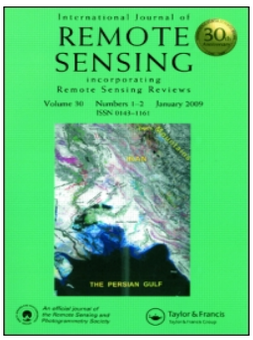

International Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery

Hongwei Guo , Jinhui Jeanne Huang , Bowen Chen , Xiaolong Guo & Vijay P. Singh

To cite this article: Hongwei Guo , Jinhui Jeanne Huang , Bowen Chen , Xiaolong Guo & Vijay P. Singh (2021) A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery, International Journal of Remote Sensing, 42:5, 1841-1866, DOI: 10.1080/01431161.2020.1846222

To link to this article: https://doi.org/10.1080/01431161.2020.1846222

INTERNATIONAL JOURNAL OF REMOTE SENSING
2020, VOL. 42, NO. 5, 1841–1866
https://doi.org/10.1080/01431161.2020.1846222

Taylor & Francis
Taylor & Francis Group

A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery

Hongwei Guo[a], Jinhui Jeanne Huang[a], Bowen Chen[a], Xiaolong Guo[b] and Vijay P. Singh[c,d]

[a]College of Environmental Science and Engineering/Sino-Canada Joint R&D Centre for Water and Environmental Safety, Nankai University, Tianjin, PR China; [b]Department of Urban Management and Eco-protection, Tianjin High-Tech Area, Tianjin, PR China; [c]Department of Biological & Agricultural Engineering, Texas A&M University, College Station, TX, USA; [d]National Water Centre, UAE University, Al Ain, UAE

ABSTRACT

Water-quality monitoring for small urban waterbodies by remote sensing gets to be difficult due to the coarse spatial resolution of remote-sensing imagery. The recently launched Sentinel-2 produces imagery with a spatial resolution of 10 × 10 m and a temporal resolution of 5 days. It provides an opportunity to conduct high-frequency water-quality monitoring for small waterbodies. Since illegal discharges are an important issue for urban water management, total phosphorous (TP), total nitrogen (TN), and chemical oxygen demand (COD) were chosen as the target water-quality parameters. TP, TN and COD, however, are non-optically active parameters. There are fairly limited previous studies on retrieving these parameters in comparison with optically active parameters, e.g. Chlorophyll-*a* etc. Based on the fact that non-optically active parameters may be highly correlated with optically active parameters, this study compared 255 possible Sentinel-2 imagery band compositions to identify the most appropriate ones for TP, TN and COD retrieval. Three machine-learning models, namely Random Forest (RF), Support Vector Regression (SVR) and Neural Networks (NN), were compared to seek the most robust ones for retrieving the above non-optically active parameters. Results showed that the most appropriate band (hereafter termed as 'B_{index}' for brevity) compositions for TP, TN, and COD retrieval were '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$', '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$', and '$B_2 + B_3 + B_5 + B_6 + B_7 + B_8$' respectively. The coefficient of determination (R^2) of TP, TN, and COD estimations by NN, RF and SVR was 0.94, 0.88, and 0.86, respectively. The retrieval performances of these non-optically active parameters were hence significantly improved by the optimized machine-learning models and imagery band selection. The developed models have limitations in applying to other areas, thus band selection and tuning parameters with new data are necessary for different areas. The water-quality mapping obtained from Sentinel-2 imagery provided a full spatial coverage of the water-quality characterization for the entire water surface, and helped identify illegal discharges to urban waterbodies. This study provides a new practical and efficient water-quality monitoring strategy for managing small waterbodies.

ARTICLE HISTORY
Received 3 June 2020
Accepted 10 October 2020

CONTACT Jinhui Jeanne Huang huangj@nankai.edu.cn B406, College of Environmental Science and Engineering, 38 Tongyan Rd., Haihe Education Park, Jinnan District, Tianjin, P.R.China, 300350

Supplemental data for this article can be accessed here.

1. Introduction

In urban areas, waterbodies, such as lakes and reservoirs, may be polluted by illegal discharges of industrial effluent and domestic sewage (Shao et al. 2006). Deterioration of water quality may increase human exposure to diseases and harmful chemicals; reduce ecosystem productivity and biodiversity; and damage aquaculture, agriculture and other water-related industries (Hoekstra, Buurman, and Van Ginkel 2018; Brönmark and Hansson 2002). Traditional water-quality monitoring methods are primarily based on water sample collection and testing or automatic in-situ measurements. Both methods are either labour intensive or very costly. In addition, most water sample testing would need reagents for testing, and the treatment of waste generated by testing is also costly. Although these methods may have high accuracy, individual samples only reflect the water quality at specific sampling points and are limited in characterizing water quality for the entire water surface (Shuchman et al. 2013; Ritchie, Zimba, and Everitt 2003; O'Reilly et al. 1998; Olmanson, Brezonik, and Bauer 2013). In many cases, the decision makers would need a full picture of water characteristics over the entire water surface for water-quality management. Remote sensing has been used to monitor water quality since the 1970s (Vignolo, Pochettino, and Cicerone 2006; Holyer 1978; Ritchie, Schiebe, and McHenry 1976). Compared with traditional methods, remote sensing can provide the full coverage required for dynamic water-quality monitoring (Duan, Ma, and Hu 2012).

Over the past several decades, scholars have carried out extensive research works on water-quality monitoring by remote sensing, and have achieved good results in estimating optically active parameters, such as Chlorophyll-*a* (Chl-*a*), suspended particulate matter (SPM), coloured dissolved organic matter (CDOM), turbidity and transparency etc. (Brezonik et al. 2015; Hou et al. 2017; Shi et al. 2015; Bugnot et al. 2018; Shahzad et al. 2018; Doña et al. 2015). However, estimating non-optically active parameters, such as total phosphorus (TP), total nitrogen (TN), and chemical oxygen demand (COD) directly from spectral characteristics is difficult, because they are less likely to impact the optical characteristics measured by satellite sensors (Deng, Zhang, and Cen 2019; Gholizadeh and Melesse 2017; Mathew, Srinivasa Rao, and Mandla 2017; Xiong et al. 2020; Ferdous, Tauhid, and Rahman 2020; Chang, Bai, and Chen 2017; Gao et al. 2015). Generally, non-optically active parameters have been estimated indirectly based on the correlation between optically active parameters and non-optically active parameters (Carlson 1977; Wu et al. 2010; Mathew, Srinivasa Rao, and Mandla 2017). For instance, Chang, Xuan, and Yang (2013) estimated TP in Tampa Bay (USA) with the Moderate-resolution Imaging Spectroradiometer (MODIS) images and genetic programming models. The results indicated that the Band 1, Band 3 and Band 4 of MODIS images were most influential for the determination of TP concentrations. Li et al. (2017a) developed empirical models to estimate TP and TN in the Xin'anjiang Reservoir (China) using Land Remote-Sensing Satellite (System, Landsat) 8 Operational Land Imager (OLI) images. The Landsat 8 OLI-derived factors (Band 1 + Band 3 + Band 4)/Band 2 and Band 4/(Band 2 + Band 5) shown a strong correlation with TP and TN concentrations, respectively. Wang et al. (2004) estimated COD in the reservoirs of Shenzhen (China) using the Landsat Thematic Mapper (TM) images. The results indicated that the TM Band 1 to Band 4 and organic pollution measurements (e.g. COD) had high correlation. Although the previous studies on non-optically active parameters were fairly limited, these studies proved the possibility of retrieving non-optically active parameters from optical characteristics.

The most widely used remote-sensing imagery in the existing research works are from Landsat TM, Enhanced Thematic Mapper Plus (ETM+) and OLI, Sea-Viewing Wide Field-of-View Sensor (SeaWiFS), MODIS, and Medium Resolution Imaging Spectrometer (MERIS) (Moses et al. 2009; Halme, Pellikka, and Mõttus 2019; Kishino, Tanaka, and Ishizaka 2005; Shenglei et al. 2016; Keith et al. 2018). However, the temporal resolution of TM, ETM+ and OLI data is 16 days, and the spatial resolution of SeaWiFS, MODIS and MERIS data is greater than or equal to 250 × 250 m, resulting in challenges in high-frequency characterization of the water quality for small waterbodies. Hyperspectral imagery contains a large number of continuous spectral information and provides more spectral characteristics for water-quality retrieval (Brando and Dekker 2003; Li et al. 2017a; Gitelson et al. 2011). However, spaceborne hyperspectral data (e.g. data of Hyperion and Compact High Resolution Imaging Spectrometer (CHRIS)) is only experimental data rather than operational at present, and airborne hyperspectral data (e.g. data of Airborne Visible Infrared Imaging Spectrometer (AVIRIS), Compact Airborne Spectrographic Imager (CASI), and Contact Image Sensors (CIS)) has very limited spatial coverage with high cost (Lunetta et al. 2009; Halme, Pellikka, and Mõttus 2019). By comparison, the recently launched Sentinel-2 produces imagery with a spatial resolution of 10 × 10 m and a temporal resolution of 5 days. It provides an opportunity to conduct high-frequency water-quality monitoring for small waterbodies.

In summary, studies on non-optically active parameters were fairly limited in the past, and most remote-sensing imagery has a too coarse spatial or temporal resolution for high-frequency water-quality monitoring of small waterbodies. Therefore, this study aims to retrieve non-optically active parameters for small urban waterbodies using Sentinel-2 imagery. TP, TN and COD were selected as the target parameters, since they may help in identifying the sources of illegal discharges from industrial effluent or domestic sewage. A total of 255 possible band compositions of eight Sentinel-2 imagery bands were compared to identify the most appropriate ones for retrieving each water-quality parameter. Three machine-learning models, namely Random Forest (RF), Support Vector Regression (SVR) and Neural Networks (NN), were introduced in the empirical methods (Wang et al. 2018; Li et al. 2017a; Le et al. 2011), and compared to seek the most robust ones for retrieving the non-optically active parameters. This study may help urban water management by providing a more practical and efficient water-quality monitoring strategy of non-optically active parameters.

2. Materials and methods

2.1. Study area

This study selected a small urban lake with an area of 0.60 km^2 (600 m × 1000 m). The lake is located in an industrial park in the City of Tianjin, China, and provides flood control and ecological balance for its surrounding areas (Figure 1). The total developed area of the industrial park is 2 km^2 with diversified land uses, including industrial, residential, and commercial. The ecosystem health of the lake is directly related to the management of the surrounding areas.

