extended to Dr. Lewis McCaffrey from NY-DEC Finger Lakes Hub for his thorough review and constructive feedback on the article. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the Fulbright Program, U.S. Department of State and Higher Education Commission, Pakistan.

REFERENCES

[1] K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery," *Remote Sens.*, vol. 8, no. 8, Aug. 2016, Art. no. 640, doi: 10.3390/rs8080640.

[2] X. Li, Y. Li, and G. Li, "A scientometric review of the research on the impacts of climate change on water quality during 1998–2018," *Environ. Sci. Pollut. Res.*, vol. 27, no. 13, pp. 14322–14341, May 2020, doi: 10.1007/s11356-020-08176-7.

[3] A. P. Yunus, Y. Masago, and Y. Hijioka, "COVID-19 and surface water quality: Improved lake water quality during the lockdown," *Sci. Total Environ.*, vol. 731, Aug. 2020, Art. no. 139012, doi: 10.1016/j.scitotenv.2020.139012.

[4] A. Romanelli, M. L. Lima, H. E. Massone, and K. S. Esquius, "Spatial decision support system for assessing lake pollution hazard: Southeastern Pampean shallow lakes (Argentina) as a case study," *Wetlands Ecol. Manage.*, vol. 22, pp. 247–265, 2014.

[5] R. M. Khan, B. Salehi, M. Mahdianpari, F. Mohammadimanesh, G. Mountrakis, and L. J. Quackenbush, "A meta-analysis on harmful algal bloom (HAB) detection and monitoring: A remote sensing perspective," *Remote Sens.*, vol. 13, no. 21, Oct. 2021, Art. no. 4347, doi: 10.3390/rs13214347.

[6] D. M. Anderson, P. M. Glibert, and J. M. Burkholder, "Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences," *Estuaries*, vol. 25, no. 4, pp. 704–726, Aug. 2002, doi: 10.1007/BF02804901.

[7] C. Acuna-Alonso, X. Alvarez, E. Valero, and F. A. L. Pacheco, "Modelling of threats that affect Cyano-HABs in an eutrophicated reservoir: First phase towards water security and environmental governance in watersheds," *Sci. Total Environ.*, vol. 809, Feb. 2022, Art. no. 152155, doi: 10.1016/j.scitotenv.2021.152155.

[8] R. E. Carlson, "A trophic state index for lakes," *Limnol. Oceanogr.*, vol. 22, no. 2, pp. 361–369, Mar. 1977, doi: 10.4319/lo.1977.22.2.0361.

[9] M. Gholizadeh, A. Melesse, and L. Reddi, "A comprehensive review on water quality parameters estimation using remote sensing techniques," *Sensors*, vol. 16, no. 8, Aug. 2016, Art. no. 1298, doi: 10.3390/s16081298.

[10] R. J. Davies-Colley, "Measuring water clarity with a black disk," *Limnol. Oceanogr.*, vol. 33, no. 4, pp. 616–623, 1988.

[11] J. T. Kirk, *Light and Photosynthesis in Aquatic Ecosystems*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[12] D. M. Anderson et al., "Harmful algal blooms and eutrophication: Examining linkages from selected coastal regions of the United States," *Harmful Algae*, vol. 8, no. 1, pp. 39–53, Dec. 2008, doi: 10.1016/j.hal.2008.08.017.

[13] A. Clinkhammer, S. Cook, L. McCaffrey, A. R. Prestigiacomo, and C. Sosa, "2017 Finger Lakes water quality report. Summary of historic Finger Lakes data and the 2017 citizens statewide lake assessment program," NYS Dept. Environ. Conserv.. Syracuse, NY, USA, Sep. 2018.

[14] A. Clinkhammer, S. Cook, L. McCaffrey, and A. R. Prestigiacomo, "2018 Finger Lakes water quality report. Summary of historic Finger Lakes data and the 2017-2018 citizens statewide lake assessment program," NYS Dept. Environ. Conserv., Syracuse, NY, USA, Nov. 2019.

