Remote Sensing of Turbidity for Lakes in Northeast China Using Sentinel-2 Images With Machine Learning Algorithms

Yue Ma, Kaishan Song, Zhidan Wen, Ge Liu, Yingxin Shang, Lili Lyu, Jia Du, Qian Yang, Sijia Li, Hui Tao, and Junbin Hou

***Abstract*—Monitoring water quality of inland lakes and reservoirs is a great concern for the public and government in China. Water turbidity is a reliable and direct indicator that can reflect the water quality. Remote sensing has become an efficient technology for monitoring large-scale water turbidity. This study aims to search an optimal regression model to accurately predict water turbidity using remote sensing data. To achieve this goal, 187 water samples were collected from field campaigns across Northeast China, in 2018, of which the samples were gathered within ±6 days of Sentinel-2 overpasses. The spectral reflectance data was used as independent variables for modeling. The simple regression, partial least squares regression, support vector regression, extreme learning machine, back-propagation neural network, classification and regression tree, gradient boosting decision tree (GBDT), random forest (RF), and K-nearest neighbor were used to compare. From model validation, we identified GBDT as the best regression model ($R^2 = 0.88$, RMSE = 9.90 NTU, MAE = 6.71 NTU). We applied GBDT to retrieve the water turbidity and obtained a satisfactory result. Feature selection technique from tree-based ensemble method was also tested. We selected B2, B3, B4, and B5 as the important variables because of their high ability to explain the variation of turbidity. These results demonstrated the significance of using a promising method to retrieve water turbidity using Sentinel-2 imagery at the regional scale. It is beneficial to monitor the spatial-temporal distribution of water turbidity; support water quality management and inland water environment protection.**

Manuscript received February 10, 2021; revised July 1, 2021 and August 15, 2021; accepted August 21, 2021. Date of publication September 3, 2021; date of current version September 22, 2021. This work was supported in part by the National Key Research and Development Project of China under Grant 2019YFC0409105, in part by the National Key R&D Program of China under Grant 2016YFB0501502, in part by the National Natural Science Foundation of China under Grant 40170304, and in part by the Outstanding Yong Scientist Foundation of Institute of Northeast Geography and Agroecology (IGA), Chinese Academy of Sciences. *(Corresponding author: Kaishan Song.)*

Yue Ma is with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, Jilin 130102, China, and also with the School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun, Jilin 130118, China. (e-mail: mayue417@hotmail.com).

Kaishan Song, Zhidan Wen, Ge Liu, Yingxin Shang, Lili Lyu, Jia Du, Sijia Li, Hui Tao, and Junbin Hou are with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, Jilin 130102, China. (e-mail: songkaishan@iga.ac.cn; wenzhidan@iga.ac.cn; liuge@iga.ac.cn; shangyingxin@iga.ac.cn; lvlili@iga.ac.cn; jiaqidu@neigae.ac.cn; lisijia@iga.ac.cn; taohui@iga.ac.cn; hou0621@foxmail.com).

Qian Yang is with the School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun, Jilin 130118, China. (e-mail: jluyangqian10@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3109292

***Index Terms*—Machine learning algorithms, remote estimation, Sentinel-2, water turbidity.**

I. Introduction

LAKES and reservoirs are important inland water resources, which play a significant role in ecological environment, industrial production, and human wellbeing [1], [2]. Freshwater resources from lakes and reservoirs are crucial for agricultural irrigation and hydropower [3], [4]. Lakes and reservoirs are capable of regulating runoff, adjusting regional climate, supporting navigation, and flood control [5]. Furthermore, they also serve other functions such as agricultural production and recreational purpose which advance the local economic efficiency [6], [7]. Given these beneficial functions, extra efforts are needed to monitor the water quality of lakes and reservoirs, protect and manage the inland water environment. Basically, water quality is easily affected by the presence of phytoplankton, suspended sediment, organic pollutants, and dissolved substances. Furthermore, many water quality parameters, such as chlorophyll-a (Chla) concentration, total suspended solids (TSS), and colored dissolved organic matters (CDOM), are often used to measure the condition of water[4], [8], [9]. These water constituent concentrations jointly modulated water turbidity. Turbid water will decrease the depth that sunlight penetrates in water. It can inhibit the photosynthesis and the underwater plant may die, which impairs the aquatic life and water quality for aquatic and human life. Therefore, it is necessary to utilize a reliable indicator that can adequately represent the complex components and water condition of inland lakes and reservoirs. Turbidity describes the amount of light scattered or blocked by suspended particles in water [8], [10]. It is a common physical water quality parameter for measuring the condition of water and describing the level of water clarity [11]–[14]. In addition, due to the strong relationship between turbidity and total suspended solid, low-cost turbidity measurement can be applied to estimate total suspended solids (instead of using the gravimetric method which requires a lengthy and costly procedure) [8], [15]. However, dissolved substances (such as CDOM) that also affect water clarity may be too small to be counted in a suspended solids concentration, but they can be part of a turbidity measurement [12]. Many factors can influence turbidity, but not as the direct indicator of water quality. Previous studies suggested that

turbidity can indirectly reflect the natural and anthropogenic conditions, such as the changes of land use and cover, soil erosion, weather, and urban expansion [16]–[20]. A remarkable increase in turbidity in the previously clear water should arouse attention. The potential problems may have a strong impact on water quality and human life. Therefore, regularly measuring and monitoring water turbidity is very essential and significant.

Turbidity is an optical property. It is closely associated with the particulate backscattering coefficient, thus can impact the water surface reflectance [13], [15]. Compared to the traditional *in situ* measurement, which is time-consuming, expensive, and limited ability to capture spatiotemporal distribution, remote sensing appears to be a promising technology that can rapidly, dynamically, effectively, and cost-efficiently monitor large-scale water turbidity [21]–[23]. Many previous studies have demonstrated that the remote sensing method can be used to estimate the water turbidity and monitor water environment. Landsat, Geostationary Ocean Color Imager (GOCI), and Moderate Resolution Imaging Spectroradiometer (MODIS) imageries with low and medium resolution are the most widely adopted multispectral data in turbidity mapping [24]–[30]. Nevertheless, few studies have investigated whether the Sentinel-2 multispectral satellite data with the fine spatial resolution could sufficiently characterize water turbidity [31], [32]. Briefly, the Sentinel-2 MultiSpectral Instrument (MSI) remote sensor is different from the Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager (OLI) sensors in terms of spectrum range and spectral resolution. It is necessary to examine the applicability of Sentinel-2 imagery for mapping water turbidity.

Regarding the estimation methods of water turbidity, the simple regression (SR) algorithm which applies the strong relative spectra reflectance based on mathematical function and statistical theory is generally used [33]–[38]. However, as a proxy of optical indicator, water turbidity is often influenced by several substances in water. It is limited to clarify the relationship between water turbidity and spectra reflectance using simple mathematical function [39]–[41]. Machine learning algorithms are well known for their prominent advantage in data mining [42]. These methods may well describe the nonlinear relationship between objective and feature variables. Due to the good generalization capability and robustness, machine learning algorithms are less affected by the noise in the data and can effectively learn the potential characteristics of data to handle new dataset well. Moreover, machine learning models are easier and more flexible to implement for most researchers who own small dataset with high-dimensional features, and it is enough to run the programs on computers with CPU. Whilst majority of machine learning algorithms [e.g., support vector regression (SVM), extreme learning machine (ELM), back-propagation neural network (BP), decision tree regression (CART), tree-based ensemble regression, and K-nearest neighbor (KNN)] can solve regression problems, they are usually applied to solve classification problems due to their high overall classification accuracy [43]–[45]. In particular, few studies evaluated the predictive performance of these machine learning methods in mapping water turbidity. However, it is necessary to take into account the problems of overfitting during the whole modeling process. The optimal hyperparameters for machine learning methods are searched in the parameter space and determined according to the best cross validation score. It may contribute to balance model complexity with dataset and effectively reduce the risk of overfitting.

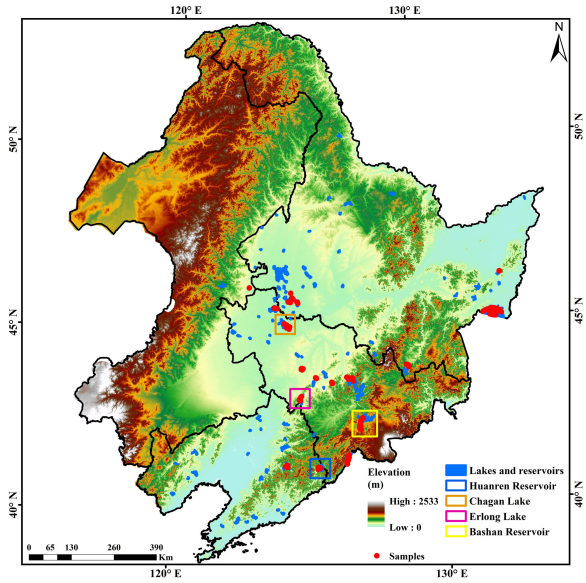

Fig. 1. Distribution of sampled lakes and reservoirs in different regions with respecting to elevation across Northeast China.

Many lakes and reservoirs in Northeast China show obvious spatial heterogeneity in terms of water turbidity [46]. These lakes and reservoirs in the flat plains are generally shallow with high turbidity, while those in the mountainous areas are usually deep and clear [47]. To better understand whether different machine learning methods can adequately predict the water turbidity of lakes and reservoirs, we examine our calculations of water turbidity for a broad range of waters in Northeast China. The main research objectives of this study are to: 1) Relate the Sentinel-2 spectral reflectance to water turbidity; 2) compare and evaluate the predictability of different machine learning regression algorithms; 3) identify the key variables for water turbidity estimation; and 4) map the water turbidity of typical lakes and reservoirs to verify the ability of model application.

II. Materials and Methods

A. Study Area

In this article, the lakes and reservoirs are distributed in Northeast China (Latitude: 42°49′ N - 49°12′ N, Longitude: 121°38′ E - 128°30′ E), covering an area of 1 240 897 km^2 (Fig. 1). The Changbai Mountain, Greater Khingan Mountains, and Lesser Khingan Mountains are on the edge of this region with high terrain, while the Songhua River, Nen River, and Liao River are in the middle, which formed a large area of plain (the Liaohe Plain, Songnen Plain, and Sanjiang Plain) with low terrain. This region is characterized by a typical semihumid monsoon climate and four distinct seasons. The annual average temperature ranges 2–6 °C, and the annual average precipitation approximately ranges 350–700 mm [47]. There are 631 lakes

and reservoirs with area greater than 1 km^2, with a total surface area of 8294 km^2. Many saline soda lakes appeared in Northeast China because of the unique geographical and climatic conditions. Most of the shallow waters are mesotrophic or eutrophic because of the rapid development of industry and agriculture, and the intense human activities [4].

In this study, four typical and important lakes and reservoirs were selected for detailed evaluation and analysis. They are Baishan Reservoir and Huanren Reservoir located in the mountainous area with high terrain, and Chagan Lake and Erlong Lake located in the middle of Northeast China, which are in the residential area with low terrain.