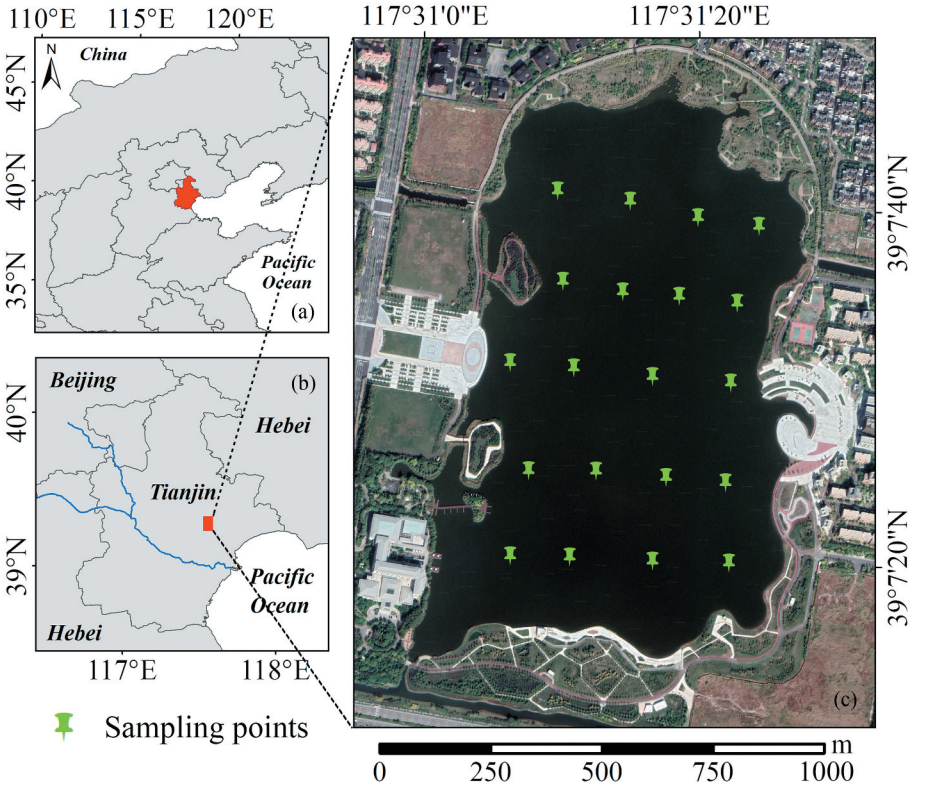

Figure 1. Locations of the City of Tianjin (a), the study area (b) and the sampling points (green pin labels) (c).

2.2. Field data

Two types of field data were used in this study: field surveys (N = 60) and biweekly measurements in Lake Simcoe (N = 33) by the Ontario Ministry of the Environment, Conservation and Parks, Canada (https://www.ontario.ca/page/ministry-environment-conservation-parks). The locations of the pre-defined stations in Lake Simcoe were shown in the supplemental material. Part of the data from the field surveys (N = 40) concurrent with Sentinel-2 measurements was used to calibrate and validate the models. The other part of the data from the field surveys (N = 20) and the measurements in Lake Simcoe (N = 33) were used to further validate the model robustness and generalization.

In the field surveys, 20 sampling points were evenly distributed on the water surface by the grid method (Figure 1). The density of the sampling points was 0.03 sites km^{-2}. To provide simultaneous sampling with Sentinel-2 overpass, water samples were collected by Yunzhou SE40 unmanned surface vessel (USV) at 10:00 to 12:00 on 16 November 2018, 20 May 2019 and 9 June 2019. The three sampling dates were selected when aquatic plants did not grow rapidly and massively, since the aquatic plants covering the lake would impact the retrieval accuracy. A total of 60 water samples were collected, with sampling depths of 30–50 cm. There was no cloud cover above the lake on the sampling dates. The time between sampling and satellite overpass was less than 4 hours. After being collected, water

samples were quickly put into amber glass bottles to avoid sunshine, and sent to the laboratory for testing. The testing method for each water-quality parameter was listed in Table 1. The field survey data on 20 May 2019 and 20 June 2019 (N = 40) constitute the ground truth data set for model calibration and validation. The field survey data on 16 November 2018 (N = 20) was selected as a repetitive experiment to validate the robustness of the developed models. The measurements in Lake Simcoe (N = 33) were used as an independent data set to validate the model generalization.

The spatial distributions of the measured water-quality parameters on the three sampling dates were visualized by ArcGIS 10.4 (Environmental Systems Research Institute, Inc., Redlands, California, USA). On 16 November 2018, the averages of TP, TN, and COD were 0.62 mg l^{-1}, 1.37 mg l^{-1}, and 31.70 mg l^{-1}, respectively. On 20 May 2019, the averages of TP, TN, and COD were 0.29 mg l^{-1}, 0.66 mg l^{-1}, and 50.45 mg l^{-1}, respectively. On 9 June 2019, the averages of TP and TN increased to 0.85 mg l^{-1} and 1.63 mg l^{-1}, respectively. The average of COD decreased to 17.45 mg l^{-1} (Figure 2).

2.3. Remote-sensing imagery processing

Since 3 December 2015, The European Space Agency (ESA) has officially provided free download services of Sentinel-2 MultiSpectral Instrument (MSI) imagery for global users. Real-time updated imagery can be downloaded freely from the Copernicus Open Access Hub (https://scihub.copernicus.eu). The imagery contains thirteen spectral bands ranging from the visible (VNIR) and near infrared (NIR) to the shortwave infrared (SWIR). The spatial resolution of Band 2 (B_2, 492.10 nm), Band 3 (B_3, 559 nm), Band 4 (B_4, 665 nm), and Band 8 (B_8, 833 nm) is 10 × 10 m. The spatial resolution of Band 5 (B_5, 703.80 nm), Band 6 (B_6, 739.10 nm), Band 7 (B_7, 779.70 nm), Band 8A (B_{8A}, 864.80 nm), Band 11 (B_{11}, 1610.40 nm), and Band 12 (B_{12}, 2185.70 nm) is 20 × 20 m. The spatial resolution of Band 1 (B_1, 442.30 nm), Band 9 (B_9, 943.20 nm) and Band 10 (B_{10}, 1376.90 nm) is 60 × 60 m. B_1, B_9 and B_{10} are three atmospheric bands, and hence were not analysed in this study.

The three images used in this study were L1C products, namely the atmospheric apparent reflectance products after ortho-rectification and sub-pixel geometric correction. The radiometric calibration and atmospheric correction were completed by sen2cor v2.8. Then, the L2A products (i.e. the bottom of the atmosphere (BOA) reflectance products) were obtained. The four bands with a spatial resolution of 20 × 20 m were resampled to 10 × 10 m using the Sentinel Application Platform (SNAP) v6.0, creating eight bands with a spatial resolution of 10 × 10 m (Figure 3).

Pixel values corresponding to the sampling points were extracted to generate the pixel value data set. The entire data set for machine-learning model development was composed of the 'ground truth data set' and 'pixel value data set'. The water area was extracted by the Normalized Difference Water Index (NDWI, Equation (1)) in ENVI 5.3 (McFeeters 1996).

Table 1. Testing method of each water-quality parameter.

Parameter	Method
TP	Acid persulphate digestion method
TN	Persulphate oxidation method
COD	Reactor digestion method

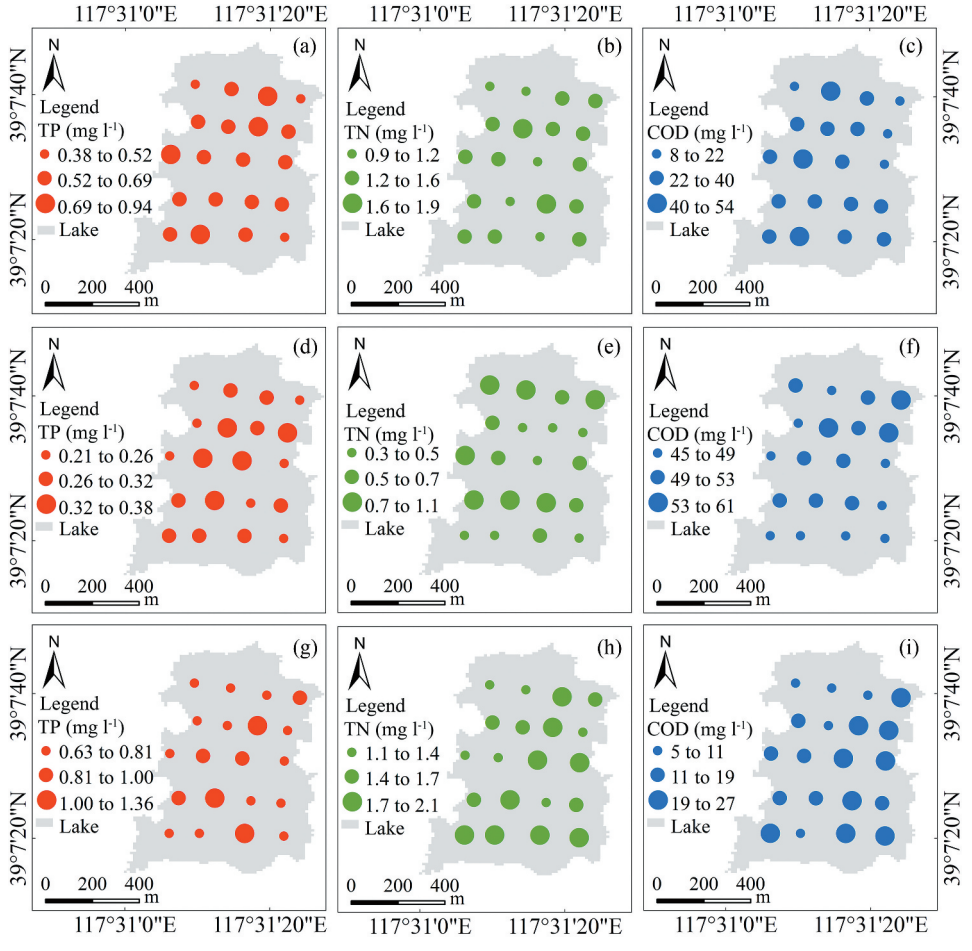

Figure 2. Spatial distributions of the measured water-quality parameters. (a–c) represent TP, TN COD, respectively on 16 November 2018; (d–f) represent TP, TN, COD, respectively on 20 May 2019; (g–i) represent TP, TN, COD, respectively on 9 June 2019. The bigger the dot size, the more serious the water pollution.

$$\mathrm{NDWI} = \frac{X_{\mathrm{Green}} - X_{\mathrm{NIR}}}{X_{\mathrm{Green}} + X_{\mathrm{NIR}}} \tag{1}$$

where X_{Green} and X_{NIR} are the pixel values of the green band and the NIR band, respectively. For Sentinel-2 imagery, the green band and the NIR band are B_3 and B_8, respectively.

2.4. Development of machine-learning models

Three machine-learning models, namely RF, SVR and NN, were selected to fit the correlation between water-quality parameters and pixel values, respectively. The coefficient of determination (R^2), mean absolute percentage error (MAPE) and root mean square percentage error (RMSPE) were used to evaluate the model performances (Xiong et al. 2020). They were defined as follows:

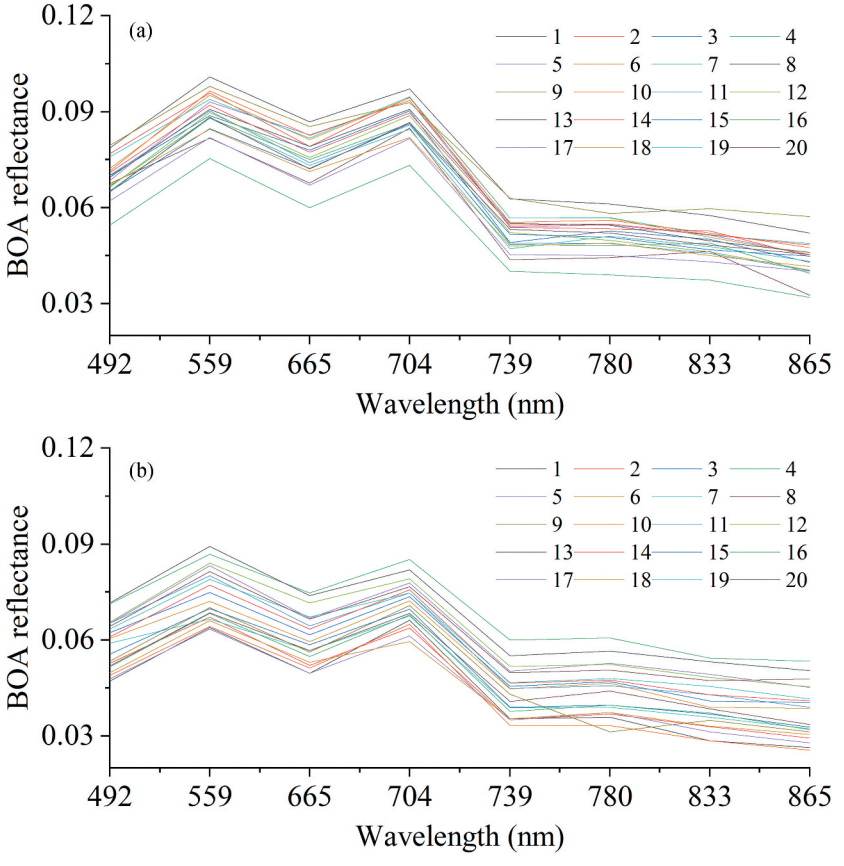

Figure 3. BOA reflectance of the sampling points on 20 May 2019 (a) and 9 June 2019 (b). Different colours represent different sampling points.