[15] L. G. Olmanson, M. E. Bauer, and P. L. Brezonik, "A 20-year Landsat water clarity census of Minnesota's 10,000 lakes," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 4086–4097, Nov. 2008, doi: 10.1016/j.rse.2007.12.013.

[16] M. H. Gholizadeh, A. M. Melesse, and L. Reddi, "Spaceborne and airborne sensors in water quality assessment," *Int. J. Remote Sens.*, vol. 37, no. 14, pp. 3143–3180, Jul. 2016, doi: 10.1080/01431161.2016.1190477.

[17] M. Bonansea et al., "Evaluating the feasibility of using Sentinel-2 imagery for water clarity assessment in a reservoir," *J. South Amer. Earth Sci.*, vol. 95, Nov. 2019, Art. no. 102265, doi: 10.1016/j.jsames.2019.102265.

[18] M. Niroumand-Jadidi, F. Bovolo, L. Bruzzone, and P. Gege, "Inter-comparison of methods for chlorophyll-a retrieval: Sentinel-2 time-series analysis in Italian lakes," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2381, doi: 10.3390/rs13122381.

[19] R. M. Khan, B. Salehi, M. Mahdianpari, and F. Mohammadimanesh, "Water quality monitoring over finger lakes region using sentinel-2 imagery on Google Earth engine cloud computing platform," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 279–283, Jun. 2021, doi: 10.5194/isprs-annals-V-3-2021-279-2021.

[20] N. T. T. Ha, N. T. P. Thao, K. Koike, and M. T. Nhuan, "Selecting the best band ratio to estimate chlorophyll-a concentration in a tropical freshwater lake using Sentinel 2A images from a case study of Lake Ba Be (Northern Vietnam)," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 9, Art. no. 290, Sep. 2017, doi: 10.3390/ijgi6090290.

[21] Y. Shang et al., "Remote estimates of CDOM using Sentinel-2 remote sensing data in reservoirs with different trophic states across China," *J. Environ. Manage.*, vol. 286, May 2021, Art. no. 112275, doi: 10.1016/j.jenvman.2021.112275.

[22] V. H. Neves, G. Pace, J. Delegido, and S. C. Antunes, "Chlorophyll and suspended solids estimation in Portuguese reservoirs (Aguieira and Alqueva) from Sentinel-2 imagery," *Water*, vol. 13, no. 18, Sep. 2021, Art. no. 2479, doi: 10.3390/w13182479.

[23] S. Bi et al., "Inland water atmospheric correction based on turbidity classification using OLCI and SLSTR synergistic observations," *Remote Sens.*, vol. 10, no. 7, Jun. 2018, Art. no. 1002, doi: 10.3390/rs10071002.

[24] N. Pahlevan et al., "ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters," *Remote Sens. Environ.*, vol. 258, Jun. 2021, Art. no. 112366, doi: 10.1016/j.rse.2021.112366.

[25] D. Wang, R. Ma, K. Xue, and S. Loiselle, "The assessment of Landsat-8 OLI atmospheric correction algorithms for Inland waters," *Remote Sens.*, vol. 11, no. 2, Jan. 2019, Art. no. 169, doi: 10.3390/rs11020169.

[26] Q. Zeng, H. Zhang, X. Chen, L. Tian, W. Li, and G. Wang, "Evaluation on the atmospheric correction methods for water color remote sensing by using MERIS image: A case study on chlorophyll-a concentration of Lake Poyang," *Hupo Kexue/J. Lake Sci.*, vol. 28, no. 6, pp. 1306–1315, 2016, doi: 10.18307/2016.0616.

[27] J. Soriano-González et al., "Towards the combination of C2RCC processors for improving water quality retrieval in Inland and Coastal areas," *Remote Sens.*, vol. 14, no. 5, Feb. 2022, Art. no. 1124, doi: 10.3390/rs14051124.

[28] I. Ogashawara et al., "The use of Sentinel-2 for chlorophyll-*a* spatial dynamics assessment: A comparative study on different lakes in Northern Germany," *Remote Sens.*, vol. 13, no. 8, Apr. 2021, Art. no. 1542, doi: 10.3390/rs13081542.