B. Field Sampling and Laboratory Analysis

Field campaigns were conducted to measure the water turbidity of the selected lakes and reservoirs in June, July, August, and September, 2018. We also acquired the Sentinel-2 satellite imagery data for the same periods of the field campaigns. During field measurement, water samples were collected approximately 0.5-m below the water surface. The samples were kept in the portable refrigerator at 4°C and they were delivered to the laboratory within seven days. The geographical coordinates of the sampling stations were recorded using a Trimble PXRS (Trimble Navigation, Inc., Sunnyvale, CA, USA) global positioning system (GPS). In the laboratory at room temperature (20 $\pm$ 2°C), water turbidity of each water sample was determined using a UV-VIS spectrophotometer (SHIMADZU UV-2600, Japan) which was widely used in turbidity measurement. The spectrophotometer had a spectral range between 185 and 900 nm and used a 3-cm quartz cell. The artificial turbid water with 400 NTU was prepared by mixing the solution of $(N_2H_4)H_2SO_4$ and $(CH_2)_6N_4$. The water without turbidity was prepared by using distilled water filtered through 0.2 glass fiber membranes. These two kinds of water as the reference were used to obtain standard turbidity solution. The absorbance curve of standard solution (0, 4, 10, 20, 40, 80, and 100 NTU) at 680 nm was used as calibration curve to obtain measured water turbidity of samples.

C. Imagery Data and Preprocessing

1) Sentinel-2 Acquisition: The Copernicus Sentinel-2 mission, which provides a global coverage of Earth's land surface every 5 or 10 days by a constellation of two satellites (Sentinel-2A and Sentinel-2B), aims at monitoring the changes of global land surface condition. The multispectral instrument contains 13 spectral bands with central wavelength ranging from 0.443 to 2.190 μm [B1 (443 nm), B2 (490 nm), B3 (560 nm), B4 (665 nm), B5 (705 nm), B6 (740 nm), B7 (783 nm), B8 (842 nm), B8A (865 nm), B9 (945 nm), B10 (1375 nm), B11 (1610 nm), and B12 (2190 nm)] [6], [48]. The Sentinel-2 sensor features high spatial resolutions, including 10, 20, and 60 m on different bands. Level-1C Sentinel-2 TOA products, $100\times100 km^2$ orthoimages in UTM/WGS84, are freely accessible for global users [31]. The images can be downloaded from the United States Geological Survey (USGS) website[1] and Sentinel's Scientific Data Hub[2]. In this study, 27 cloud-free Sentinel-2A images were selected in accordance with the dates of field surveys. Due to the frequent satellite revisit time, the time window for Sentinel-2A over passing determined $\pm$6 days that the water turbidity is relatively stable.

2) Image Preprocessing: Sen2cor is a prototype processor that performs Atmospheric Correction (AC, including Cirrus clouds and terrain correction) for Sentinel-2 MSI products. A large database of look-up tables (LUTs) has been compiled using an atmospheric radiative transfer model based on libRadtran1. Level-1C (L1C) top-of-atmosphere (TOA) image data were converted into an orthoimage Level-2A (L2A) bottom-of-atmosphere (BOA) reflectance product by executing the software via the windows command line. The Cirrus band 10 without surface information was omitted in the L2A output [49]. The generated L2A images are resampled with 20-m spatial resolution for 12 bands using the Sentinel-2 Toolbox (Sentinel Application Platform, SNAP), and then these images were transformed to ENVI standard format [23]. Image mosaicking and spatial subset were executed using ENVI 5.3 software (Exelis Visual Information Solutions, Inc., Boulder, CO, USA) to ensure the completeness of the lakes and reservoirs. The spectral reflectance data extracted from Sentinel-2 L2A resampled imageries was used as variables for machine learning modeling.

Satellite images were classified into water and nonwater areas by using MNDWI (Modified Normalized Difference Water Index) [50]. The MNDWI value extracted through ROIs (Region of Interest) of water was greater than 0. However, the thresholds should be respectively slightly modified in different images. The extracted water bodies were converted into polygons in a shapefile using ArcGIS 10.2 software (ESRI Inc. Redlands, CA, USA). To reduce the classification error and obtain accurate water boundary, the polygons were manually examined and corrected. The shapefiles of lakes and reservoirs were used as water masks to extract the turbidity map derived from Sentinel-2 imagery data. The data preprocessing procedure is outlined in Fig. 2.

D. Regression Algorithms

To identify the optimal method for estimating water turbidity, we picked nine algorithms and comprehensively compared their performances. These algorithms are the SR, partial least squares regression (PLSR), SVR, ELM, BP, classification and regression tree (CART), tree-based ensemble methods [including random forest (RF) and gradient boosting decision tree (GBDT)], and K-nearest neighbor regression (KNN). All analyses including water turbidity prediction and experimental hyperparameters optimization were performed in the Python 3.7 programming environment with its contributed packages. The spectral reflectance derived from Sentinel-2 MSI bands was independent variable, while the corresponding in situ water turbidity measurement was dependent variable.

[1][Online]. Available: https://earthexplorer.usgs.gov/

[2][Online]. Available: https://scihub.copernicus.eu/

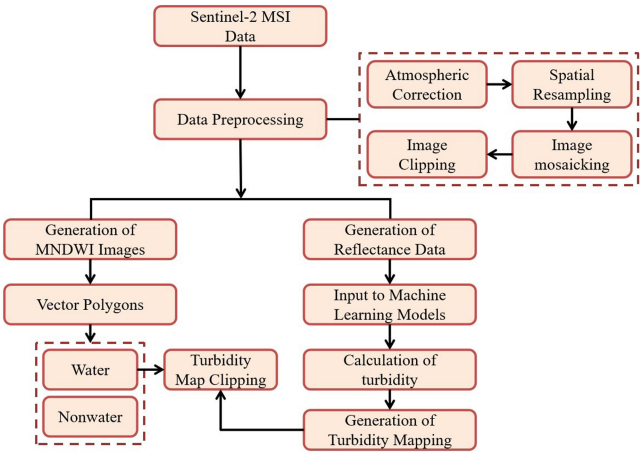

Fig. 2. Methodology flow diagram of data processing.

SR and PLSR are classical and widely used statistical regression methods. The SR method explains the relationship between response variable and predictor variables using simple mathematical function, while the PLSR method finds the latent variables to modeling the covariance relations between *X* (Sentinel-2 spectral reflectance) and *Y* (water turbidity) matrices [51], [52]. The PLSR method is suitable for the high-dimensional predictors and able to reduce the multicollinearity among X values. To establish effective regression models, the SR method requires more relevant variables, while the PLSR method needs appropriate number of components.

SVR is able to transform an original dataset [(x_1, y_1), (x_2, y_2), …, (x_n, y_n)] into a potentially higher dimensional feature space by the function Φ(x) using a kernel function [53]. The kernel function is a crucial factor for nonlinear regression tasks, and the most suitable and effective kernel function is typically selected through experience and many experiments [56]. In the new high-dimensional space, a nonlinear regression function *f*(x) can fit the dataset and able to output continuous prediction values [54], [55]. *f*(x) can be defined as follows:

$$f(x) = \omega \cdot \Phi(x) + b \tag{1}$$

where the variables ω and b represents the slope and offset of the regression function. SVR solves the following regression problem:

$$\min_{\omega,\mathrm{b},\xi,\xi^*} Q\left(\omega, \mathrm{b}, \xi, \xi^*\right) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{2}$$

subject to

$$\begin{aligned} &y_i - \omega \cdot \Phi(x_i) - b \le \varepsilon + \xi_i, \\ &\omega \cdot \Phi(x_i) + b - y_i \le \varepsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \ge 0, \mathrm{i} = 1, \ldots, \mathrm{n} \end{aligned} \tag{3}$$

where *C* is penalty factor used to control the empirical risk and confidence range, ξ, ξ^* are relaxation factors used to modify the convergence speed, $\varepsilon$is loss function applied to estimate the prediction accuracy.

The back propagation (BP) neural network and ELM are both flexible algorithms for modeling based on multilayer perceptron (MLP). Neural networks, which are made up of many artificial neurons, consist of input, hidden, and output layers. The input layer represents the input features (e.g., Sentinel-2 spectral reflectance). Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation and bias, followed by a nonlinear activation function. The output layer receives the values from the last hidden layer and transforms them into the output value (e.g., water turbidity) [57], [58].The feed-forward propagation can be defined as (4). The BP trains dataset using back-propagation in the output layer and the values of weight and bias which decide the performance of network dynamically update according to the back-propagating errors [59], [60]. The back-propagating errors for regression can be defined as (5). By contrast, the ELM is more effective and it is characterized by the constant weight and bias, which are randomly and initially assigned [61]. The type of activation function, hidden layer size, and learning rate are the key empirical parameters of the networks.

$$o_j^l = f\left(\sum_{i=1}^{S_{l-1}} o_i^{l-1}\omega_{ji}^l + b_j^l\right) \tag{4}$$

where $f(\bullet)$ is the activation function, ω_{ji}^l is the weight coefficient from the *j*th neuron in the layer *l* to the *i*th neuron in the layer $l-1$, b_j^l is the bias for the *j*th neuron in the layer *l*, o_i^{l-1} is the output of *i*th neuron in the layer $l-1$, o_j^l is the output of *j*th neuron in the layer *l*, S_{l-1} is the number of neurons in the layer $l-1$

$$e_k = \frac{1}{2}\sum_{k=1}^{M}\left(d_k - o_k\right)^2 \tag{5}$$

where d_k and o_k is the expectation output and the network output of *k*th neuron, *M* is the number of neurons in the output layer, e_k is the network prediction error used for BP.

The CART, RF, and GBDT are all tree-based machine learning methods. The RF and GBDT are promising and widely applied ensemble methods, which consist of a large number of CART that constructs binary trees by recursively dividing the features space at each node to group the similar target. For regression, the loss function often uses mean squared error (6) as the criteria to minimize for splitting each node

$$H(Q_m) = \frac{1}{N_m}\sum_{y\in Q_m}\left(y - \overline{y}_m\right)^2 \tag{6}$$

where *H()* is the loss function that mean square error, Q_m represents the datasets at node *m*, N_m is the number of samples at node *m*, *y*, and $\bar{y}_m$, respectively, represent the original target value and mean target value at node *m*.