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

$$\text{MAPE}(\%) = \frac{100}{N} \times \sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3}$$

$$\text{RMSPE}(\%) = 100 \times \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2} \tag{4}$$

Where y_i, $\bar{y}_i$, and $\hat{y}_i$ are the measured, the mean, and the estimated water-quality parameters, respectively; N is the number of the sampling points.

Learning curves were used to tune model parameters. A learning curve shows model performance on the test set in the y-axis and different values of a model parameter in the x-axis. According to learning curves, the optimal value of each model parameter was determined. To prevent the over fitting and improve model generalization, a 10-fold cross validation was used to calculate the evaluation metrics of the model performances

(Rapinel et al. 2019). For each cross validation, the whole data set was randomly split into 70% training set ($N = 28$) and 30% test set ($N = 12$). A model was fitted on the training set and the model performance was evaluated on the test set. The final R^2, MAPE, RMSPE were the averages of the 10 fold cross validation. The principles of the three machine-learning models were described in the following subsections.

2.4.1. Random Forest

RF is an ensemble learning method (Wang et al. 2019). It uses data to construct multiple models, and integrates the modelling results of all models by voting (classification problem) or averaging (regression problem), so that the results of the entire model have high accuracy and generalization performance. In the implementation of RF, an equal number of training samples are randomly extracted from all samples, then M sub-feature sets are selected from all input features, and finally, an optimal attribute is selected from the sub-feature set for node splitting.

When solving the regression problem, each decision tree is a regression tree, which is divided by the minimum mean square deviation. In other words, for splitting feature A, the corresponding arbitrary splitting point s splits the data set into two data sets (D_1 and D_2), so that the mean square deviation of the two data sets after splitting is the minimum:

$$\min_{A,s}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2\right] \tag{5}$$

where c_1 and c_2 are the mean sample output values of D_1 and D_2, respectively; yi is the measured value.

Then, at the two nodes after splitting, the splitting continues according to the above principle. The final prediction result is the mean of all decision trees. In this study, RF was implemented by scikit-learn 0.21.3 of Python 3.7. Using learning curves, the main parameters were set as Table 2:

2.4.2. Support Vector Regression

SVR maps x in low-dimensional feature space to a high-dimensional feature space $\varphi(x)$ with a non-linear function φ, and seeks a linear regression hyperplane in high-dimensional feature space to solve the non-linear problems in low-dimensional feature space (Vapnik 1995; Mountrakis, Im, and Ogole 2011). Linear functions in high-dimensional feature space can be constructed as follows:

$$y = \langle w, \varphi(x)\rangle + b \tag{6}$$

where x and y are the input and output; operator '$\langle\rangle$' denotes the inner product computation, and the weight vector w and bias constant b can be obtained by minimizing the risk function:

Table 2. Main parameter setting of RF.

Parameter	TP	TN	COD
N estimators	80	36	40
Max depth	6	5	3
Max features	5	2	6

$$\min\left(\frac{w^2}{2}+C\sum_{i=1}^{N}L_\varepsilon(x_i,y_i,f)\right) \tag{7}$$

where operator '||||' denotes the Euclidean norm computation.

In Equation (7),

$$L_\varepsilon(x_i,y_i,f)=\begin{cases}|y_i-f(x_i)|-\varepsilon, |y_i-f(x_i)|\geq\varepsilon\\ 0, otherwise\end{cases} \tag{8}$$

where L_ε is the loss function; C is a pre-set penalty coefficient, which is used to punish errors greater than ε; ε is the deviation between the estimated and the measured values; N is the number of sampling points. Introducing slack variables ξ_i and ξ_i^*, the solution of Equation (7) can be transformed to Equation (9):

$$\min\left(\frac{w^2}{2}+C\sum_{N}^{i=1}\left(\xi_i+\xi_i^*\right)\right) \tag{9}$$

The constraint condition is as follows:

$$\begin{cases}y_i-[w,x_i+b]\leq\varepsilon+\xi_i\\ [w,x_i+b]-y_i\leq\varepsilon+\xi_i^*\\ \xi_i^*,\xi_i\geq 0\end{cases} \tag{10}$$

The Lagrange multipliers a_i and a_i^* are introduced to establish the Lagrange function and then solve the dual problem of the original problem. Finally, the regression function of the optimal hyperplane is obtained:

$$f(x) = \sum_{i=1}^{N}\left(a_i-a_i^*\right)K(x_i,x)+b \tag{11}$$

where $K(x_i,x)$ is the kernel function:

$$K(x_i,x)=\varphi x_i,x\rangle = \varphi(x_i)^{\mathrm{T}}\varphi(x) \tag{12}$$

where T is the transpose operator.

The most commonly used kernel functions include linear kernel, polynomial kernel, sigmoid kernel, radial basis function (RBF) kernel, etc. (Chen et al. 2019). Using learning curves, the RBF kernel was selected as the kernel function in this study:

$$K(x_i,x)=\exp\left(-\frac{x-x_i}{\sigma^2}\right) \tag{13}$$

In this study, SVR was implemented by scikit-learn 0.21.3 of Python 3.7. Using learning curves, the main parameters were set as Table 3:

Table 3. Main parameter setting of SVR.

Parameter	TP	TN	COD
C	1.23	1.23	7.35
Gamma	0.06	0.10	0.06

2.4.3. Back-propagation Neural Networks

NN is a multi-layer perceptron model, which is trained by error back propagation algorithm. The model is developed according to the internal relations of the data itself, and has good non-linear approximation ability and comprehensive processing ability for cluttered information (Wu et al. 2009). The model structure consists of one input layer, one or more hidden layers and one output layer. The upper and lower layers are fully connected, and there is no connection between neurons in each layer.

The numbers of nodes in the input layer (N), hidden layer (L), and output layer (M) are determined by the sequence of input and output (x, y). The training process of the model employs the following steps:

(1) *Initialization*. Initialize $w_{i,j}$ and $w_{j,k}$. The $w_{i,j}$ is the connection weight between the jth neuron in the input layer and the ith neuron in the hidden layer. The $w_{j,k}$ is the connection weight between the kth neuron in the hidden layer and the jth neuron in the output layer. Meanwhile, the threshold a of the hidden layer, threshold b of the output layer, activation functions, and learning efficiency η are preset.
(2) *Hidden layer output calculation*. The output H of the hidden layer is calculated according to the input variable x, $w_{i,j}$ and a:

$$H_j = f\left(\sum_{i=1}^{N} w_{i,j}x_i - a_j\right) j = 1, 2, \ldots, L \tag{14}$$

(1) *Output layer output calculation*. The predicted output O is calculated according to H, $w_{j,k}$, and b:

$$O_k = \sum_{j=1}^{L} H_j w_{j,k} - b_k k = 1, 2, \ldots, M \tag{15}$$

(1) *Error calculation*. The prediction error e is calculated according to O and the expected output y. If the error meets the requirement, the training will be completed, otherwise step 5 will be repeated.
(2) *Updating weights and thresholds*. Turn back to step 2 after updating $w_{i,j}$, $w_{j,k}$, a, and b according to e.

The common activation functions in NN include ReLU function, sigmoid function and tanh function. Among them, ReLU function is most commonly used (Kurt et al. 2008). ReLU function is a piecewise linear function, and well makes up for the gradient disappearance problem of sigmoid function and tanh function. The ReLU function is as follows:

$$g(z) = \begin{cases} z, z > 0 \\ 0, z < 0 \end{cases} \tag{16}$$

In this study, NN was implemented in Keras 2.2.4 of Python 3.7. The number of neurons in the input layer was set to six, for there were six input imagery bands. By analysing the model performances and calculation efficiency, the number of the hidden layers was set to four. For TP and TN, the number of neurons in each hidden layer was set to 300. For COD, the number of neurons in each hidden layer was set to 200.

3. Results

3.1. Sentinel-2 imagery band selection

The average spectral shapes across the TP, TN and COD concentrations were shown in Figure 4. The average BOA reflectance on the two sampling dates showed significant differences. The BOA reflectance increased with the increase of the TP, TN and COD

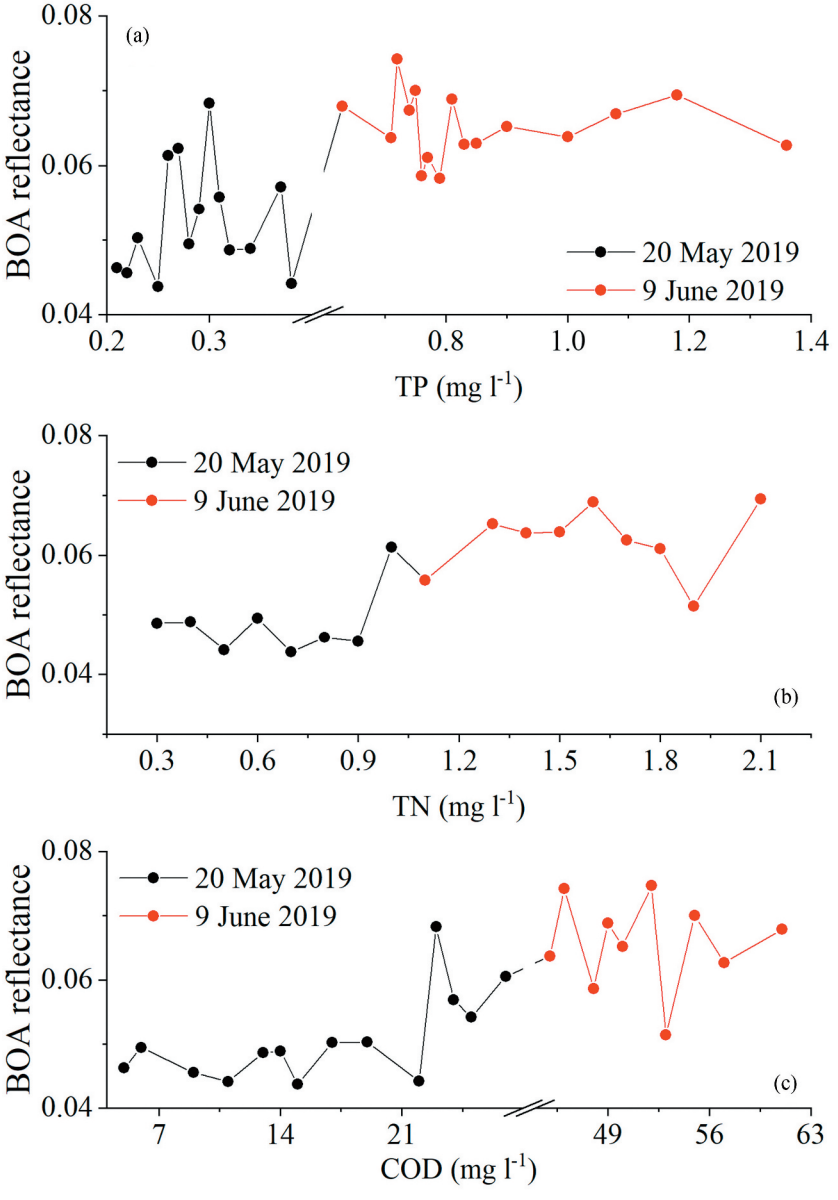

Figure 4. Average spectral shapes across the concentrations of TP (a), TN (b), and COD (c). The BOA reflectance was the average of the eight selected imagery bands in this study. When plotting, different BOA reflectance corresponding to the same concentration was averaged.

concentrations. On each sampling date, BOA reflectance fluctuated significantly with the changes of TP, TN and COD concentrations. The results indicated the possibility to estimate TP, TN and COD from spectral characteristics.