[29] X. Sòria-Perpinyà et al., "Assessment of Sentinel-2-MSI atmospheric correction processors and in situ spectrometry waters quality algorithms," *Remote Sens.*, vol. 14, no. 19, Sep. 2022, Art. no. 4794, doi: 10.3390/rs14194794.

[30] Z. Lee et al., "Secchi disk depth: A new theory and mechanistic model for underwater visibility," *Remote Sens. Environ.*, vol. 169, pp. 139–149, Nov. 2015, doi: 10.1016/j.rse.2015.08.002.

[31] M. Niroumand-Jadidi, F. Bovolo, L. Bruzzone, and P. Gege, "Physics-based bathymetry and water quality retrieval using PlanetScope imagery: Impacts of 2020 COVID-19 lockdown and 2019 extreme flood in the Venice lagoon," *Remote Sens.*, vol. 12, no. 15, Jul. 2020, Art. no. 2381, doi: 10.3390/rs12152381.

[32] M. A. Matus-Hernandez, N. Y. Hernandez-Saavedra, and R. O. Martinez-Rincon, "Predictive performance of regression models to estimate chlorophyll-a concentration based on Landsat imagery," *PLoS One*, vol. 13, no. 10, Oct. 2018, Art. no. e0205682, doi: 10.1371/journal.pone.0205682.
[33] J. D. Olden, J. J. Lawler, and N. L. Poff, "Machine learning methods without tears: A primer for ecologists," *Quart. Rev. Biol.*, vol. 83, no. 2, pp. 171–193, Jun. 2008, doi: 10.1086/587826.
[34] S. N. Topp, T. M. Pavelsky, D. Jensen, M. Simard, and M. R. V. Ross, "Research trends in the use of remote sensing for inland water quality science: Moving towards multidisciplinary applications," *Water*, vol. 12, no. 1, Jan. 2020, Art. no. 169, doi: 10.3390/w12010169.
[35] D. A. Maciel, C. C. F. Barbosa, E. M. L. de Moraes Novo, R. F. Júnior, and F. N. Begliomini, "Water clarity in Brazilian water assessed using Sentinel-2 and machine learning methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 182, pp. 134–152, 2021, doi: 10.1016/j.isprsjprs.2021.10.009.
[36] M. Saberioon, J. Brom, V. Nedbal, P. Soucček, and P. Císar, "Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters," *Ecol. Indicators*, vol. 113, 2020, Art. no. 106236, doi: 10.1016/j.ecolind.2020.106236.
[37] L. F. Arias-Rodriguez, Z. Duan, R. Sepúlveda, S. I. Martinez-Martinez, and M. Disse, "Monitoring water quality of Valle de Bravo reservoir, Mexico, using entire lifespan of MERIS data and machine learning approaches," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1586, doi: 10.3390/rs12101586.
[38] M. Niroumand-Jadidi, F. Bovolo, M. Bresciani, P. Gege, and C. Giardino, "Water quality retrieval from Landsat-9 (OLI-2) imagery and comparison to Sentinel-2," *Remote Sens.*, vol. 14, no. 18, Sep. 2022, Art. no. 4596, doi: 10.3390/rs14184596.
[39] Y. Zhang et al., "Improving remote sensing estimation of Secchi disk depth for global lakes and reservoirs using machine learning methods," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 1367–1383, 2022, doi: 10.1080/15481603.2022.2116102.
[40] Y. Ma et al., "Remote sensing of turbidity for lakes in Northeast China using Sentinel-2 images with machine learning algorithms," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9132–9146, 2021, doi: 10.1109/JSTARS.2021.3109292.
[41] R. M. Khan, B. Salehi, M. Niroumand-Jadidi, and M. Mahdianpari, "Quantification and mapping of water clarity for freshwater lakes using Sentinel-2 data and random forest regression model: Application on Finger Lakes, New York," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 2890–2893.
[42] J. A. Bloomfield, *Lakes of New York, NY, USA State. Volume 1. Ecology of the Finger Lakes*. New York, NY, USA: Academic, 1978.
[43] C. W. Callinan, J. P. Hassett, J. B. Hyde, R. A. Entringer, and R. K. Klake, "Proposed nutrient criteria for water supply lakes and reservoirs," *J. AWWA*, vol. 105, no. 4, pp. E157–E172, 2013.
[44] R. M. Khan, B. Salehi, and M. Mahdianpari, "Machine learning methods for water quality monitoring over Finger Lakes using Sentinel-2," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 6316–6319, doi: 10.1109/IGARSS46834.2022.9883230.