The RF produces individual CART by bagging technology which uses bootstrap samples of dataset instead of the total original data and yields final result using the averaged prediction of the individual trees [62]–[64]. However, the GBDT is different from the RF due to the use of boosting technology to sequentially generated individual CART based on the residuals

of the preceding tree [65], [66]. For the CART method, the key parameters (including the maximum depth of the tree, the minimum number of samples splitting an internal node, and the minimum number of samples being at a leaf node) determine the tree structure and affect the performance of the method [67]. For the ensemble methods, except for the same parameters as the CART, the number of trees and learning rate are the important hyperparameters. Moreover, the tree-based ensemble can be utilized to evaluate the relative importance of features with respect to the predictability of target variable. The feature importance is defined to be the mean decrease in impurity when a single feature value is randomly shuffled. Due to the variation of feature set, the precision decrease indicates how much the model depends on the feature [63], [65]. The out-of-bag (OOB) error as the accuracy evaluation criteria is calculated by using the about 37% samples (called OOB data) without training model. The variable importance can be defined as (7)

$$V(X^j) = \sum_{t=1}^{N} (e_t^j - e_t)/N \tag{7}$$

where e_t represents the OOB error, e_t^j represent the OOB error after shuffling the *j*th feature X^j, N is the number of CARTs in the ensemble model, V is the importance of feature X^j.

The KNN is a simple and easy method to predict for continuous data with multivariate [68], [69]. The prediction result of each testing sample is computed based on the weighted average of the response variable of the k nearest samples in the training set, where k is an integer value specified by user [70]. The square of distance between training and testing sample in each feature space is calculated using a given distance metric. Then, the weight is defined as the inverse of the square root of the distance sum in all feature space. The k value and distance metric are important parameters of the KNN method, which affect the performance and efficiency of the method.

E. Hyperparameters Optimization

Hyperparameters are not directly learnt within models. These parameters are provided when constructing a model and may be optimized by automatically searching the hyperparameter space. The performance and efficiency of models can be dramatically improved by passing appropriate arguments to the model developer. Some important and influential hyperparameters for modeling can be optimized using grid search strategy, which exhaustively considers all parameter combinations for given values [71]. This search strategy needs to create a multidimensional grid in the hyperparameter space. Each dimension of grid represents a kind of hyperparameter to be optimized. Each point grid represents a parameter combination value. The machine learning process is repeated using different parameter combinations until the iterations are achieved. The optimal hyperparameter combination is determined according to the cross validation R^2 score. Whilst this search strategy is easy to implement and understand, it is inefficient when the number of parameters is large [72]. In this study, the hyperparameters of models we used were provided in scikit-learn based on Python 3.7. While the important parameters require optimization in detail, others can be set to their default values. In order to improve computational efficiency and precisely search the optimal parameter combination, we empirically specified a wide initial range of hyperparameter values for coarse searching, and then narrowed the range and designed an adaptive search step for fine searching based on the cross-validation score from initial searching.

F. Model Validation

Here model development and performance evaluation are important steps. The parameters of algorithms and the predictive performance were evaluated using a fivefold cross validation approach. The model accuracies are determined by the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute errors (MAE) using validation samples. Briefly, RMSE and MAE describe the overall error of prediction, while R^2 quantifies the amount of variation explained by the developed relationships [73]. Models which produce the highest R^2 value with the lowest RMSE and MAE are considered suitable for estimating water turbidity. These measurements are defined as follows:

$$\mathrm{RMSE} = \sqrt{\sum\nolimits_{i=1}^{\mathrm{n}} (y_i - y_i')^2 \Big/ n} \tag{8}$$

$$\mathrm{R}^2 = 1 - \sum\nolimits_{i=1}^{\mathrm{n}} (y_i - y_i')^2 \Big/ \sum\nolimits_{i=1}^{\mathrm{n}} (y_i - \bar{y}_i)^2 \tag{9}$$

$$\mathrm{MAE} = \frac{1}{n} \sum\nolimits_{i=1}^{n} |y_i - y_i'| \tag{10}$$

where y_i and y_i' are the observed and predicted value for the *i*th observation; $\bar{y}$ is the average observed value; n is the number of calibration and validation samples.

III. Results

A. Descriptive Statistics

The statistics of water turbidity samples are shown in Table I. The dataset shows a large water turbidity range (0.83–112.26 NTU), with a mean of 32.43 NTU and a standard deviation of 28.68 NTU. The SXKH, HMTR, CGL, GXKL, YLL, NYR, TPCR, and TMJR (see the full name in Table I) exhibit higher turbidity values (average: Above 30 NTU). These lakes and reservoirs are located in a relatively flat region, where the average elevation is about 100 m. Due to the ample water resource and flat landscape, the fertile land around these lakes is used for farming. The extra human activities could be one of the reasons that the water is turbid. Further, CGL, YLL, NYR are saline soda lakes with water turbidity varying from 42.04 to 49.22 NTU. The water turbidity decreased with the rise of terrain. YFR exhibits the lowest water turbidity (average: 2.40 NTU), with the highest elevation (767.90 m). YFR is located in the mountains and human activity has less impact on the water quality. Among all the sampled lakes and reservoirs, GXKL covers a largest area of 4412.2 km^2, with higher turbidity (average: 48.5 NTU). While the area of XXSR was smallest (30.2 km^2), with lower turbidity (average: 10.76 NTU). However, it does not mean that the smaller area of water body exhibits lower turbidity.

TABLE I.
DESCRIPTIVE STATISTICS OF THE SAMPLES USED IN THE PREDICTION OF THE WATER TURBIDITY

Sites	SD	SN	LA (km^2)	LE (m)	MIN (NTU)	MAX (NTU)	MEAN (NTU)	STD (NTU)
BSR	2018-10-14	28	105.5	593.05	0.94	8.26	3.63	2.20
CGL	2018-10-11	12	307.1	131.13	25.91	80.91	49.22	14.39
ELL	2018-7-17	11	170	268.13	5.96	11.46	9.08	1.81
GYGR	2018-10-11	7	61.8	435.28	1.22	4.06	2.45	1.05
GXKL	2018-10-18	30	4412.2	122.10	14.23	80.70	48.50	12.32
HRR	2018-10-12	9	98.6	384.10	1.46	4.49	2.55	0.84
HMTR	2018-7-31	3	38.8	74.07	63.53	66.68	65.24	1.30
JPL	2018-10-20	10	91.5	438.32	10.48	71.69	20.81	17.32
NYR	2018-10-12	9	220.1	130.64	13.58	80.28	42.04	23.16
SHL	2018-10-21	6	216.2	297.09	12.50	30.86	6.45	19.15
SXKL	2018-7-30	29	162.1	122.10	32.78	112.26	72.46	22.95
TMJR	2018-9-13	1	53.6	167.71	30.61	30.61	-	-
TPCR	2018-6-15	10	34.3	193.83	19.69	61.41	41.45	10.57
XXSR	2018-10-14	4	30.2	328.72	9.18	12.69	10.76	1.27
YFR	2018-10-13	12	69.3	767.90	0.83	9.61	2.40	2.26
YLL	2018-10-11	6	180.8	130.51	34.61	51.20	43.66	5.49
Total		187	-	-	0.83	112.26	32.43	28.68
Calibration set		131	-	-	0.83	112.26	32.86	29.00
Validation set		56	-	-	0.94	103.43	31.42	27.89

Note: SD = Sampling dates, SN = Sample numbers, LA = Lake area, LE = Mean elevation around lake, MIN = Minimum value, MAX = Maximum value, MEAN = Average value, STD = Standard deviation value, BSR = Baishan Reservoir, CGL = Chagan Lake, ELL = Erlong Lake, GYGR = Guanyinge Reservoir, HMTR = Hamatong Reservoir, HRR = Huanren Reservoir, JPL = Jingpo Lake, NYR = Nanyin Reservoir, SHL = Songhua Lake, TPCR = Taipingchi Reservoir, TMJR = Tumuji Reservoir, XXSR = Xingxingshao Reservoir, SXKL = Small Xingkai Lake, GXKL = Great Xingkai Lake, YLL = Yueliang Lake, YFR = Yunfeng Reservoir; -, denotes not available.

The area of HMTR and TPCR are below 40 km^2, with higher water turbidity (HMTR average: 65.24 NTU, TPCR average: 41.45 NTU). Water turbidity may be influenced by several nature and anthropogenic factors. We separated the total 187 samples into two subsets, i.e., a calibration subset with 131 samples and a validation subset with 56 samples. Water turbidity varied from 0.83 to 112.26 NTU, with a mean of 32.86 NTU and a standard deviation of 29.00 NTU in the calibration subset. It is similar to the validation samples that ranged from 0.94 to 103.43 NTU, where the mean and standard deviation were 31.42 and 27.89 NTU, respectively. From above, we can see that the calibration and validation subsets are quite similar.

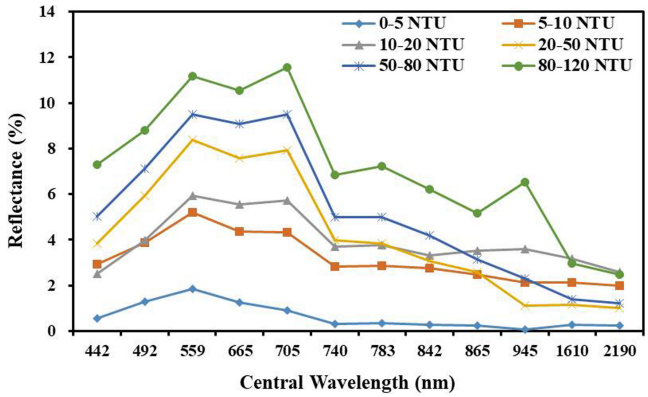

Fig. 3. Mean spectra reflectance at Sentinel-2 bands for sampling stations with various water turbidity ranges.

B. Spectral Response to Turbidity Variation

The mean values of the reflectance of Sentinel-2 spectra for water turbidity in different value ranges are shown in Fig. 3. Overall, the spectral curve was similar in every value range. For water with higher turbidity, the mean spectral reflectance was also higher. In 0–5-NTU regions, the spectral reflectance peak with lowest value was 1.9%; while in 80–120-NTU regions, the peak with highest value was 11.6%. The reflectance increased monotonously from 442 to 559 nm, and gradually decreased from 705 to 2190 nm. Two spectral peaks are at 559 and 705 nm. Moreover, the spectral variability as a response to water turbidity is more obvious from 492 to 705 nm. The significant difference of reflectance gradually reduced after 745-nm band between 5 and 80 NTU. The result here is similar to the analysis of GOCI imagery by Wang *et al.* [30], which reported that the mean spectrum of reflectance first increases from 412 to 680 nm, and then gradually decreases from 680 to 865 nm. The spectral response is the weakest from 945 to 2190 nm, which the water is characterized by powerful absorption of electromagnetic spectrum.