Then, the standard deviations of the BOA reflectance in each band across the TP, TN and COD concentrations were calculated and compared (Figure 5). For TP and TN, the BOA reflectance of B_3, B_4, B_5, B_6, B_7 and B_8 fluctuated more obviously across the concentration ranges. For COD, the BOA reflectance of B_2, B_3, B_5, B_6, B_7 and B_8 fluctuated more obviously across the concentration ranges. The results suggested that the band compositions of the above-mentioned bands could be used for the TP, TN and COD retrieval.

In order to further validate whether the above-mentioned band compositions were most appropriate, this study used all possible band compositions (a total of 255) to retrieve each water-quality parameter by multiple linear regression. R^2 was selected to evaluate the model performances (Figure 6).

According to Figure 6, with the increase of band number, the average R^2 increased, and reached the maximum when the band number was 6. The average R^2 of seven bands compositions was greater than that of eight bands compositions. The most influential bands were B_3, B_4 and B_5, namely, the green and red bands (B_5 was the vegetation red edge with a wavelength of 705 nm). The most appropriate band compositions for TP, TN, and COD retrieval were '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$', '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$' and '$B_2 + B_3 + B_5 + B_6 + B_7 + B_8$', respectively.

3.2. Performances of machine-learning models

The most appropriate band composition of each water-quality parameter mentioned in Subsection 3.1. was used to develop RF, SVR and NN models. Comparing the water-quality parameters estimated by the machine-learning models with the ground truth, Figure 7 was obtained.

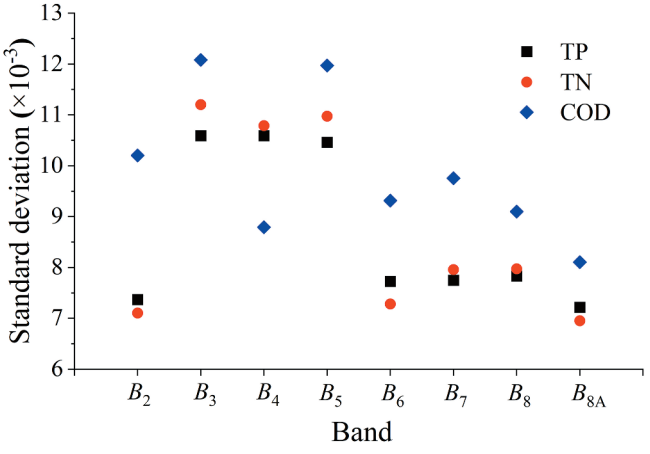

Figure 5. Standard deviations of the BOA reflectance in each band across the TP, TN and COD concentrations. The black squares, red dots and blue diamonds represent TP, TN and COD, respectively.

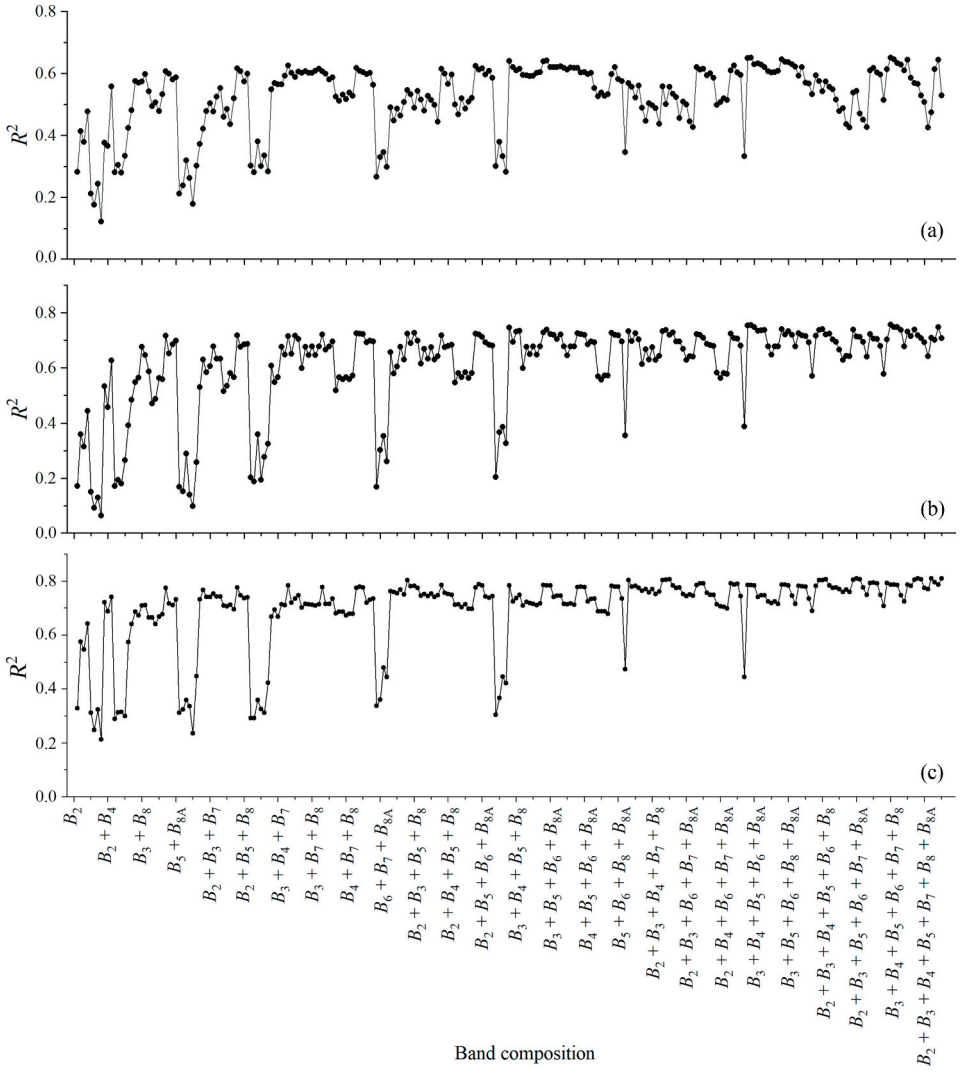

Figure 6. R^2 of TP (*a*), TN (*b*) and COD (*c*) retrieval using different band compositions by multiple linear regression.

The optimal models of TP, TN and COD retrieval were different. For TP, the performance of NN was good. R^2 reached 0.94, and the MAPE and RMSPE were 12.43% and 16.80%, respectively. For TN, the performance of RF was good. R^2 reached 0.88, and the MAPE and RMSPE were 18.39% and 29.64%, respectively. For COD, the performance of SVR was good. R^2 reached 0.86, and the MAPE and RMSPE were 12.55% and 18.75%, respectively.

Taking the performances of multiple linear regression as a comparison, machine learning significantly improved the retrieval accuracy of each water-quality parameter (Table 4).

Furthermore, using the Kriging spatial interpolation (Carletti, Picci, and Romano 2000; Beaulant et al. 2008; Wang et al. 2019) of ArcGIS 10.4, each water-quality parameter was interpolated with the ground truth values. Figure 8 showed the comparisons of water-quality parameters estimated by the spatial interpolations and machine-learning models.

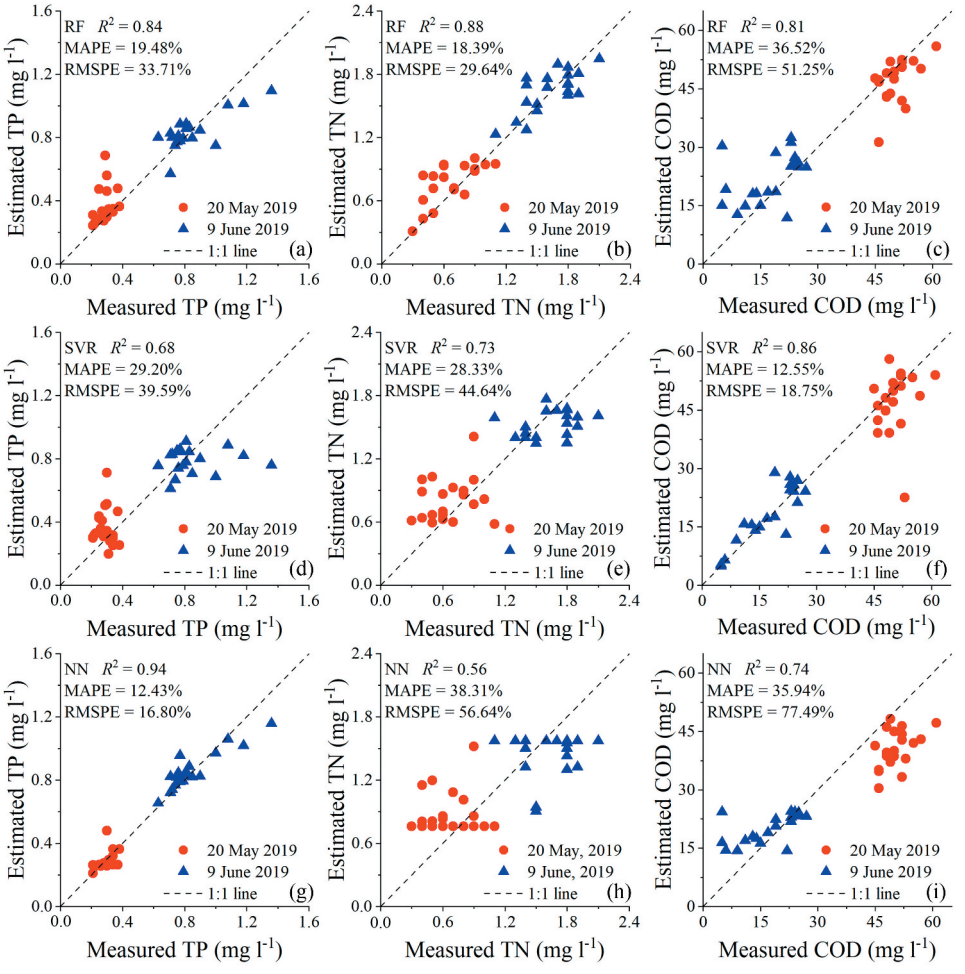

Figure 7. The model performances of TP, TN, and COD retrieval. (a–c) represent the accuracy of RF; (d–f) represent the accuracy of SVR; and (g–i) represent the accuracy of NN. The red dots and blue triangles represent the sampling points on 20 May 2019 and 9 June 2019, respectively.