[45] C. Brockmann, R. Doerffer, M. Peters, S. Kerstin, S. Embacher, and A. Ruescas, "Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters," in *Proc. Living Planet Symp.*, 2016, p. 54.
[46] Y. Zhang, Y. Zhang, K. Shi, Y. Zhou, and N. Li, "Remote sensing estimation of water clarity for various lakes in China," *Water Res.*, vol. 192, Mar. 2021, Art. no. 116844, doi: 10.1016/j.watres.2021.116844.
[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[48] G. Ridgeway, D. Madigan, and T. S. Richardson, "Boosting methodology for regression problems," in *Proc. 7th Int. Workshop Artif. Intell. Statist.*, 1999.
[49] J. C.-W. Chan and D. Paelinckx, "Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2999–3011, Jun. 2008, doi: 10.1016/j.rse.2008.02.011.
[50] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.
[51] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation," *Remote Sens.*, vol. 13, no. 21, Nov. 2021, Art. no. 4405, doi: 10.3390/rs13214405.
[52] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, pp. 199–222, 2004.
[53] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 123–140.
[54] O. Kramer and O. Kramer, "Scikit-learn," in *Machine Learning for Evolution Strategies*. New York, NY, USA: Springer, 2016, pp. 45–53.
[55] B. H. Menze et al., "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinf.*, vol. 10, 2009, Art. no. 123.
[56] B. N. Seegers, R. P. Stumpf, B. A. Schaeffer, K. A. Loftin, and P. J. Werdell, "Performance metrics for the assessment of satellite data products: An ocean color case study," *Optica Express*, vol. 26, no. 6, Mar. 2018, Art. no. 7404, doi: 10.1364/OE.26.007404.
[57] G. Ali et al., "Spatial–temporal characterization of rainfall in Pakistan during the past half-century (1961–2020)," *Sci. Rep.*, vol. 11, no. 1, Mar. 2021, Art. no. 6935, doi: 10.1038/s41598-021-86412-x.
[58] C. E. Binding, J. H. Jerome, R. P. Bukata, and W. G. Booty, "Suspended particulate matter in Lake Erie derived from MODIS aquatic colour imagery," *Int. J. Remote Sens.*, vol. 31, no. 19, pp. 5239–5255, Oct. 2010, doi: 10.1080/01431160903302973.
[59] M. Bonansea, M. C. Rodriguez, L. Pinotti, and S. Ferrero, "Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina)," *Remote Sens. Environ.*, vol. 158, pp. 28–41, Mar. 2015, doi: 10.1016/j.rse.2014.10.032.
[60] Y. Y. Adusei, J. Quaye-Ballard, A. A. Adjaottor, and A. A. Mensah, "Spatial prediction and mapping of water quality of Owabi reservoir from satellite imageries and machine learning models," *Egypt. J. Remote Sens. Space Sci.*, vol. 24, no. 3, pp. 825–833, 2021, doi: 10.1016/j.ejrs.2021.06.006.
[61] M. Pereira-Sandoval et al., "Calibration and validation of algorithms for the estimation of chlorophyll-a concentration and Secchi depth in inland waters with Sentinel-2," *Limnetica*, vol. 38, no. 1, pp. 471–487, 2019, doi: 10.23818/limn.38.27.
[62] C. J. Legleiter, D. A. Roberts, and R. L. Lawrence, "Spectrally based remote sensing of river bathymetry," *Earth Surf. Processes Landforms*, vol. 34, no. 8, pp. 1039–1059, 2009.
[63] B. A. Gilman and K. Olwany, "Health of Canandaigua Lake and tributary streams." *Canandaigua Lake Watershed Council*, Canandaigua, NY, USA, 2006.
[64] A. R. Prestigiacomo, R. M. Gorney, J. B. Hyde, C. Davis, and A. Clinkhammer, "Patterns and impacts of cyanobacteria in a deep, thermally stratified, oligotrophic lake," *AWWA Water Sci.*, vol. 5, no. 2, 2023, Art. no. e1326.
[65] E. M. Ostos, L. Cruz-Pizarro, A. Basanta, C. Escot, and D. George, "Algae in the motion: Spatial distribution of phytoplankton in thermally stratified reservoirs," *Limnetica*, vol. 25, no. 1/2, pp. 205–216, 2006.
[66] R. M. Gorney, S. G. June, K. M. Stainbrook, and A. J. Smith, "Detections of cyanobacteria harmful algal blooms (cyanoHABs) in New York, NY, USA State, United States (2012–2020)," *Lake Reservoir Manage.*, vol. 39, no. 1, pp. 21–36, 2023.