C. Correlation of Spectral Parameters and Water Turbidity

The Pearson's correlation coefficients are used to describe the relationship between water turbidity and spectral reflectance,

TABLE II.
PEARSON'S CORRELATION BETWEEN WATER TURBIDITY AND SPECTRAL REFLECTANCE VARIABLES OF SINGLE-BAND AND BAND COMBINATION

Variables	r	Variables	r
B1	0.76**	B3+B5	0.83**
B2	0.80**	B5-B8	0.77**
B3	0.81**	B3*B5	0.83**
B4	0.79**	B5/B3	0.62**
B5	0.83**	(B3-B5)/(B3+B9)	-0.66**
B6	0.66**	(B3*B5)/(B4+B12)	0.86**
B7	0.64**	(B8/B4)/(B8+B9)	-0.61**
B8	0.63**	(B2-B12)/(B2*B3)	-0.51**
B8A	0.43**	(B5+B8)/(B3*B5)	-0.62**
B9	0.34**	(B3+B5)/(B2/B3)	0.80**
B11	0.12	B5/(B3+B12)	0.70**
B12	0.13	(B3-B5)/B3	-0.62**

Note: r is Pearson's correlation coefficient; **is the significant at 1% probability; B1-B12 represented the Sentinel-2 MSI band.

and all outcomes are significant at the 99% confidence level. The analysis of two kinds of spectral variables (i.e., the original Sentinel-2 spectral reflectance of each band, and the mathematic form calculated by all possible combinations from Sentinel-2 twelve bands) are shown in Table II. In terms of original spectral bands, water turbidity was strongly linear correlated with most of the spectral band variables, except with B11 (r = 0.12) and B12 (r = 0.13), which were the short wave infrared bands from 1610 to 2190 nm. The correlation coefficients were larger than 0.76 from B1 to B5. B5 exhibit the highest value of 0.83. For combination forms of bands, all of the variables were significantly correlated, with higher correlation coefficients compared to the single-band variables. It corresponded to the previous study that the calculation using more bands was able to have better performance [30]. Specifically, the band combination using B3, B4, B5, and B12 exhibited the highest correlation coefficient, with the r value of 0.86. It better described the strongest correlation between water turbidity and spectral reflectance.

D. Comparison of Regression Algorithms

1) Hyperparameters Optimization: We used cross-validation grid search for hyperparameters optimization. The values of hyperparameters were determined as the optimal results among all combinations of parameters according to the R^2 and RMSE values of cross validation in the validation subset. Table III shows the type of kernel function, gamma (kernel coefficient), and C (regularization parameter) optimized in SVR. They determine the function for transforming into high-dimensional space and may resist overfitting. For the RBF function, we used 1 for gamma value and 4 for C value. They are the optimal parameters for estimating water turbidity in SVR ($R^2 = 0.80$ and RMSE = 0.11 -NTU).

Both ELM and BP contain the crucial parameters (including activation functions and the number of hidden-layer neurons). The small number of hidden-layer neurons was not enough to train the neural networks and caused the decrease of model accuracy. In contrast, a large number of hidden-layer neurons might cost the operation efficiency. The results show that the numbers of ELM and BP were 25 and 50, respectively, which produced relatively stable cross-validation accuracy. The rectified linear unit (relu) function was better for BP, while the sigmoid function was more appropriate for ELM. These models with hyperparameters generated better results ($R^2 = 0.76$ and RMSE = 0.12 NTU for ELM; $R^2 = 0.71$ and RMSE = 0.13 NTU for BP).

The GBDT and RF are both CART-based ensemble methods and these three methods contain similar hyperparameters due to the decision tree method, including the minimum number of samples being at a leaf node and the maximum depth of the individual regression trees. Optimization of the number of trees is important for GBDT and RF. In Table III, the minimum number of samples for leaf node in CART, GBDT, and RF method were 10, 3, and 1, respectively. The maximum depths of CART, GBDT, and RF were 8, 3, and 9, respectively. The numbers of trees of GBDT and RF were quite close at 600 and 720, respectively. In GBDT and RF, the square root of the number of total input variables were used as the number of split features at each node. In terms of optimal accuracy, the GBDT ($R^2 = 0.79$, RMSE = 0.11 NTU) and RF ($R^2 = 0.76$, RMSE = 0.12 NTU) are similar and better than the CART ($R^2 = 0.62$, RMSE = 0.15 NTU). The distance metric for weights and the number of neighbors are the hyperparameters in KNN. The results indicate that the distance using Minkowski metric and 5 for neighbor numbers generated desirable simulation results ($R^2 = 0.79$ and RMSE = 0.11).

2) Statistic Regression Models: As illustrated in Table IV, the statistic regression models including two SR models and PLSR were used for turbidity modeling. The independent variables were selected from the variable subset (see Table II). The regression models using selected variables generated good fit between predicted and observed water turbidity. B3 and the band combination using B3, B5, B4, and B12 were selected variables, with r^2 values of 0.79 and 0.86, respectively (Table IV). According to the regression equation, exponential function was identified as the best mathematical model for water turbidity prediction. As a comparison, the model with band combination (model SR2) outperformed the model with single-band (model SR1) by producing better calibration result. In validation mode, SR2 also yielded better results, with R^2 value of 0.73 and lower RMSE and MAE values [Table IV and Fig. 4(b)]. The PLSR with five principal components gave the highest validation accuracy, with R^2 value of 0.79 [Fig. 4(c)] and lowest RMSE and MAE for both calibration and validation subsets (Table IV).

3) Nonlinear Machine Learning Models: Table V and Fig. 5 present the prediction accuracies of different machine learning methods. Considering the calibration quality, all methods gave satisfactory results and the R^2 values were higher than 0.8. As expected, the GBDT and RF yielded highest R^2 values of 0.99 and 0.98, respectively, with the lowest RMSE and MAE values (Table V). It should point out that the KNN produced R^2 value of 1 because the calibration samples were merely used to calculate the distance between validation samples (different from other methods that should be used for modeling). In terms

TABLE III.
RESULTS FOR THE FOR HYPERPARAMETERS OPTIMIZATION USING CROSS-VALIDATION GRID SEARCH

Model	Hyper-parameters	Initial references		Optimal value	Validation	
		Search range	Step		$RMSE_{CV}$	R^2_{CV}
SVR	Kernel function	poly, rbf, sigmoid		rbf	0.11	0.80
	Gamma	$[2^{-5}, 2^5]$	$2^{1/2}$	1		
	C	$[2^{-5}, 2^5]$	$2^{1/2}$	4		
ELM	Activation function	radbas, sigmoid, tanh		sigmoid	0.12	0.76
	Hidden layer size	[5, 50]	1	25		
BP	Activation function	Logistic, sigmoid, relu		relu	0.13	0.71
	Hidden layer size	[5, 50]	1	50		
	Learning rate	0.001, 0.01, 0.1		0.01		
CART	Min samples leaf	[5, 100]	5	10	0.15	0.62
	Max depth	[2, 15]	1	8		
GBDT	Tree number	[100, 1500]	100	600	0.11	0.79
	Learning rate	0.01, 0.03, 0.05		0.03		
	Min samples leaf	[1, 50]	1	3		
	Max depth	[3, 19]	2	3		
RF	Tree number	[20, 1500]	20	720	0.12	0.76
	Min samples leaf	[1, 50]	1	1		
	Max depth	[3, 19]	2	9		
KNN	Distance metric	Euclidean, Manhattan Minkowski		Minkowski	0.11	0.79
	Neighbors number	[1, 20]	1	5		

TABLE IV.
PERFORMANCE SUMMARY OF SIMPLE REGRESSION METHODS AND PLSR

Model	Variables	Equation	r^2	Calibration			Validation	
				MAE	RMSE	R^2	MAE	RMSE
SR1	B3	$y=1.559e^{35.533x}$	0.79	12.33	19.13	0.63	12.26	18.11
SR2	(B3*B5)/(B4+B12)	$y=1.879e^{37.745x}$	0.86	12.22	20.13	0.65	10.94	16.89
PLSR	5 PC			9.91	13.63	0.78	10.04	13.35

Note: r^2 represented the degree of fitting between observed and predictive data; PC represented the principal component using in PLSR.

TABLE V.
PERFORMANCE SUMMARY OF NONLINEAR MACHINE LEARNING MODELS

Model	Calibration			Validation	
	MAE	RMSE	R^2	MAE	RMSE
SVR	7.99	10.52	0.87	9.76	13.26
ELM	9.05	12.44	0.82	9.00	13.20
BP	8.89	12.64	0.81	10.29	14.12
CART	7.84	11.97	0.83	9.67	14.57
GBDT	1.22	1.82	0.99	6.71	9.90
RF	3.00	4.70	0.98	6.84	10.54
KNN	0.00	0.00	1.000	6.88	11.13

of validation, the CART performed poorly, with the lowest R^2 value of 0.73 and the highest RMSE and MAE values (RMSE = 14.57 NTU, MAE = 9.67 NTU) [Table V and Fig. 5(d)]. The ELM, GBDT, RF, and KNN yielded high R^2 values (all above 0.8). Especially, the

GBDT and RF produced better R^2 values of 0.88 and 0.86, respectively, with the lower RMSE and MAE values [Table V, Fig. 5 (e) and (f)]. The KNN produced a slightly lower R^2 value of 0.85, with the RMSE value of 11.13 NTU, and the MAE value of 6.88 NTU [Table V and Fig. 5(g)].

E. Variables Importance and Models Optimization

The RF and GBDT as tree-based ensemble models can be used to evaluate the relative importance of features with respect to the predictability of target variable [74]. The higher the importance fractions produce, the more important the variable is. As shown in Fig. 6, the ranking of variables was slightly different. For the RF model, the sequence of the variable importance fractions above 0.1 were B3, B5, B4, and B2 (sorted in ascending order).

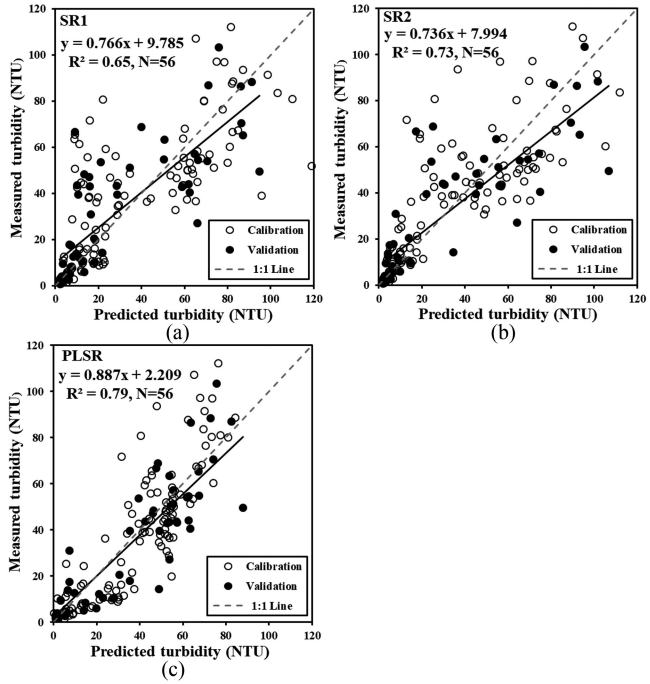

Fig. 4. Model performance comparison between simple linear regression method and PLSR model. (a) Simple linear regression model using B3. (b) Simple linear regression model using band combination. (c) PLSR model. The calculations of band combinations are shown in Table IV.