Table 4. Comparison of accuracy between machine learning and multiple linear regression.

	Machine learning			Multiple linear regression		
Parameter	R^2	MAPE (%)	RMSPE (%)	R^2	MAPE (%)	RMSPE (%)
TP	0.94	12.43	16.80	0.65	30.65	39.83
TN	0.88	18.39	29.64	0.76	22.14	36.24
COD	0.86	12.55	18.75	0.81	39.18	71.65

According to Figure 8, machine-learning models recreated the high dynamic ranges of the measured water-quality parameters in detail. Especially for 1–20 sampling points of TP and TN, the estimations by the spatial interpolations showed almost no fluctuation. Further accuracy comparison indicated that the R^2, RMSPE, and MAPE of the TP estimation were higher by 30.84%, 26.85% and 9.32% using the machine-learning model (NN) than the spatial interpolation. For the TN estimation, the R^2, RMSPE and MAPE were higher by

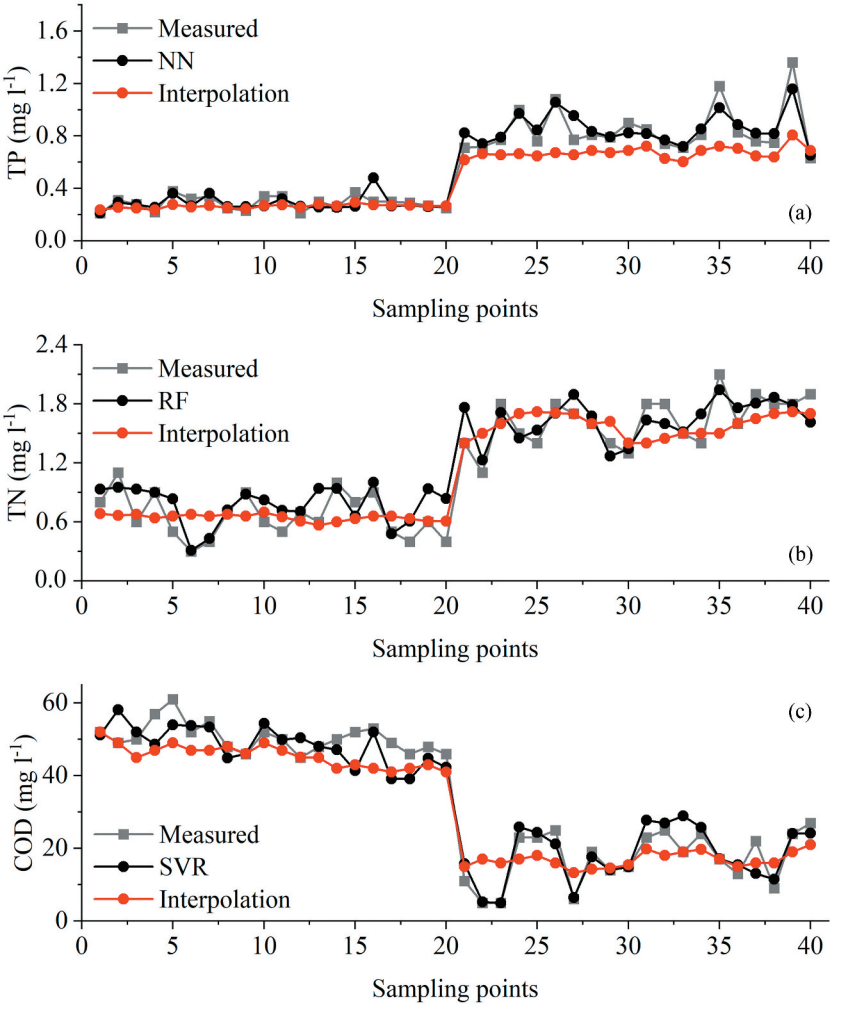

Figure 8. Comparisons of water-quality parameters estimated by the spatial interpolations and machine-learning models. (a–c) represent TP, TN, COD, respectively. The black and red dots represent the water-quality parameters estimated by the machine-learning models and the spatial interpolations, respectively. The grey squares represent the measured water-quality parameters.

23.41%, 17.83% and 6.80% using the machine-learning model (RF) than the spatial interpolation. For the COD estimation, the R^2, RMSPE and MAPE were higher by 13.68%, 40.67% and 30.61% using the machine-learning model (SVR) than the spatial interpolation. The results proved that compared to the spatial interpolation, machine learning could recreate the dynamic ranges of the measured water-quality parameters in more detail, and significantly improve the estimation accuracy.

3.3. Water-quality mapping

Using the Geospatial Data Abstraction Library (GDAL) v3.0.4 in Python 3.7, the imagery pixel values of the entire water surface were read and imported into the developed

machine-learning models to estimate TP, TN and COD. Then the estimated results were output into the imagery to generate the water-quality distributions (Figure 9).

The spatial distributions of TP and TN tended to be consistent, which might be due to the fact that TP and TN were from the same pollution sources. For example, domestic sewage from residential areas was discharged into the lake. TP and TN in the south and west of the lake were higher than those in the north and east of the lake. On 20 May 2019, the high values of TP and TN were distributed in the southwest of the lake, while on 9 June 2019, the area expanded from the southwest of the lake to the north of the lake. Accordingly, the averages of TP and TN increased from 0.29 mg l^{-1} and 0.66 mg l^{-1} on 20 May 2019 to 0.85 mg l^{-1} and 1.63 mg l^{-1} on 9 June 2019, respectively. COD in the east of the lake was higher than that in the west of the lake. On 20 May 2019, the high values of COD were distributed in most areas except the southwest of the lake. By 9 June 2019, the area of high values contracted eastward to the centre and east of the lake. Meanwhile, the average of COD decreased from 50.45 mg l^{-1} on 20 May 2019 to 17.45 mg l^{-1} on 9 June 2019.

According to the mapping of TP and TN, domestic sewage containing N and P might be continuously discharged into the lake. It was observed from the remote-sensing imagery that there was a piece of farmland with an area about 1.25 km^2 as well as a subdivision of a residential area adjacent to the lake on the southwest. Therefore, the increases of TP and TN in the south and west of the lake were likely related to the application of chemical fertilizer and the discharges of domestic sewage. The high values of COD were widely distributed in

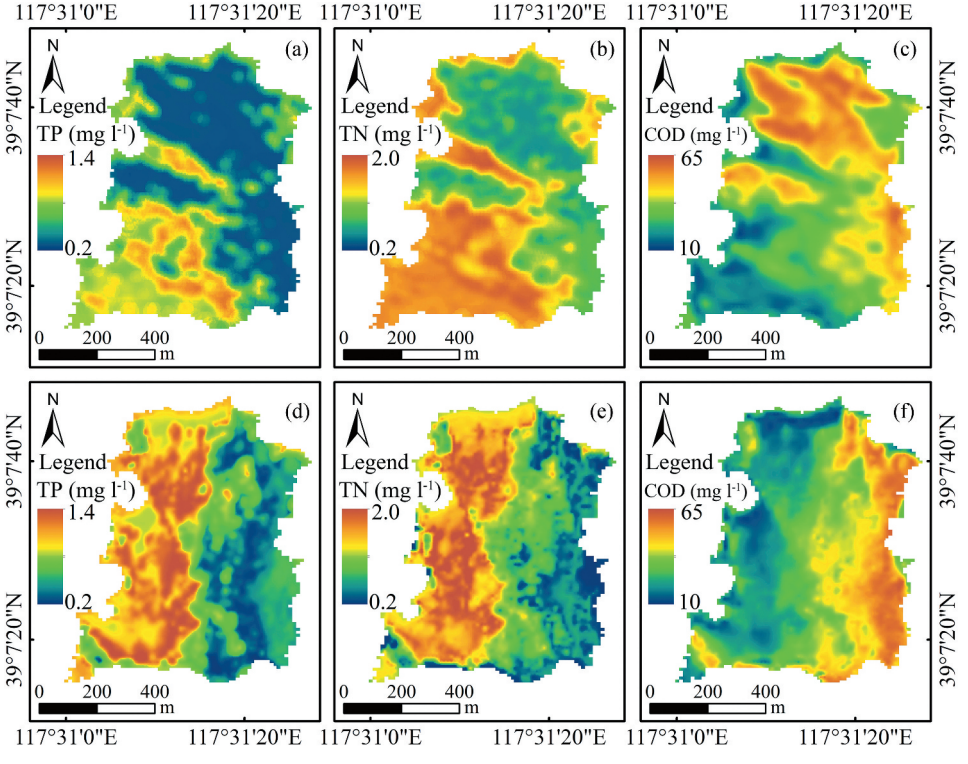

Figure 9. Water-quality distributions on the two sampling dates. (a–c) represent TP, TN, COD, respectively on 20 May 2019; (d–f) represent TP, TN, COD, respectively on 9 June 2019.

the centre and east of the lake. From 20 May 2019 to 9 June 2019, there was an obvious contraction process to the east of the lake. According to this change, industrial effluent or domestic sewage might be discharged into the east of the lake, which may originate from the nearby residential areas and a pharmaceutical factory in the east of the lake.

From the perspective of optical characteristics, the spectrum of lake water is mainly affected by three optically active components: SPM, phytoplankton and CDOM (Xiong et al. 2020; Wang et al. 2020). Many previous studies confirmed that different components had different absorption characteristics. For instance, phytoplankton has obvious absorption peaks at the blue band (430 to 500 nm) and the red band (650–750 nm) (Ma et al. 2006; Pahlevan et al. 2020). CDOM has a strong absorption at the ultraviolet band (280–400 nm), and the absorption shows an exponential decrease from the ultraviolet to visible wavelengths (Mannino et al. 2014; Brezonik et al. 2015). In the west of the lake, algae and aquatic plants grew continuously due to the high TP and TN, and the optical characteristics of the lake water were dominated by phytoplankton. While in the east of the lake, when industrial effluent or domestic sewage containing plenty of organic matter entered the lake, the optical characteristics of the lake water were dominated by CDOM and SPM. Therefore, the spatial distribution of water quality estimated from spectral characteristics showed an obvious difference between TP and TN concentrations and COD. This result was consistent with the above analysis on the source tracing of pollutants in the lake.

It also could be observed that the water quality was not evenly distributed, although the waterbody was fairly small (the surface area was only 0.60 km^2). The study of water-quality retrieval based on high spatial resolution remote-sensing imagery was therefore crucial for many water management issues, e.g. identifying illegal discharges to urban waterbodies and spills on the shore etc.

4. Discussion

4.1. Model robustness, generalization and limitations

The novelty of the approach proposed in this study is to determine the optimal band composition of TP, TN and COD retrieval by analysing the correlation between 255 band compositions and each water-quality parameter. Moreover, three machine-learning models, i.e. RF, SVR and NNs, were constructed for each water-quality parameter to seek the most appropriate one. During the model training, the learning curves were used to tune each model parameter to ensure the optimal model performance. The evaluation metrics of the model performance were the averages of a 10 fold cross validation. In each cross validation, the test set is new to the model, which can improve the model robustness and generalization to a certain extent. Figure 10 showed the model accuracy on the training set and test set. There was no significant difference between the estimation errors of the two data sets.

Furthermore, we compared the satellite-derived results in the field survey data on 16 November 2018 (Figure 11). All R^2 of TP, TN and COD decreased, but kept above 0.60. For TP and TN, the MAPE and RMSPE decreased, mainly because the concentration ranges narrowed on 16 November 2018. The MAPE and RMSPE of COD increased to 37.41% and 30.49%, respectively. The results proved the model robustness and generalization in the local area.