Rabia Munsaf Khan (Graduate Student Member, IEEE) received the B.S. (Hons.) degree in space science from the University of the Punjab, Lahore, Pakistan, in 2013, and the M.S. degree in remote sensing and geographic information science from the Institute of Space Technology, Islamabad, Pakistan, in 2017. She is currently working toward the Ph.D. degree in geospatial information science and engineering from the Department of Environmental Engineering, State University of New York College of Environmental Science and Forestry (SUNY ESF), Syracuse, NY, USA.

Her research interests include remote sensing and image analysis, regression modeling, machine learning, deep learning, geo big data analysis, and sustainable development goals for environmental applications with a particular focus on water quality.

Ms. Khan is currently a reviewer for journals including IEEE REMOTE SENSING AND GEOSCIENCE LETTERS and the *Canadian Journal of Remote Sensing*. She is a student member of the American Society of Photogrammetry and Remote Sensing (ASPRS) and the American Geophysical Union and a member of Geo Aquawatch Early Career Society. She is currently working with the IEEE GRSS Young Professional (YP) Ambassador, YP representative for the GRSS Remote Sensing, Environment, Analysis and Climate Technologies (REACT) technical committee and IEEE Climate and Sustainability Task Force. She also secured second place in the women category in GRSS Inspire Us photo competition and first place in smart city ideas, future leader congress in Bangkok, Thailand. She was the recipient of numerous awards and honors including the prestigious Fulbright scholarship, Fulbright travel grant, professional development grant from SUNY ESF, student grant for the ASPRS conference, laptops from the government of Pakistan, and merit scholarships throughout her academic journey.

Bahram Salehi (Senior Member, IEEE) received the B.Sc. degree in geomatics engineering from University of Tehran, Tehran, Iran, in 2002 and the M.Sc. degree in remote sensing engineering from K.N. Toosi University of Technology, Tehran, Iran in 2005, and the Ph.D. degree in geomatics engineering with a focus on remote sensing from the University of New Brunswick, Fredericton, NB, Canada, in 2012.

He is currently an Assistant Professor of Remote Sensing Engineering with the State University of New York College of Environmental Science and Forestry, Syracuse, NY, USA, and the Director of Salehi-Geolab. He mentored nearly 15 postdoctoral, Ph.D., and M.Sc. students, along with supervising several interns and undergraduate students. He has 20 years of combined academic and industry research and development experience in the U.S., Canada, and Iran. His expertise lies in multisensor (SAR, optical, and lidar) data fusion integrated with machine and deep learning for various environmental applications. He has coauthored more than 70 journal articles and book chapters with his graduate students and holds a Google Scholar h-index of 33 as of January 2024. Furthermore, his lab actively explores UAV photogrammetry and lidar point cloud processing through deep learning for the purpose of forest 3-D modeling. His research interests include forest biomass and carbon storage estimation, wetland classification and monitoring at regional and national scales, and water quality monitoring of freshwater lakes

Dr. Salehi was the General Chair of the IEEE-STRATUS 2022 and 2024, a UAV remote sensing conference to be held in Syracuse, NY, USA. Additionally, he holds the position of the National Director for the American Society for Photogrammetry and Remote Sensing, UAS division. His contributions to the field have been recognized with several awards, including the 2019 Early Career Achievement Medal from the Canadian Remote Sensing Society. In 2023, he was recognized as the third highly-ranked scholar in remote sensing in the United States according to ScholarGPS.