TABLE VI.
PERFORMANCE SUMMARY OF GBDT AND RF MODEL WITH SELECTED VARIABLES

Model	Calibration			Validation	
	MAE	RMSE	R^2	MAE	RMSE
GBDT_VS	1.65	2.50	0.99	7.28	10.79
RF_VS	3.26	4.99	0.97	7.49	10.90

Similarly, the order was B2, B5, B4, B3, and B6 in GBDT model. For the RF and GBDT models, four spectral variables (including B2, B3, B4, and B5) were more important than the other variables. Further, B9, B11, and B12 exhibited lower importance fractions, with the value of approximately 0.03.

To improve the efficiency of models and reduce noise variables, variables selection (VS) was implemented according to the importance of variables. All variables were sorted (in descending order) by the importance fractions. One variable was added into the model for training per iteration, the variables were selected when the cross-validation accuracy trended to be stable. The cross-validation accuracy curves of the two models were similar (Fig. 7), the variations of R^2 and RMSE were significant before five variables were added to the model. As shown in Table VI and Fig. 8, in calibration subset, the two models provided good results and the R^2 values exceeded 0.95. During model validation, the GBDT (R^2 = 0.86, RMSE = 10.79 NTU, MAE = 7.28 NTU) and RF (R^2 = 0.85, RMSE = 10.90 NTU, MAE = 7.49 NTU) models with five variables yielded good results (Table VI and Fig. 8).

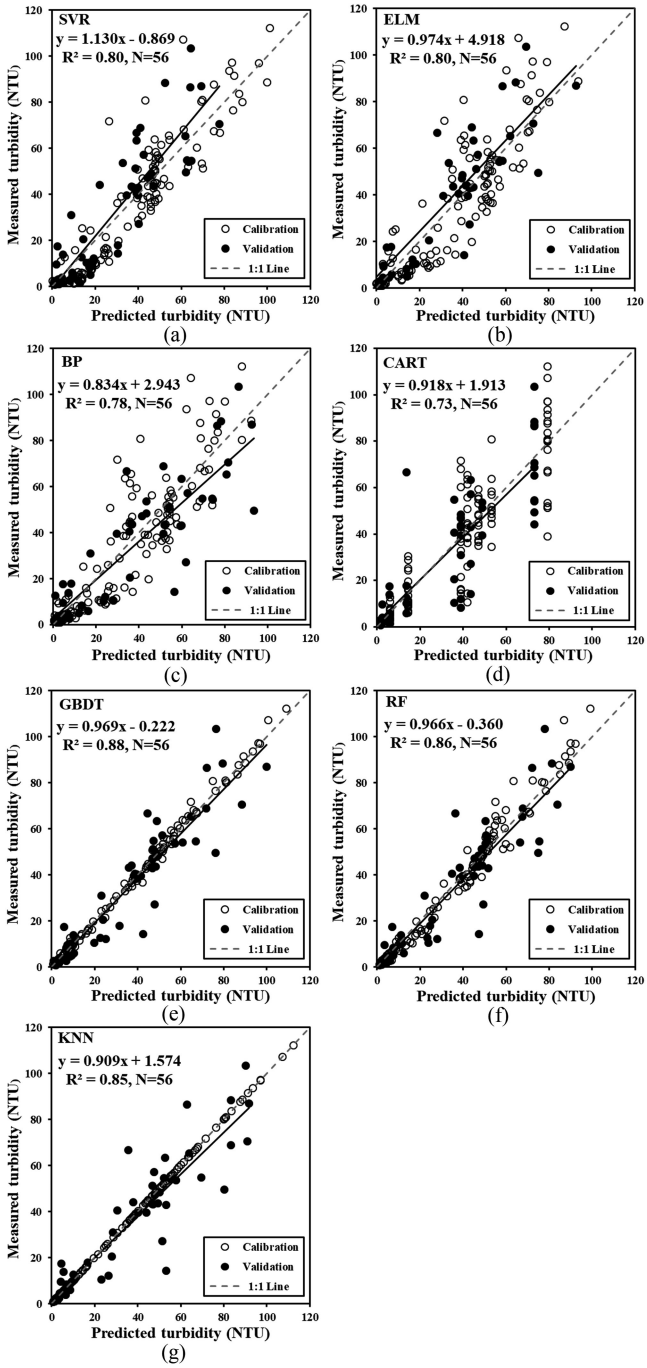

Fig. 5. Model performance comparison of nonlinear regression methods. (a) SVR model. (b) ELM model. (c) BP model. (d) CART model. (e) GBDT model. (f) RF model. (g) KNN model.

F. Water Turbidity Mapping

The water turbidity map was produced by applying the models with good predictability. Two kinds of predictive algorithms were implemented, which included the nonlinear methods (GBDT and RF) and linear method (PLSR). Further, in order to evaluate the predictability of tree-based ensemble methods with selected variables, the RF and GBDT with selected five important variables were applied to mapping water turbidity.

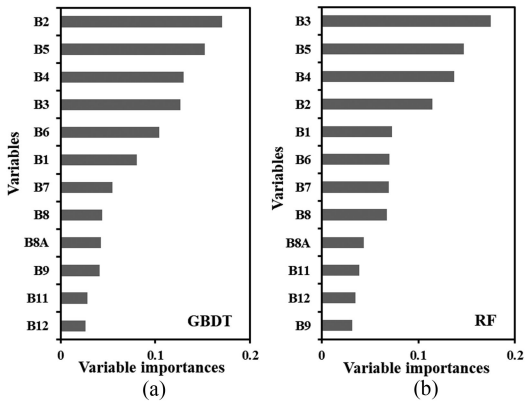

Fig. 6. Variables importance calculated by the mean decrease in impurity in GBDT and RF. (a) Variables importance ranking in GBDT. (b) Variables importance ranking in RF.

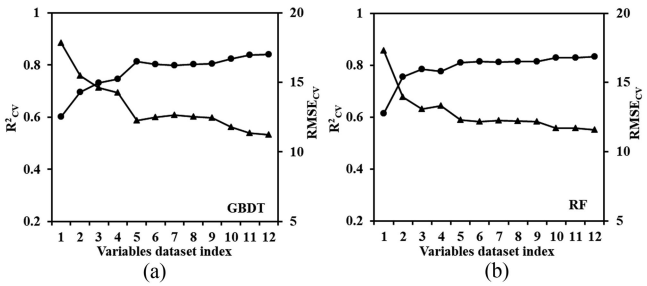

Fig. 7. Cross-validation accuracy of GBDT and RF using one variable in per iteration. (a) R^2 and RMSE variation of GBDT. (b) R^2 and RMSE variation of RF.

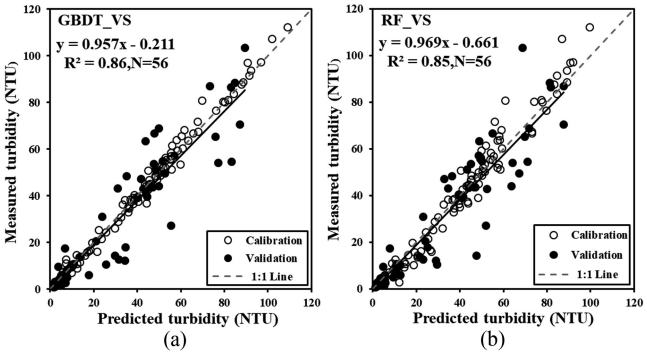

Fig. 8. Model performance comparison of RF and GBDT with selected variables. (a) GBDT model with top 5 selected variables. (b) RF model with top 5 selected variables.

Figs. 9 –12 demonstrate that the extracted results are reliable and consistent with the measured data. For the Baishan Reservoir, the water turbidity ranged 1.6–10.8 NTU based on the GBDT and 1.7–14.2 NTU based on RF model; while the variation was larger in the PLSR (0.2–16.4 NTU) (Fig. 9). The higher turbidity was observed in the central part of the reservoir with turbidity above 7.5 NTU. Due to the villages and cropland located in close proximity to water body, domestic and agricultural water might inflow into reservoir, which brings in turbidity, reducing

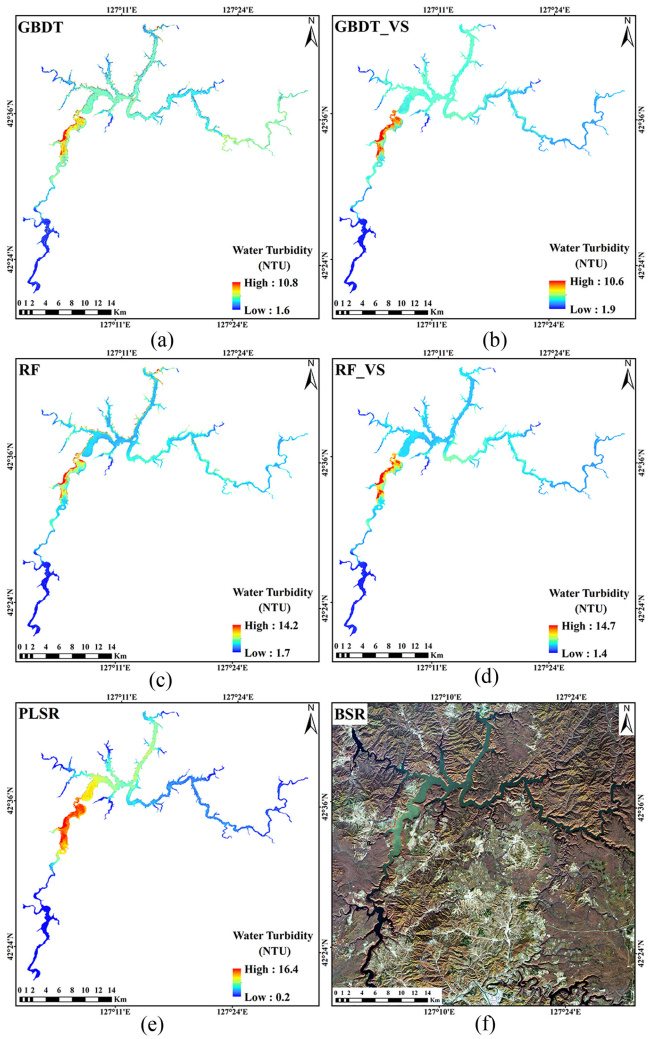

Fig. 9. Water turbidity mapping result for Baishan Reservoir (BSR) using different models. (a) GBDT model. (b) GBDT model with selected variables. (c) RF model. (d) RF model with selected variables. (e) PLSR model. (f) True color remote sensing image.

water quality. For the Chagan Lake, the water turbidity was higher in RF model (33.4–68.5 NTU) and varied widely (29.6–81.3 NTU) when the GBDT was used. The turbidity varied the most (17.4–111.6 NTU) based on PLSR (Fig. 10). For the Erlong Lake, the water turbidity were similar in RF (8.7–67.6 NTU) and GBDT (7.7–65.9 NTU). In the PLSR model, the variation was wider (17.5–73.5 NTU) (Fig. 11). Turbidity of Chagan Lake and Erlong Lake was highest. These lakes located in the Songnen Plain of Northeast of China where the agricultural and human activities were more frequent. Mesotrophic or eutrophic water causes algae multiply greatly and turbidity increased. In addition, the higher turbidity was observed in the downstream of Erlong Lake, it might because that silt has silted up the sluggish lake area. For the Huanren Reservoir, the water turbidity ranged 1.6–7.1 NTU based on RF; 0.3–8.2 NTU using GBDT; and 0.18–11.9 NTU in PLSR (Fig. 12). Higher turbidity for one quarter of Huanren Reservoir water was close to the dam near the towns, and another area with high turbidity was downstream