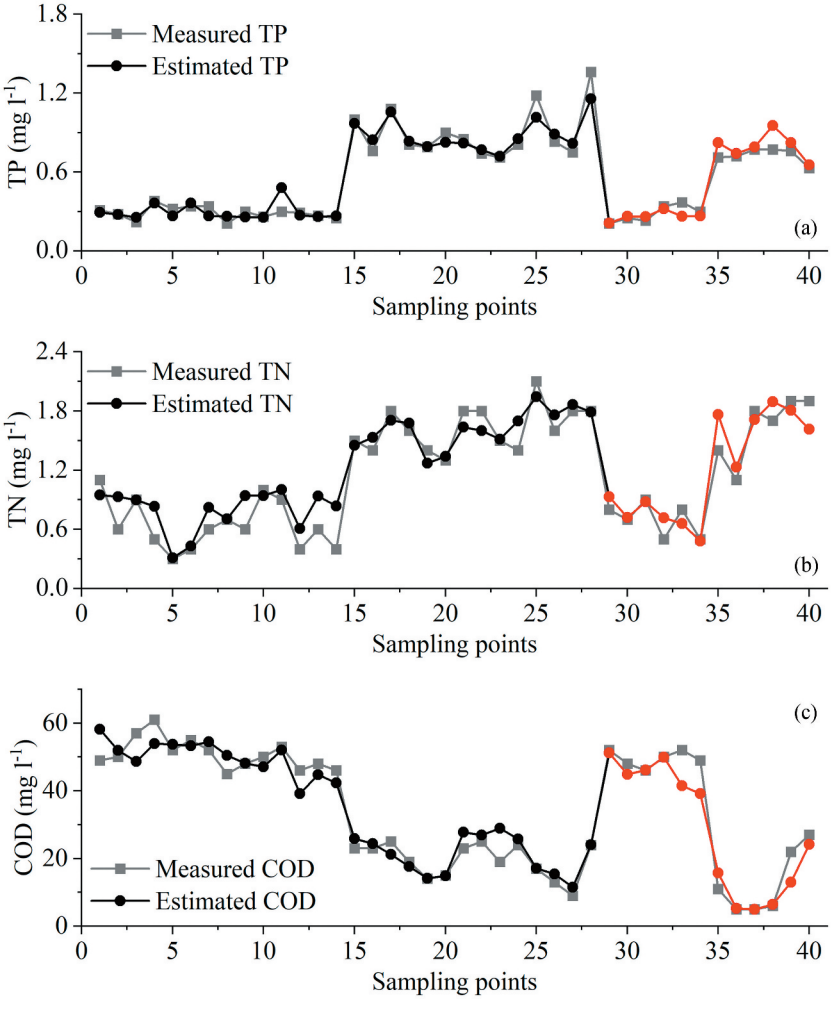

Figure 10. Comparisons of the model accuracy on the training set and test set. (a–c) represent TP, TN, COD, respectively. The black and red dots represent the estimated values from the training set and test set, respectively. The grey squares represent the measured values.

To validate whether the developed models work well in other areas, we compared the satellite-derived results in the field measurements of Lake Simcoe in 2018. Since COD is not a regular parameter, no matches between satellite-derived results and field measurements were generated. For TP and TN, 33 samples were matched, respectively (Figure 12). According to Figure 12, the model of TP was completely failed due to the huge gap in the concentration ranges. The R^2, MAPE and RMSPE of TN were 0.53, 31.69% and 59.32%, respectively. The model performance also decreased significantly. The results were consistent with the research work of Cao et al. (2020). In addition, since the optical characteristics of different waterbodies are different, the band composition is also referred to as one of the reasons that affect the model performance. Therefore, the developed models are capable of providing reliable results in local areas, but also have limitations in applying

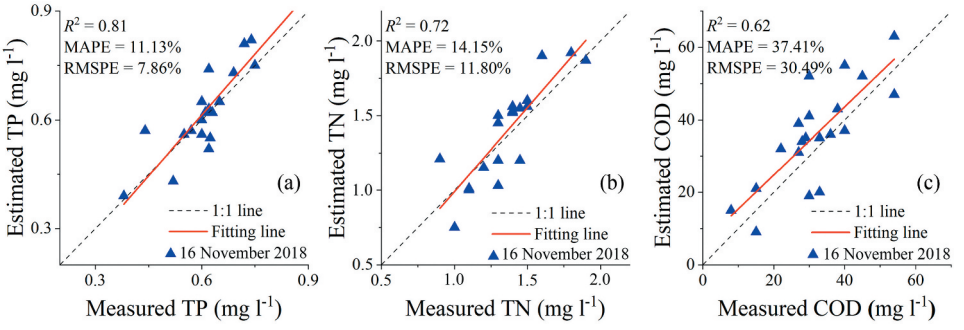

Figure 11. The model performances of estimating TP, TN, and COD on 16 November 2018. (a–c) represent TP, TN, COD, respectively.

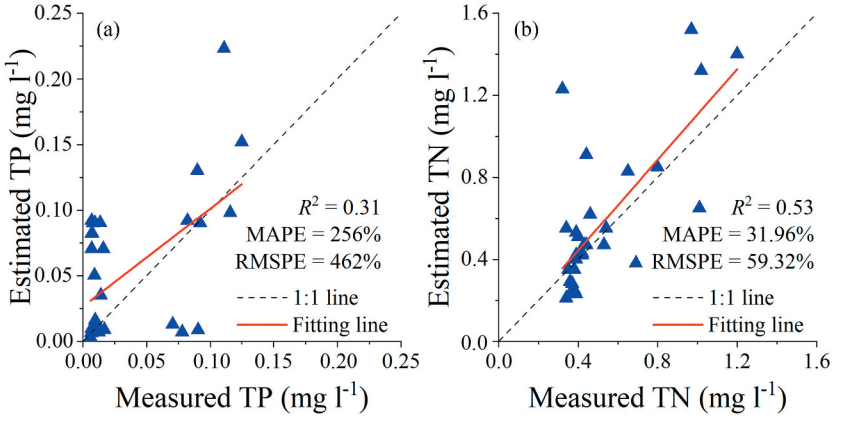

Figure 12. The model performances of estimating TP and TN of Lake Simcoe in 2018. (a) and (b) represent TP and TN, respectively.

to other areas. Band selection and tuning parameters with new data are necessary for different areas.

4.2. Water-quality classification

For water management, another focus is on water-quality classification. Based on the estimated water-quality parameters by machine-learning models, this study further analysed and discussed the water-quality classification. Figure 13 showed the comparisons between the estimated water-quality parameters with the environmental quality standards for surface water (MEE 2002). The environmental quality standards for surface water classify each water-quality parameter into five levels, i.e. Class I to V, indicating water quality from good to poor.

According to Figure 13, TP on both 20 May 2019 and 9 June 2019 was worse than Class V. Serious phosphorus pollution existed in the entire lake surface. TN of the lake surface on both 20 May 2019 and 9 June 2019 covered Class II to V. The average TN on both 20 May 2019 and 9 June 2019 was subject to Class IV. On 20 May 2019, 1.05%, 40.37%,

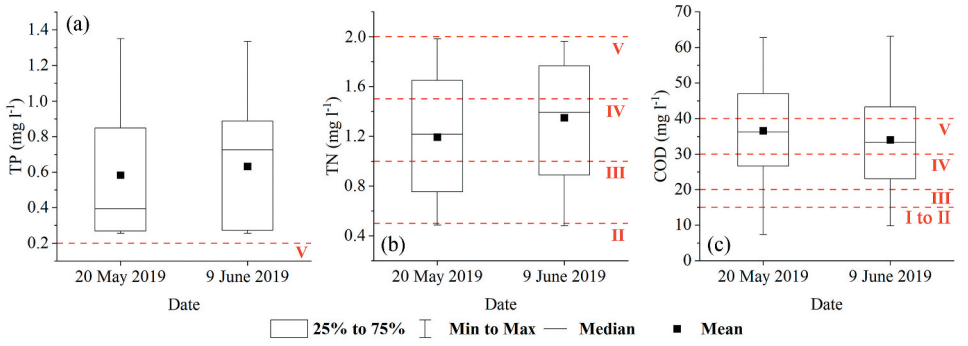

Figure 13. Comparisons of the estimated water-quality parameters and the environmental quality standards for surface water (MEE 2002). (a–c) represent TP, TN, COD, respectively. The red dashed lines represent different water-quality classifications (noted with Roman numerals).

21.33% and 37.25% of the lake surface were subject to Class II, Class III, Class IV and Class V, respectively. On 9 June 2019, the water-quality deteriorated. The Class III, Class IV, and Class V area expanded to 28.63%, 25.65% and 74.35%, respectively. The Class II area almost disappeared. COD of the lake surface on both 20 May 2019 and 9 June 2019 covered Class I to V. In the part of the lake surface, COD was worse than Class V. The average COD on both 20 May 2019 and 9 June 2019 was subject to Class V. On 20 May 2019, 41.21% of the lake surface was worse than Class V. 4.29%, 9.10%, 19.09% and 26.30% of the lake surface were subject to Class I to II, Class III, Class IV and Class V, respectively. On 9 June 2019, the area with COD worse than Class V decreased to 34.49%. The area of Class V and Class II decreased to 22.43% and 3.80%, respectively. The area of Class III and Class IV increased to 11.44% and 27.84%, respectively.

These results indicated the potential feasibility of water-quality classification by remote sensing. Visualization of water-quality classification can be used for integrating water-quality online monitoring and early warning platforms. This helps the water management grasp the water quality in real time and make reasonable decisions. In terms of retrieval models, using machine-learning regression models, water-quality parameters can be estimated in specific values. Then by comparing the estimated results with the water-quality evaluation standards, the spatial distribution of water-quality classification can be acquired. In the future research, machine-learning classification models (e.g. Convolutional Neural Networks (CNN), Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) etc.) can also be considered to directly classify water quality by remote sensing (Mountrakis, Im, and Ogole 2011; Maxwell, Warner, and Fang 2018). In this way, the model complexity will be reduced by a certain extent, and consequently, the cost of model development will be reduced.

4.3. Feasibility of retrieving non-optically active parameters

When plenty of domestic and agricultural wastewater containing N and P enter water-bodies, excessive nutrient elements lead to aquatic plants and algae blooms (Jun Chen and Quan 2012; Chang, Xuan, and Yang 2013). Chl-*a* is one of the most important pigments in phytoplankton, and exists in all eukaryotic algae (Moses et al. 2012). According to the research work of Tan, Cherkauer, and Chaubey (2016), the spectrum

from phytoplankton dominated water experienced low reflectance in red (650 to 750 nm) wavelengths due to the absorption by Chl-*a* and other pigments. Meanwhile, other relevant studies also show that there is a significant correlation between Chl-*a* and the blue (430 to 500 nm), green (543 to 578 nm), red (650 to 750 nm) and NIR (780 to 1100 nm) bands (Ma et al. 2006; 2009; Gitelson et al. 2011; Chang, Xuan, and Yang 2013; Li et al. 2017a). The increase of TP and TN thus changes the optical characteristics of waterbodies.