Milad Niroumand-Jadidi (Member, IEEE) received the B.Sc. degree in geomatics engineering from the University of Tabriz, Tabriz, Iran, in 2009, the M.Sc. degree in remote sensing engineering from K. N. Toosi University of Technology, Tehran, Iran, in 2013, and the Ph.D. degree in civil and environmental engineering from the University of Trento, Italy, and Freie Universität Berlin, Germany, in 2017.

Since 2017, he has been a postdoctoral researcher with the Remote Sensing for Digital Earth unit of the Digital Society Center, Fondazione Bruno Kessler, Trento, Italy. He authored or coauthored more than ten peer-reviewed articles in top journals and presented numerous papers at international conferences. He focuses on both physics-based and machine-learning models to retrieve information on water quality and bathymetry. His research interests include the development of methods and applications for remote sensing of inland and coastal waters from optical data.

Dr. Niroumand-Jadidi is on the Editorial Board for *Remote Sensing*, where he has also been a special issue Editor. He is a referee for several international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Remote Sensing of Environment*, *Remote Sensing*, *Journal of Hydrology*, *Water*, and *International Journal of Remote Sensing*. He was the recipient of several international awards. Among those, he has been awarded a DLR-DAAD fellowship for spending three months (May–August 2022) as a visiting scientist at the German Aerospace Center (DLR). He was also the recipient of the Best Young Author Award from the International Society for Photogrammetry and Remote Sensing in 2021. Furthermore, he was the recipient of the Best Paper Award from SPIE Remote Sensing conferences for three years in a row (2015 in Toulouse, 2016 in Edinburgh, and 2017 in Warsaw). He is also a recipient of the award for potential long-range contribution to the field of optics and photonics from SPIE (the International Society for Optics and Photonics) in 2016. Moreover, he is the recipient of the Alexander Goetz Instrument Support Award 2016, which gave him access to a field spectroradiometer during the Ph.D. research.

Masoud Mahdianpari (Senior Member, IEEE) received the B.S. degree in surveying and geomatics engineering and the M.Sc. degree in remote sensing engineering from the Department of Engineering, University of Tehran, Tehran, Iran, in 2010 and 2013, respectively, and the Ph.D. degree in electrical engineering from the Department of Engineering and Applied Science, Memorial University, St. John's, NL, Canada, in 2019.

He was an Ocean Frontier Institute postdoctoral fellow with Memorial University, St. John's, NL, Canada, and C-CORE. He is currently a Remote Sensing Technical Lead with C-CORE and a Cross Appointed Professor with the Faculty of Engineering and Applied Science, Memorial University. He is the author or coauthor of more than 150 publications, including peer-reviewed journal articles, conference papers, books, and book chapters. His research interests include remote sensing and image analysis, with a special focus on PolSAR image processing, multimodal data analytics, machine learning, geo Big Data, and deep learning.

Dr. Mahdianpari is an Editorial Team Member for *Remote Sensing*, IEEE GEOSCIENCE AND REMOTE SENSING, and *Frontiers in Environmental Science*. He was a recipient of the Research and Development Corporation Ocean Industries Student Research Award, organized by Newfoundland Industry and Innovation Center, amongst more than 400 submissions, in 2016, the T. David Collett Best Industry Paper Award organized by IEEE in 2016, the Com-Adv Devices, Inc., Scholarship for Innovation, Creativity, and Entrepreneurship from the Memorial University in 2017, the Artificial Intelligence for Earth Grant organized by Microsoft in 2018, the Graduate Academic Excellence Award organized by Memorial University in 2019, and the Best Industry Paper Award organized by the IEEE Newfoundland Electrical and Computer Engineering Conference in 2020 and 2023. In 2023, he achieved a global ranking in the top 1% of scientists, as reported by Stanford and Elsevier.