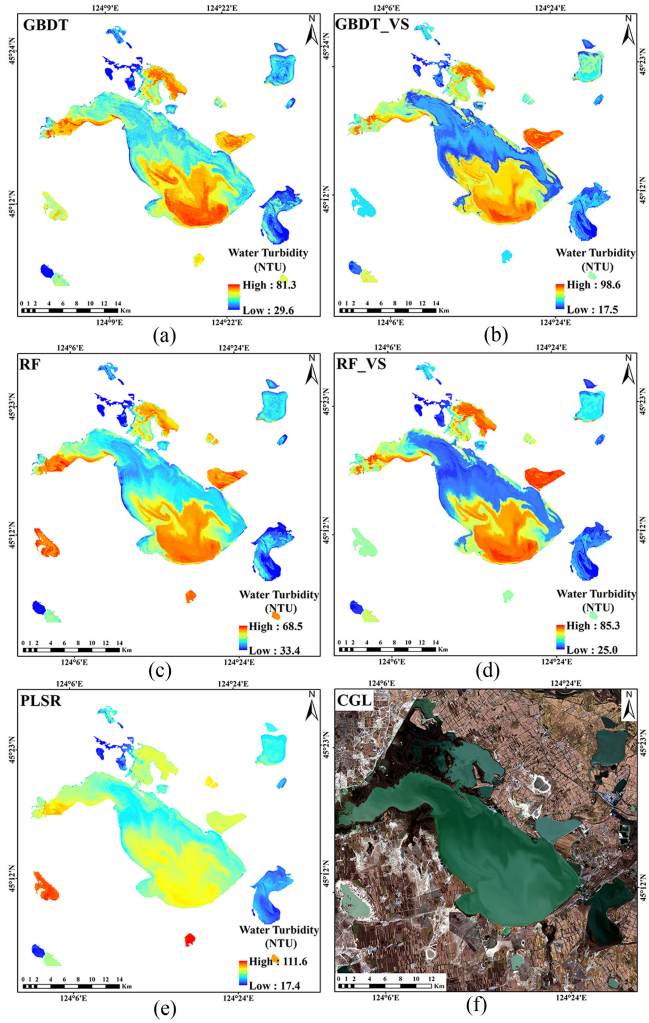

Fig. 10. Water turbidity mapping result for Chagan Lake (CGL) using different models. (a) GBDT model. (b) GBDT model with selected variables. (c) RF model. (d) RF model with selected variables. (e) PLSR model. (f) True color remote sensing image.

of reservoir. By comparing the mapping results of these lakes and reservoirs based on RF and GBDT with selected variables, the water turbidity had a similar variation when the model used all variables.

IV. Discussion

Whilst many statistical and machine learning regression techniques can be used for estimating water turbidity, it is necessary to contribute to exploring their performance in a comparative perspective. This study made some effort to comprehensively evaluate the techniques for water turbidity based on remote sensing data.

A. Variable Importance

In terms of single spectral bands of Sentinel-2 data, the Pearson's correlation coefficients showed that the visible bands from 442 to 559 nm were highly related to water turbidity, especially, 559 and 705 nm. On the contrary, the short wave infrared bands at 945, 1610, and 2190 nm were less related to water turbidity. It was coincident with the peak and trough patterns of the spectral reflectance curve. For band combination variables, the band ratio calculated by the spectral bands with significant differences can describe the relationship between reflectance and water turbidity.

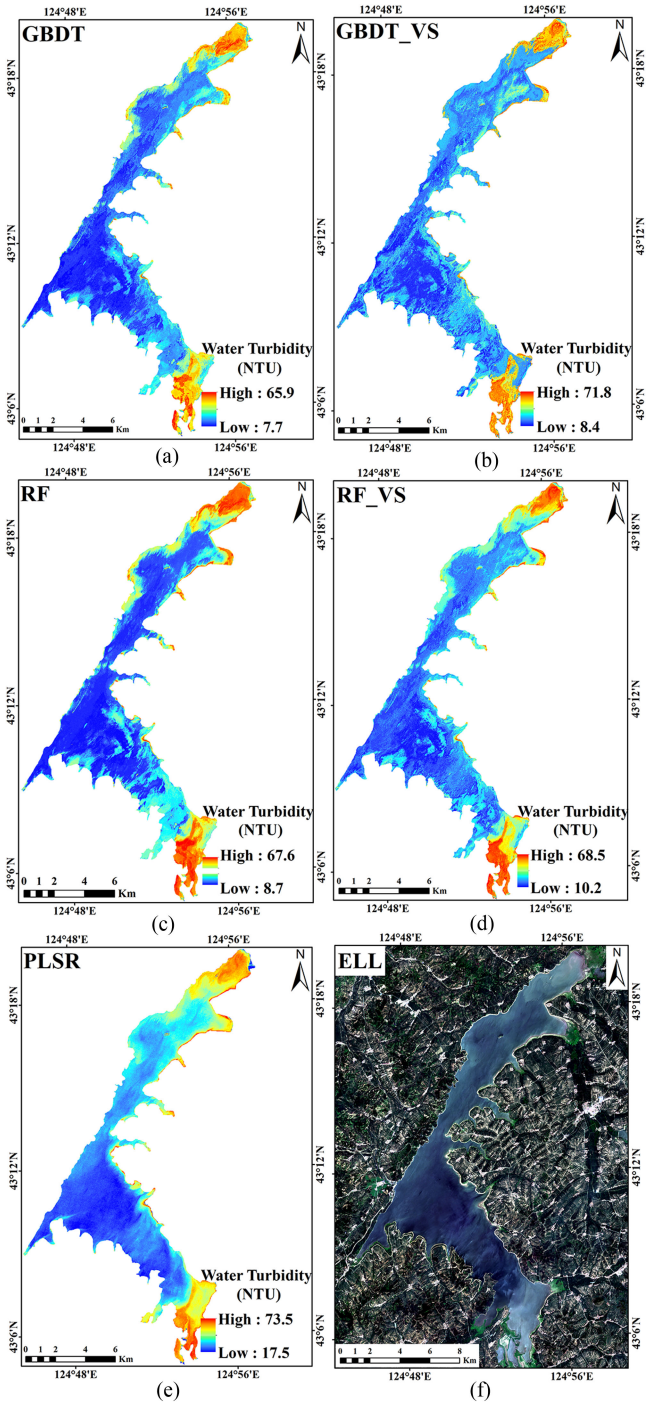

Fig. 11. Water turbidity mapping result for Erlong Lake (ELL) using different models. (a) GBDT model. (b) GBDT model with selected variables. (c) RF model. (d) RF model with selected variables. (e) PLSR model. (f) True color remote sensing image.

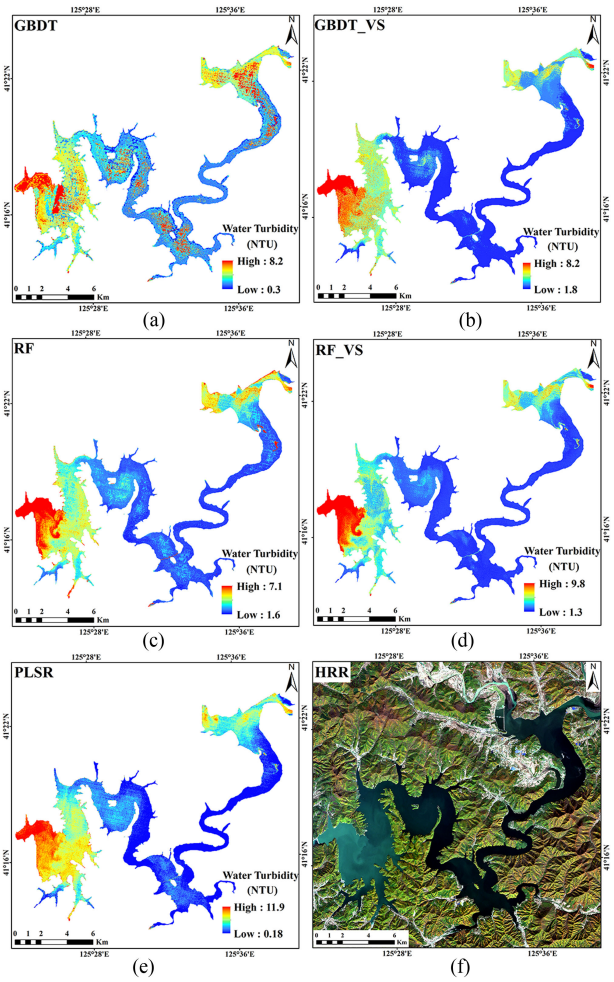

Fig. 12. Water turbidity mapping result for Huanren Reservoir (HRR) using different models. (a) GBDT model. (b) GBDT model with selected variables. (c) RF model. (d) RF model with selected variables. (e) PLSR model. (f) True color remote sensing image.

Moreover, the impurity-based feature importance from the GBDT and RF models were also used for evaluating the spectral variables. According to the results, the rank of variables importance were similar for both models, which B2, B3, B4, and B5 were the more important variables for estimating water turbidity. By comparing the Pearson's correlation coefficients and spectral characteristics, the variables selected by the impurity-based feature importance were reliable. Moreover, the spectral signals of B3, B4, and B5 were related to the suspended sediment and phytoplankton in turbid water. Water has absorption property in B2. The selected variables are coincident with the actual physical properties [75]. In other words, the GBDT and RF methods might be able to deal with the variables and extract the characteristic information.

B. Comparison of Regression Methods

We compared the performances of statistic and machine learning regression techniques. During the modeling process, statistic models were easier to build, but the prediction accuracy of water turbidity was not satisfactory. Specifically, the PLSR outperformed the SR model with good predictability. Machine learning regression models were difficult to build. To obtain a model with good performance, more efforts would be needed in searching for suitable hyperparameter space and training the model. However, the machine learning models have strong ability to extract variables and describe the nonlinear relationship between spectral variables and water turbidity.

The tree-based ensemble regression methods (i.e., GBDT and RF) yielded results much better than other methods. It means that the ensemble methods which combined the several base estimators (e.g., CART) can produce better prediction results and improve the generalization and robustness over a single estimator. By comparing the GBDT and RF, GBDT slightly outperformed RF in the task of water turbidity estimation. The GBDT is a typical boosting algorithm, which comprised of numerous CART sequentially constructed from "pseudo" residuals (negative gradient of the loss function) [76], [77]. Due to the decrease of deviance gradually, the GBDT can predict more accurately. By contrast, the RF is a promising bagging algorithm, which integrated several CART. The mean predictive result was calculated by all trees and the variance significantly decreased. Among the evaluated methods, the CART was the worst method for estimating water turbidity. It even produced an accuracy lower than the PLSR. The key disadvantage of ensemble methods is the high computation complexity. To enhance the efficiency, we constructed models using selected variables depending on their importance. The results demonstrated that the accuracy was slightly lower than before. We deduced that using ensemble models with selected variables was effective to reduce model complexity and maintain relatively high accuracy.