When industrial effluent containing plenty of organic matter is discharged into waterbodies, insoluble substances directly lead to the turbidity increase. On the other hand, aerobic microorganisms consume oxygen in the water to degrade organic matter. At a certain depth, the dissolved oxygen might gradually decrease to 0, which results in anaerobic condition. In an anaerobic environment, the cyclic state of Fe^{3+} from Fe_2O_3 and $Fe(OH)_3$ in water could be destroyed, and a certain amount of Fe^{2+} accumulates. Meanwhile, S from sulphate and organic sulphur is reduced to H_2S, and the anaerobic environment prevents microorganisms from assimilating H_2S to organic sulphur compounds. Unassimilated H_2S might react with Fe^{2+} to form FeS. The FeS turns the water to black by absorbing on the suspended solids (SS), or being raised into the water by the bubbles generated from anaerobic decomposition (Duan et al. 2014) (Figure 14). The increase of COD thus changes the optical characteristics of waterbodies, such as the increase of reflectance in the green (543–578 nm) and red (650–750 nm) wavelengths (Tan, Cherkauer, and Chaubey 2016). Based on the above analysis, it is feasible to retrieve TP, TN and COD from spectral characteristics.

5. Conclusions

This research developed a machine learning-based strategy for non-optically active water-quality parameters retrieval of small urban waterbodies based on Sentinel-2 imagery. Compared with Landsat TM/ETM+, MODIS and other remote-sensing imagery, Sentinel-2 imagery with high spatiotemporal resolution makes it possible to retrieve water-quality

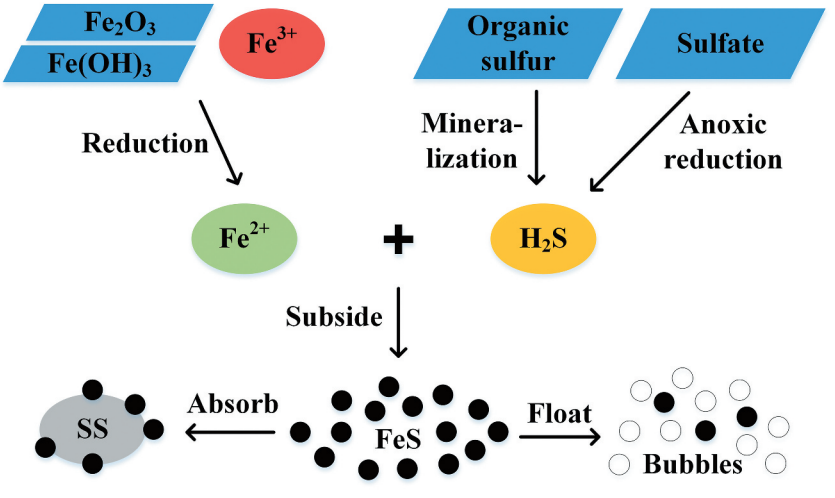

Figure 14. Formation mechanism of black water.

parameters for small urban waterbodies. The most influential Sentinel-2 imagery bands for TP, TN and COD retrieval were B_3, B_4 and B_5. The optimal retrieval accuracy of TP, TN and COD was obtained from the band compositions of '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$', '$B_3 + B_4 + B_5 + B_6 + B_7 + B_8$' and '$B_2 + B_3 + B_5 + B_6 + B_7 + B_8$', respectively. Compared to the spatial interpolation and multiple linear regression, the retrieval performances for non-optically active parameters were significantly improved by the optimized machine-learning models and imagery band selection, especially for TP and TN. The developed models are capable of providing reliable results in local areas, but also have limitations in applying to other areas. Band selection and tuning parameters with new data are necessary for different areas. According to the water-quality mapping by remote-sensing imagery and the interviews of the residents in the neighbourhood, the pollutants, especially the illegal discharges of industrial effluent and domestic sewage, were traced back to the source. Water-quality classification based on the water-quality parameter estimations helps in the integration of water-quality online monitoring and early warning systems. Machine-learning classification models can alternatively be considered for water-quality classification by remote sensing in future research. This study provides a new practical and efficient water-quality monitoring method for managing small waterbodies.

Disclosure statement

The authors declared that they had no conflict of interest over any part or the entirety of the presented study.

Funding

This work was supported by the National Key Research and Development Program of China under [Grant 2016YFC0400709]; Ministry of Science and Technology of the People's Republic of China; and Science and Technology Commission of Tianjin Binhai New Area under [Grant BHXQKJXM-PT-ZJSHJ -2017001].

ORCID

Hongwei Guo http://orcid.org/0000-0003-3663-5908

Data availability statement

The field survey data that support the findings of this study are available from the corresponding author, JJH, upon reasonable request.

References

Beaulant, A. L., G. Perron, J. Kleinpeter, C. Weber, T. Ranchin, and L. Wald. 2008. "Adding Virtual Measuring Stations to a Network for Urban Air Pollution Mapping." *Environment International* 34 (5): 599–605. doi:10.1016/j.envint.2007.12.004.

Brando, V. E., and A. G. Dekker. 2003. "Satellite Hyperspectral Remote Sensing for Estimating Estuarine and Coastal Water Quality." *IEEE Transactions on Geoscience and Remote Sensing* 41: 1378–1387. doi:10.1109/TGRS.2003.812907.

Brezonik, P. L., L. G. Olmanson, J. C. Finlay, and M. E. Bauer. 2015. "Factors Affecting the Measurement of CDOM by Remote Sensing of Optically Complex Inland Waters." *Remote Sensing of Environment*. doi:10.1016/j.rse.2014.04.033.

Brönmark, C., and L. A. Hansson. 2002. "Environmental Issues in Lakes and Ponds: Current State and Perspectives." *Environmental Conservation* 29: 290–307. doi:10.1017/S0376892902000218.

Bugnot, A. B., M. B. Lyons, P. Scanes, G. F. Clark, S. K. Fyfe, A. Lewis, and E. L. Johnston. 2018. "A Novel Framework for the Use of Remote Sensing for Monitoring Catchments at Continental Scales." *Journal of Environmental Management* 217: 939–950. Elsevier Ltd. doi:10.1016/j.jenvman.2018.03.058.

Cao, Z., R. Ma, H. Duan, N. Pahlevan, J. Melack, M. Shen, and K. Xue. 2020. "A Machine Learning Approach to Estimate Chlorophyll-A from Landsat-8 Measurements in Inland Lakes." *Remote Sensing of Environment* 248: 111974. doi:10.1016/j.rse.2020.111974.

Carletti, R., M. Picci, and D. Romano. 2000. "Kriging and Bilinear Methods for Estimating Spatial Pattern of Atmospheric Pollutants." *Environmental Monitoring and Assessment* 63 (2): 341–359. doi:10.1023/A:1006293110652.

Carlson, R. E. 1977. "A Trophic State Index for Lakes." *Limnology and Oceanography* 22: 361–369. doi:10.4319/lo.1977.22.2.0361.

Chang, N. B., K. Bai, and C. F. Chen. 2017. "Integrating Multisensor Satellite Data Merging and Image Reconstruction in Support of Machine Learning for Better Water Quality Management." *Journal of Environmental Management* 201: 227–240. doi:10.1016/j.jenvman.2017.06.045.

Chang, N. B., Z. Xuan, and Y. J. Yang. 2013. "Exploring Spatiotemporal Patterns of Phosphorus Concentrations in a Coastal Bay with MODIS Images and Machine Learning Models." *Remote Sensing of Environment* 134: 100–110. doi:10.1016/j.rse.2013.03.002.

Chen, J., K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, et al. 2019. "A Comparison of Linear Regression, Regularization, and Machine Learning Algorithms to Develop Europe-Wide Spatial Models of Fine Particles and Nitrogen Dioxide." *Environment International* 130: 104934. February. doi:10.1016/j.envint.2019.104934.

Chen, J., and W. Quan. 2012. "Using Landsat/TM Imagery to Estimate Nitrogen and Phosphorus Concentration in Taihu Lake, China." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5: 273–280. doi:10.1109/JSTARS.2011.2174339.

Deng, C., L. Zhang, and Y. Cen. 2019. "Retrieval of Chemical Oxygen Demand through Modified Capsule Network Based on Hyperspectral Data." *Applied Sciences (Switzerland)*. doi:10.3390/app9214620.

Doña, C., N. B. Chang, J. M. Vicente Caselles, A. C. Sánchez, J. Delegido, and B. W. Vannah. 2015. "Integrated Satellite Data Fusion and Mining for Monitoring Lake Water Quality Status of the Albufera de Valencia in Spain." *Journal of Environmental Management* 151: 416–426. doi:10.1016/j.jenvman.2014.12.003.

Duan, H., R. Ma, and C. Hu. 2012. "Evaluation of Remote Sensing Algorithms for Cyanobacterial Pigment Retrievals during Spring Bloom Formation in Several Lakes of East China." *Remote Sensing of Environment* 126: 126–135. doi:10.1016/j.rse.2012.08.011.

Duan, H., R. Ma, S. A. Loiselle, Q. Shen, H. Yin, and Y. Zhang. 2014. "Optical Characterization of Black Water Blooms in Eutrophic Waters." *Science of the Total Environment* 482–483: 174–183. doi:10.1016/j.scitotenv.2014.02.113.

Ferdous, J., M. Tauhid, and U. Rahman. 2020. "Developing an Empirical Model from Landsat Data Series for Monitoring Water Salinity in Coastal Bangladesh." *Journal of Environmental Management* 255: 109861. November 2019. doi:10.1016/j.jenvman.2019.109861.

Gao, Y., J. Gao, H. Yin, C. Liu, T. Xia, J. Wang, and Q. Huang. 2015. "Remote Sensing Estimation of the Total Phosphorus Concentration in a Large Lake Using Band Combinations and Regional Multivariate Statistical Modeling Techniques." *Journal of Environmental Management* 151: 33–43. doi:10.1016/j.jenvman.2014.11.036.

Gholizadeh, M. H., and A. M. Melesse. 2017. "Study on Spatiotemporal Variability of Water Quality Parameters in Florida Bay Using Remote Sensing." *Journal of Remote Sensing & GIS* 6 (3). doi:10.4172/2469-4134.1000207.

Gitelson, A. A., B. C. Gao, R. R. Li, S. Berdnikov, and V. Saprygin. 2011. "Estimation of Chlorophyll-a Concentration in Productive Turbid Waters Using a Hyperspectral Imager for the Coastal Ocean - The Azov Sea Case Study." *Environmental Research Letters*. doi:10.1088/1748-9326/6/2/024023.

Halme, E., P. Pellikka, and M. Mõttus. 2019. "Utility of Hyperspectral Compared to Multispectral Remote Sensing Data in Estimating Forest Biomass and Structure Variables in Finnish Boreal Forest." *International Journal of Applied Earth Observation and Geoinformation* 83: 101942. doi:10.1016/j.jag.2019.101942.

Hoekstra, A. Y., J. Buurman, and K. C. H. Van Ginkel. 2018. "Urban Water Security: A Review." *Environmental Research Letters* 13: 053002. doi:10.1088/1748-9326/aaba52.

Holyer, R. J. 1978. "Toward Universal Multispectral Suspended Sediment Algorithms." *Remote Sensing of Environment* 7: 323–338. doi:10.1016/0034-4257(78)90023-8.

Hou, X., L. Feng, H. Duan, X. Chen, D. Sun, and K. Shi. 2017. "Fifteen-Year Monitoring of the Turbidity Dynamics in Large Lakes and Reservoirs in the Middle and Lower Basin of the Yangtze River, China." *Remote Sensing of Environment* 190: 107–121. doi:10.1016/j.rse.2016.12.006.