The SVR and multilayer perceptron methods (ELM and BP) produced less accurate results. The results indicated that it was difficult for SVR to search for an appropriate function to accurately predict water turbidity. Thus, the BP and ELM are relatively inefficient. Although weight optimization is the key step for multilayer perceptron methods, the neural networks are initialized by random weights which can easily cause local minimization problem. It was hard to optimize the network to describe the relationship between water turbidity and spectral reflectance. The KNN was identified as another potential regression technique in our comparison, and it was easier to build. Moreover, the water turbidity mapping results showed the estimation ability of all models. The GBDT, RF, and PLSR demonstrated better performance among nonlinear and linear methods. The water turbidity variation was similar and reliable in the GBDT and RF models, whereas the variation was wider in the PLSR model.

C. Importance of Hyperparameters Optimizations

It is important to optimize hyperparameters for building machine learning models. With good hyperparameters, models can be improved and become more efficient with little risk of overfitting. GBDT and RF are both tree-based ensemble methods. However, in these two methods, the maximum depth of the tree and the minimum number of samples being at a

leaf node are obviously different. These two hyperparameters decide the size of base learner (CART). Due to the simple structure of individual CART, the GBDT method can be able to effectively avoid overfitting that is easily produced. ELM and BP are multilayer perceptron algorithms. These two methods obtain different activation function and the number of artificial neurons in the hidden layer. ELM uses fewer hidden layer neurons and may be able to learn faster and improve model efficiency. The optimal value of C and gamma can lead to SVR model with better generalization ability and higher prediction accuracy. For KNN method, the distance metrics and k value directly influence the prediction accuracy and model efficiency. Further, we note that the grid search strategy is a practical method to search the best combination of hyperparameters. The application of important parameters selected from a large number of hyperparameters can effectively improve the robustness and generalization of models.

V. Conclusion

This study showed the feasibility of estimating the water turbidity using Sentinel-2 imagery in a large study area. We combined the Pearson's correlation analysis and the feature importance from tree-based ensemble methods. We identified B2, B3, B4, and B5 as the most significant spectral variables for estimating water turbidity. This study demonstrated that the machine learning methods have a strong advantage in water turbidity prediction compared to linear regression. For the tree-based ensemble methods, the GBDT slightly outperformed the RF with the same dataset, and significantly better than other machine learning methods. The CART was the worst for turbidity estimation. The PLSR as a statistical linear method produced results that were better than the SR algorithm. Moreover, hyperparameters optimization is a key step for building machine learning methods. The risk of overfitting can be effectively reduced by using good hyperparameters. It is effective to use grid search strategy. Overall, our results demonstrated the effectiveness and reliability of using the GBDT and RF machine regression methods to estimate the large-scale water turbidity using Sentinel-2 imagery data.

Acknowledgment

The authors would like to thank Dr. C. Fang, Y. Zhao, T. Shao, J. Ma, and M. Wang for assistance with data collection and processing. The authors would also like to thank anonymous reviewers for the constructive comments and recommendations, which are very helpful in improving and strengthening this paper.

References

[1] P. H. Gleick, "Global freshwater resources: Soft-path solutions for the 21st century," *Science*, vol. 302, no. 5650, pp. 1524–1528, Nov. 2003.

[2] Z. Duan and W. G. M. Bastiaanssen, "Estimating water volume variations in lakes and reservoirs from four operational satellite altimetry databases and satellite imagery data," *Remote Sens. Environ.*, vol. 134, pp. 403–416, Jul. 2013.

[3] J. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, Dec. 2016.

[4] K. Song *et al.*, "Quantification of lake clarity in China using landsat OLI imagery data," *Remote Sens. Environ.*, vol. 243, Jun. 2020, Art. no. 111800.

[5] I. Haddeland, T. Skaugen, and D. P. Lettenmaier, "Anthropogenic impacts on continental surface water fluxes," *Geophys. Res. Lett.*, vol. 33, no. 8, Apr. 2006, Art. no. L08406.

[6] C. Kuhn *et al.*, "Performance of landsat-8 and sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity," *Remote Sens. Environ.*, vol. 224, pp. 104–118, Apr. 2019.

[7] X. Deng *et al.*, "Remote sensing estimation of catchment-scale reservoir water impoundment in the Upper Yellow River and implications for river discharge alteration," *J. Hydrol.*, vol. 585, Jun. 2020, Art. no. 124791.

[8] E. Ayana, *Determinants of Declining Water Quality*. Washington, DC, USA: World Bank, 2019.

[9] Z. Wen *et al.*, "Quantifying the trophic status of lakes using total light absorption of optically active components," *Environ. Pollut.*, vol. 245, pp. 684–693, Feb. 2019.

[10] R. J. Davies-Colley and D. G. Smith, "Turbidity suspended sediment, and water clarity: A review," *J. Amer. Water Resour. Assoc.*, vol. 37, no. 5, pp. 1085–1101, Oct. 2001.

[11] M. Potes, M. J. Costa, and R. Salgado, "Satellite remote sensing of water turbidity in Alqueva Reservoir and implications on lake modelling," *Hydrol. Earth Syst. Sci.*, vol. 16, pp. 1623–1633, Jun. 2012.

[12] C. Kemker, 'Turbidity, total suspended solids and water clarity,' Fundamentals of Environmental Measurements, Fondriest Environmental, Ohio, OH, USA. Accessed: Jun. 13, 2014. [Online]. Available: https://www.fondriest.com/environmental-measurements/parameters/water quality/turbidity-total-suspended-solids-water-clarity/

[13] S. Constantin, D. Doxaran, and S. Constantinescu, "Estimation of water turbidity and analysis of its spatio-temporal variability in the Danube River plume (Black Sea) using MODIS satellite data," *Continental Shelf Res.*, vol. 112, pp. 14–30, Jan. 2016.

[14] B. Zhou *et al.*, "Long-term remote tracking the dynamics of surface water turbidity using a density peaks-based classification: A case study in the Three Gorges Reservoir, China," *Ecological Indicators*, vol. 116, Sep. 2020, Art. no. 106539.

[15] A. I. Dogliotti, K. G. Ruddick, B. Nechad, D. Doxaran, and E. Knaeps, "A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters," *Remote Sens. Environ.*, vol. 156, pp. 157–168, Jan. 2015.

[16] W. Wischmeier and D. Smith, "Predicting rainfall erosion losses–A guide to conservation planning," in *Agriculture Handbooks (USA)*, Beltsville, MD, USA: U.S. Dept. Agriculture, Sci. Educ. Admin., 1978, p. 62.

[17] S. Yang, J. Wang, W. Cong, Z. Cai, and F. Ouyang, "Utilization of nitrite as a nitrogen source by botryococcus braunii," *Biotechnology Lett.*, vol. 26, no. 3, pp. 239–243, Feb. 2004.

[18] T. Ouyang, Z. Zhu, and Y. Kuang, "Assessing impact of urbanization on river water quality in the Pearl River Delta economic zone, China," *Environmental Monit. Assesnts.*, vol. 120, no. 1, pp. 313–325, Jun. 2006.

[19] M. Uriarte, C. Yackulic, Y. Lim, and J. Arce-Nazario, "Influence of land use on water quality in a tropical landscape: A multi-scale analysis," *Landscape Ecology*, vol. 26, no. 8, pp. 1151–1164, Aug. 2011.

[20] M. Mccarthy, F. Muller-Karger, D. Otis, and P. Méndez-Lázaro, "Impacts of 40 years of land cover change on water quality in Tampa Bay, Florida," *Cogent Geosci.*, vol. 4, no. 1, Jan. 2018, Art. no. 1422956.

[21] L. Sipelgas, U. Raudsepp, and T. Kõuts, "Operational monitoring of suspended matter distribution using MODIS images and numerical modelling," *Adv. Space Res.*, vol. 38, no. 10, pp. 2182–2188, Dec. 2006.

[22] G. Liu *et al.*, "An OLCI-based algorithm for semi-empirically partitioning absorption coefficient and estimating chlorophyll a concentration in various turbid case-2 waters," *Remote Sens. Environ.*, vol. 239, Mar. 2020, Art. no. 111648.

[23] K. T. Peterson, V. Sagan, and J. J. Sloan, "Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing," *Giscience Remote Sens.*, vol. 57, no. 4, pp. 510–525, Mar. 2020.

[24] D. Doxaran, P. Castaing, and S. J. Lavender, "Monitoring the maximum turbidity zone and detecting fine-scale turbidity features in the gironde estuary using high spatial resolution satellite sensor (SPOT HRV, Landsat ETM+) data," *Int. J. Remote Sens.*, vol. 27, no. 11, pp. 2303–2321, Feb. 2006.

[25] D. Doxaran, J. Froidefond, P. Castaing, and M. Babin, "Dynamics of the turbidity maximum zone in a macrotidal estuary (the Gironde, France): Observations from field and MODIS satellite data," *Estuarine Coasta Shelf Sci.*, vol. 81, no. 3, pp. 321–332, Feb. 2009.

[26] J. Bustamante, F. Pacios, R. Díaz-Delgado, and D. Aragonés, "Predictive models of turbidity and water depth in the Doñana Marshes using

Landsat TM and ETM+ images," *J. Environm. Manage.*, vol. 90, no. 7, pp. 2219–2225, May 2009.
[27] C. Petus, G. Chust, F. Gohin, D. Doxaran, J. Froidefond, and Y. Sagarminaga, "Estimating turbidity and total suspended matter in the Adour river plume (South Bay of Biscay) using MODIS 250-m imagery," *Continental Shelf Res.*, vol. 30, no. 5, pp. 379–392, Mar. 2010.
[28] X. Jiang, B. Lu, and Y. He, "Response of the turbidity maximum zone to fluctuations in sediment discharge from river to Estuary in the Changjiang Estuary (China)," *Estuarine Coasta Shelf Sci.*, vol. 131, pp. 24–30, Oct. 2013.
[29] X. Shen and Q. Feng, "Statistical model and estimation of inland riverine turbidity with landsat 8 OLI images: A case study," *Environmental Eng. Sci.*, vol. 35, no. 2, pp. 132–140, Jun. 2017.
[30] S. Wang, Y. Mao, L. Zheng, Z. Qiu, M. Bilal, and D. Sun, "Remote sensing of water turbidity in the eastern China seas from geostationary ocean colour imager," *Int. J. Remote Sens.*, vol. 41, no. 11, pp. 4080–4101, Jan. 2020.
[31] I. Caballero, R. Stumpf, and A. Meredith, "Preliminary assessment of turbidity and chlorophyll impact on bathymetry derived from sentinel-2A and Sentinel-3A satellites in south florida," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 645.
[32] M. Sebastiá-Frasquet, J. A. Aguilar-Maldonado, E. Santamaría-Del-Ángel, and J. Estornell, "Sentinel 2 analysis of turbidity patterns in a coastal lagoon," *Remote Sens.*, vol. 11, no. 24, Dec. 2019, Art. no. 2926.
[33] V. K. Choubey, "Correlation of turbidity with Indian remote sensing Satellite-1A data," *Hydrological Sci. J.*, vol. 37, no. 2, pp. 129–140, Dec. 1992.
[34] D. G. Goodin, J. A. Harrington, M. D. Nellis, and D. C. Rundquist, "Mapping reservoir turbidity patterns using SPOT-HRV data," *Geocarto Int.*, vol. 11, no. 4, pp. 71–78, Sep. 1996.
[35] Z. Chen, C. Hu, and F. Muller-Karger, "Monitoring turbidity in Tampa bay using MODIS/Aqua 250-m imagery," *Remote Sens. Environ.*, vol. 109, no. 2, pp. 207–220, Jul. 2007.
[36] S. Ouillon *et al.*, "Optical algorithms at satellite wavelengths for total suspended matter in tropical coastal waters," *Sensors*, vol. 8, no. 7, pp. 4165–4185, Jul. 2008.
[37] J. Choi, Y. J. Park, B. R. Lee, J. Eom, J. Moon, and J. Ryu, "Application of the geostationary ocean color imager (GOCI) to mapping the temporal dynamics of coastal water turbidity," *Remote Sens. Environ.*, vol. 146, pp. 24–35, Apr. 2014.
[38] Z. Qiu, L. Zheng, Y. Zhou, D. Sun, S. Wang, and W. Wu, "Innovative GOCI algorithm to derive turbidity in highly turbid waters: A case study in the Zhejiang coastal area," *Opt. Exp.*, vol. 23, no. 19, pp. A1179–A1193, Sep. 2015.
[39] A. M. Kalteh, P. Hjorth, and R. Berndtsson, "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application," *Environ. Modelling Softw.*, vol. 23, no. 7, pp. 835–845, Jul. 2008.
[40] T. Chon, "Self-organizing maps applied to ecological sciences," *Ecological Inform.*, vol. 6, no. 1, pp. 50–61, Jan. 2011.
[41] S. Wang, M. Shen, Y. Ma, G. Chen, Y. You, and W. Liu, "Application of remote sensing to identify and monitor seasonal and interannual changes of water turbidity in Yellow River Estuary, China," *J. Geophys. Res. Oceans*, vol. 124, no. 7, pp. 4904–4917, Jul. 2019.
[42] R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1, pp. 46–54, Aug. 2010.
[43] K. Millard and M. Richardson, "Wetland mapping with lidar derivatives, SAR polarimetric decompositions, and Lidar-SAR fusion using a random forest classifier," *Can. J. Remote Sens.*, vol. 39, no. 4, pp. 290–307, Jun. 2014.
[44] M. Dalponte, H. O. Ørka, L. T. Ene, T. Gobakken, and E. Næsset, "Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data," *Remote Sens. Environ.*, vol. 140, pp. 306–317, Jan. 2014.
[45] L. Xu, J. Li, and A. Brenning, "A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery," *Remote Sens. Environ.*, vol. 141, pp. 14–23, Feb. 2014.
[46] K. Song *et al.*, "Remote estimation of Kd (PAR) using MODIS and landsat imagery for turbid inland waters in northeast China," *ISPRS J. Photogrammetry*, vol. 123, pp. 159–172, Jan. 2017.
[47] K. Song *et al.*, "Characterization of CDOM in saline and freshwater lakes across China using spectroscopic analysis," *Water Res.*, vol. 150, pp. 403–417, Mar. 2019.
[48] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.,* vol. 120, pp. 25–36, May 2012.
[49] *Sen2Cor Configuration and User Manual*, Telespazio VEGA Deutschland GmbH, Darmstadt, Germany, 2016.
[50] H. Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *J. Remote Sens.*, vol. 9, no. 5, pp. 589–595, Jan. 2005.
[51] S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils (Lecture Notes in Mathematics)*, vol. 973, B. Kågström, A. Ruhe, Ed. Berlin, Germany: Springer, 1983, pp. 286–293.
[52] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chimica Acta*, vol. 185, no. 1, pp. 1–17, Dec. 1986.
[53] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995, pp. 267–290.
[54] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[55] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, " Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.
[56] R. E. Fan, P. H. Chen, and C. Lin, "Working set selection using second order information for training SVM," *J. Mach. Learn. Res.*, vol. 6, pp. 1889–1918, Jan. 2005.
[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Readings in Cognitive Science*, A. Collins and E. E. Smith, Ed. San Mateo, CA, USA: Morgan Kaufmann, 1988, pp. 399–421.
[58] G. E. Hinton, "Connectionist learning procedures," *Artif. Intell.*, vol. 40, no. 1/3, pp. 185–234, Sep. 1989.
[59] S. I. Gallant, *Neural Network Learning and Expert System*, Cambridge, MA, USA: MIT Press, 1993, pp. 211–229.
[60] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, Jan. 2010.
[61] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, Dec. 2006.
[62] L. Breiman, "Pasting small votes for classification in large databases and on-line," *Mach. Learn.*, vol. 36, no. 1, pp. 85–103, Jul. 1999.
[63] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[64] L. Breiman, "Random forests, machine learning," *J. Clin. Microbiol.*, vol. 2, pp. 199–228, Jan. 2001.
[65] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[66] T. Hastie, R. J. Tibshirani, and J. H. Friedman, "The elements of statistical learning: Springer," *Elements*, vol. 1, no. 3, pp. 267–268, Jan. 2009.
[67] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees (CART)," *Biometrics*, vol. 40, no. 3, Sep. 1984, Art. no. 358.
[68] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 607–616, Jun. 1996.
[69] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 513–520.
[70] R. E. McRoberts, "Estimating forest attribute parameters for small areas using nearest neighbors techniques," *Forest Ecology Manag.*, vol. 272, pp. 3–12, May 2012.
[71] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, Feb. 2012.
[72] J. Dufour and J. Neves, "Chapter 1 - Finite-sample inference and nonstandard asymptotics with monte carlo tests and R," in *Handbook of Statistics*, H. D. Vinod and C. R. Rao, Ed. Amsterdam, Netherlands: Elsevier, 2019, pp. 3–31.
[73] X. Jin, J. Du, H. Liu, Z. Wang, and K. Song, "Remote estimation of soil organic matter content in the Sanjiang plain, northest China: The optimal band algorithm versus the GRA-ANN model," *Agr. Forest Meteorol.*, vol. 218–219, pp. 250–260, Mar. 2016.
[74] C. S. Chagas, W. de Carvalho Junior, S. B. Bhering, and B. C Filho, "Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions," *Catena*, vol. 139, pp. 232–240, Apr. 2016.
[75] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du, "Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling," *Ecological Informat.*, vol. 10, pp. 25–36, Jul. 2012.

[76] T. Dube and O. Mutanga, "Evaluating the utility of the medium-spatial resolution landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa," *ISPRS J. Photogramm.*, vol. 101, pp. 36–46, Mar. 2015.
[77] S. M. Ghosh and M. D. Behera, "Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest," *Appl. Geography*, vol. 96, pp. 29–40, Jul. 2018.

Yue Ma received the B.S. degree in geographic information science from Jilin Jianzhu University, Changchun, China, in 2013, and the Ph.D. degree in geographic information engineering from Jilin University, Changchun, China, in 2018.

She is currently a Postdoctoral Researcher for remote sensing of environment applications with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China, and a Lecturer with the School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun, China. Her research interests include remote sensing on water environment and data mining on remote sensing.

Kaishan Song received the M.S. degree in GIS and cartography from Northeast Normal University, Changchun, China, in 2002, and the Ph.D. degree in remote sensing application from Northeast Institute of Geography and Agroecology (IGA), Chinese Academy of Sciences (CAS), Changchun, China, in 2005.

He is currently a Full Professor for Remote Sensing of Environment Applications with IGA, CAS. His research interests include biooptical properties of inland waters, remote sensing of water quality, and impact of climatic and anthropogenic driving forces on water quality spatiotemporal variations with remotely sensed imagery data.

Zhidan Wen received the B.S. degree in bioengineering from Northeast Forestry University, Harbin, China, in 2007, the M.S. degree in microbiology from Shenyang Agricultural University, Shenyang, China, in 2010, and the Ph.D. degree in environmental science and engineering from the Harbin Institute of Technology, Harbin, in 2014.

She is currently with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. Her research interests include carbon cycle in water and greenhouse gas emissions from inland water bodies.

Ge Liu received the B.S. degree in geographic science from Liaocheng University, Liaocheng, China, in 2011, and the Ph.D. degree in environmental remote sensing from Nanjing Normal University, Nanjing, China, in 2017.

He is currently with the Northeast Institute of Geography and Agricultural Ecology, Chinese Academy of Sciences, Changchun, China. His research interests include water color remote sensing and optical properties in optically complex water bodies.

Yingxin Shang received the B.S. degree in environmental science from the Changchun University of Science and Technology, Changchun, China, in 2012, and the Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2020.

She is currently with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. Her research interests include water color remote sensing and optical properties in optically complex water bodies.

Lili Lyu received the master's degree in inorganic chemistry from the College of Chemistry, Northeast Normal University, Changchun, China, in 2011. She is currently working toward the Ph.D. degree in cartography and geographic information system with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun.

Her research interests include inversion of water quality parameters, evaluation of water nutritional status, measurement of water optical parameters, data mining, etc.

Jia Du received the B.S. degree in geography science and the M.S. degree in cartography and geographic information system from Northeast Normal University, Harbin, China, in 2004 and 2007, respectively, and the Ph.D. degree in cartography and geographic information system from the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China, in 2010.

He is currently with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences. His research interests include quantitative remote sensing, remote sensing monitoring of heat flux of surface water, and remote sensing monitoring of conservation tillage.

Qian Yang received the B.S. degree in geomatics engineering from Jilin University, Changchun, China, in 2010, and Ph.D. degree in geographic information system from Jilin University, Changchun, China, in 2015.

She is currently an Associate Professor of School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun, China, since August 2015. Her research interest includes the remote sensing on cryosphere and water environment.

Sijia Li received the B.S. degree in geographic science from Jilin Normal University, Changchun, China, in 2012, and the Ph.D. degree in environmental science from Northeast Normal University, Changchun, in 2020.

She is currently with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. Her research interests include eutrophication monitoring, water color remote sensing, and lake carbon cycle in optically complex water bodies.

Hui Tao received the B.S. and M.S. degrees in geographic information system from Yanbian University, Jilin, China, in 2015 and 2018, respectively. She is currently working toward the Ph.D. degree in cartography and geographic information system with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China.

Her research interests include water clarity and total suspended matter.

Junbin Hou received the M.S. degrees in physical geography from Jilin Normal University, Changchun, China, in 2017.

He is currently with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. His research interests include remote sensing on water environment.