Keith, D., J. Rover, J. Green, B. Zalewsky, M. Charpentier, G. Thursby, and J. Bishop. 2018. "Monitoring Algal Blooms in Drinking Water Reservoirs Using the Landsat-8 Operational Land Imager." *International Journal of Remote Sensing* 39 (9): 2818–2846. doi:10.1080/01431161.2018.1430912.

Kishino, M., A. Tanaka, and J. Ishizaka. 2005. "Retrieval of Chlorophyll A, Suspended Solids, and Colored Dissolved Organic Matter in Tokyo Bay Using ASTER Data." *Remote Sensing of Environment* 99 (1–2): 66–74. doi:10.1016/j.rse.2005.05.016.

Kurt, A., B. Gulbagci, F. Karaca, and O. Alagha. 2008. "An Online Air Pollution Forecasting System Using Neural Networks." *Environment International* 34 (5): 592–598. doi:10.1016/j.envint.2007.12.020.

Le, C., Y. Li, Y. Zha, D. Sun, C. Huang, and H. Zhang. 2011. "Remote Estimation of Chlorophyll a in Optically Complex Waters Based on Optical Classification." *Remote Sensing of Environment* 115 (2): 725–737. doi:10.1016/j.rse.2010.10.014.

Li, J., C. Hu, Q. Shen, B. B. Barnes, B. Murch, L. Feng, M. Zhang, and B. Zhang. 2017a. "Recovering Low Quality MODIS-Terra Data over Highly Turbid Waters through Noise Reduction and Regional Vicarious Calibration Adjustment: A Case Study in Taihu Lake." *Remote Sensing of Environment*. doi:10.1016/j.rse.2017.05.027.

Li, Y., Y. Zhang, K. Shi, G. Zhu, Y. Zhou, Y. Zhang, and Y. Guo. 2017b. "Monitoring Spatiotemporal Variations in Nutrients in a Large Drinking Water Reservoir and Their Relationships with Hydrological and Meteorological Conditions Based on Landsat 8 Imagery." *Science of the Total Environment*. doi:10.1016/j.scitotenv.2017.05.075.

Lunetta, R. S., J. F. Knight, H. W. Paerl, J. J. Streicher, B. L. Peierls, T. Gallo, J. G. Lyon, T. H. Mace, and C. P. Buzzelli. 2009. "Measurement of Water Colour Using AVIRIS Imagery to Assess the Potential for an Operational Monitoring Capability in the Pamlico Sound Estuary, USA." *International Journal of Remote Sensing*. doi:10.1080/01431160802552801.

Ma, R., D. Pan, H. Duan, and Q. Song. 2009. "Absorption and Scattering Properties of Water Body in Taihu Lake, China: Backscattering." *International Journal of Remote Sensing* 30 (9): 2321–2335. doi:10.1080/01431160802549385.

Ma, R., J. Tang, J. Dai, Y. Zhang, and Q. Song. 2006. "Absorption and Scattering Properties of Water Body in Taihu Lake, China: Absorption." *International Journal of Remote Sensing* 27 (19): 4277–4304. doi:10.1080/01431160600851835.

Mannino, A., M. G. Novak, S. B. Hooker, K. Hyde, and D. Aurin. 2014. "Algorithm Development and Validation of CDOM Properties for Estuarine and Continental Shelf Waters along the Northeastern U.S. Coast." *Remote Sensing of Environment* 152: 576–602. doi:10.1016/j.rse.2014.06.027.

Mathew, M. M., N. Srinivasa Rao, and V. R. Mandla. 2017. "Development of Regression Equation to Study the Total Nitrogen, Total Phosphorus and Suspended Sediment Using Remote Sensing Data in Gujarat and Maharashtra Coast of India." *Journal of Coastal Conservation* 21: 917–927. doi:10.1007/s11852-017-0561-1.

Maxwell, A. E., T. A. Warner, and F. Fang. 2018. "Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review." *International Journal of Remote Sensing* 39: 2784–2817. doi:10.1080/01431161.2018.1433343.

McFeeters, S. K. 1996. "The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features." *International Journal of Remote Sensing* 17 (7): 1425–1432. doi:10.1080/01431169608948714.

MEE (Ministry of ecology and environment of the people's Republic of China). 2002. *Environmental Quality Standard for Surface Water (GB3838-2002)*. Beijing: China Environmental Science Press.

Moses, W. J., A. A. Gitelson, R. L. Perk, D. Gurlin, D. C. Rundquist, B. C. Leavitt, T. M. Barrow, and P. Brakhage. 2012. "Estimation of Chlorophyll-a Concentration in Turbid Productive Waters Using Airborne Hyperspectral Data." *Water Research* 46: 993–1004. doi:10.1016/j.watres.2011.11.068.

Moses, W. J., A. A. Gitelson, S. Berdnikov, and V. Povazhnyy. 2009. "Satellite Estimation of Chlorophyll-a Concentration Using the Red and NIR Bands of MERISThe Azov Sea Case Study." *IEEE Geoscience and Remote Sensing Letters*. doi:10.1109/LGRS.2009.2026657.

Mountrakis, G., J. Im, and C. Ogole. 2011. "Support Vector Machines in Remote Sensing: A Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (3): 247–259. doi:10.1016/j.isprsjprs.2010.11.001.

O'Reilly, J. E., S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain. 1998. "Ocean Color Chlorophyll Algorithms for SeaWiFS." *Journal of Geophysical Research: Oceans* 103: 24937–24953. doi:10.1029/98JC02160.

Olmanson, L. G., P. L. Brezonik, and M. E. Bauer. 2013. "Airborne Hyperspectral Remote Sensing to Assess Spatial Distribution of Water Quality Characteristics in Large Rivers: The Mississippi River and Its Tributaries in Minnesota." *Remote Sensing of Environment* 130: 254–265. doi:10.1016/j.rse.2012.11.023.

Pahlevan, N., B. Smith, J. Schalles, C. Binding, Z. Cao, R. Ma, K. Alikas, et al. 2020. "Seamless Retrievals of Chlorophyll-A from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in Inland and Coastal Waters: A Machine-Learning Approach." *Remote Sensing of Environment* 240:111604. doi:10.1016/j.rse.2019.111604.

Rapinel, S., C. Mony, L. Lecoq, B. Clément, A. Thomas, and L. Hubert-Moy. 2019. "Evaluation of Sentinel-2 Time-Series for Mapping Floodplain Grassland Plant Communities." *Remote Sensing of Environment* 223: 115–129. January. Elsevier. doi:10.1016/j.rse.2019.01.018.

Ritchie, J. C., F. R. Schiebe, and J. R. McHenry. 1976. "Remote Sensing of Suspended Sediments in Surface Waters." *Photogrammetric Engineering & Remote Sensing*. 42 (12): 1539–1545.

Ritchie, J. C., P. V. Zimba, and J. H. Everitt. 2003. "Remote Sensing Techniques to Assess Water Quality." *Photogrammetric Engineering and Remote Sensing* 69: 695–704. doi:10.14358/PERS.69.6.695.

Shahzad, M. I., M. Meraj, M. Nazeer, I. Zia, A. Inam, K. Mehmood, and H. Zafar. 2018. "Empirical Estimation of Suspended Solids Concentration in the Indus Delta Region Using Landsat-7 ETM+ Imagery." *Journal of Environmental Management* 209: 254–261. doi:10.1016/j.jenvman.2017.12.070.

Shao, M., X. Tang, Y. Zhang, and W. Li. 2006. "City Clusters in China: Air and Surface Water Pollution." *Frontiers in Ecology and the Environment* 4 (7): 353–361.

Shenglei, W., L. Junsheng, Z. Bing, S. Qian, Z. Fangfang, and L. Zhaoyi. 2016. "A Simple Correction Method for the MODIS Surface Reflectance Product over Typical Inland Waters in China." *International Journal of Remote Sensing* 37 (24): 6076–6096. doi:10.1080/01431161.2016.1256508.

Shi, K., Y. Zhang, G. Zhu, X. Liu, Y. Zhou, H. Xu, B. Qin, G. Liu, and Y. Li. 2015. "Long-Term Remote Monitoring of Total Suspended Matter Concentration in Lake Taihu Using 250 M MODIS-Aqua Data." *Remote Sensing of Environment* 164: 43–56. doi:10.1016/j.rse.2015.02.029.

Shuchman, R. A., M. Sayers, G. L. Fahnenstiel, and G. Leshkevich. 2013. "A Model for Determining Satellite-Derived Primary Productivity Estimates for Lake Michigan." *Journal of Great Lakes Research* 39: 46–54. doi:10.1016/j.jglr.2013.05.001.

Tan, J., K. A. Cherkauer, and I. Chaubey. 2016. "Developing A Comprehensive Spectral-Biogeochemical Database of Midwestern Rivers for Water Quality Retrieval Using

Remote Sensing Data: A Case Study of the Wabash River and Its Tributary, Indiana." *Remote Sensing* 8 (6): 517. doi:10.3390/rs8060517.

Vapnik, V. N. 1995. "Adaptive and Learning Systems for Signal Processing, Communications and Control." In *The Nature of Statistical Learning Theory*, edited by Michael Jordan, 138–167. New York: Springer-Verlag. doi:10.2307/1271368.

Vignolo, A., A. Pochettino, and D. Cicerone. 2006. "Water Quality Assessment Using Remote Sensing Techniques: Medrano Creek, Argentina." *Journal of Environmental Management* 81 (4): 429–433. doi:10.1016/j.jenvman.2005.11.019.

Wang, J., M. Hu, B. Gao, H. Fan, and J. Wang. 2019. "A Spatiotemporal Interpolation Method for the Assessment of Pollutant Concentrations in the Yangtze River Estuary and Adjacent Areas from 2004 to 2013." *Environmental Pollution*. doi:10.1016/j.envpol.2019.05.132.

Wang, S., J. Li, B. Zhang, E. Spyrakos, A. N. Tyler, Q. Shen, F. Zhang, et al. 2018. "Trophic State Assessment of Global Inland Waters Using a MODIS-Derived Forel-Ule Index." *Remote Sensing of Environment*. doi:10.1016/j.rse.2018.08.026.

Wang, S., J. Li, B. Zhang, Z. Lee, E. Spyrakos, L. Feng, C. Liu, et al. 2020. "Changes of Water Clarity in Large Lakes and Reservoirs across China Observed from Long-Term MODIS." *Remote Sensing of Environment*. doi:10.1016/j.rse.2020.111949.

Wang, Y., H. Xia, J. Fu, and G. Sheng. 2004. "Water Quality Change in Reservoirs of Shenzhen, China: Detection Using LANDSAT/TM Data." *Science of the Total Environment*. doi:10.1016/j.scitotenv.2004.02.020.

Wu, C., J. Wu, Q. Jiaguo, L. Zhang, H. Huang, L. Lou, and Y. Chen. 2010. "Empirical Estimation of Total Phosphorus Concentration in the Mainstream of the Qiantang River in China Using Landsat TM Data." *International Journal of Remote Sensing* 31: 2309–2324. doi:10.1080/01431160902973873.

Wu, M., W. Zhang, X. Wang, and D. Luo. 2009. "Application of MODIS Satellite Data in Monitoring Water Quality Parameters of Chaohu Lake in China." *Environmental Monitoring and Assessment* 148 (1–4): 255–264. doi:10.1007/s10661-008-0156-2.

Xiong, Y., Y. Ran, S. Zhao, H. Zhao, and Q. Tian. 2020. "Remotely Assessing and Monitoring Coastal and Inland Water Quality in China: Progress, Challenges and Outlook." *Critical Reviews in Environmental Science and Technology* 50 (12): 1266–1302. doi:10.1080/10643389.2019.1656511.