Sustainable Environment Research

RESEARCH Open Access

Machine-learning-estimation of high-spatiotemporal-resolution chlorophyll-*a* concentration using multi-satellite imagery

Wachidatin Nisaul Chusnah[1], Hone-Jay Chu[1*], Tatas[1] and Lalu Muhamad Jaelani[2]

Abstract

Chlorophyll-*a* concentration for quantifying phytoplankton biomass is commonly used as an indicator for evaluating the trophic level of lakes and water quality. This research aimed to develop a high spatiotemporal-resolution model for the retrieval of chlorophyll-*a* in inland water. Firstly, the machine learning based models considering Sentinel-2 Multispectral Instrument and Sentinel-3 Ocean and Land Color Instrument (OLCI) images were applied to estimate chlorophyll-*a* concentrations ($R^2 = 0.873$ and 0.822, respectively). The spatiotemporal fusion was performed to fuse the OLCI and MSI chlorophyll-*a* images with low temporal resolution but fine spatial-resolution, and with high temporal resolution but coarse spatial-resolution. The random forest was applied to fuse images from two distinct sensors, and to refine the spatial resolution of OLCI estimations to be the same as those of Sentinel-2 MSI. Results showed that the spatiotemporal fusion can estimate dense-temporal 10 m spatial resolution chlorophyll-*a* concentration in the Tsengwen Reservoir (Root-Mean-Square Error, RMSE = 1.25–1.47 μg L^{-1}). The spatiotemporal fusion model was effectively applied to determine high spatiotemporal-resolution chlorophyll-*a* measurements in the aquatic system.

Keywords Chlorophyll-*a* estimation, Spatiotemporal fusion, Machine learning, Band ratio

1 Introduction

An inland water system is a significant water source, particularly for human necessities. Global inland freshwater ecosystems undergo extensive changes due to the increase in human demand for fresh water in the last century. Agricultural runoff, industrial waste, excrement, and other wastes generated by human activities across water bodies have increased the anthropogenic input of nitrogen and phosphorus, resulting in extensive eutrophication [1]. An evident indication of eutrophication is the fast growth and increased amount of suspended algae or phytoplankton. Harmful algal bloom is an environmental problem because it causes the discoloration of affected waters and imbalance among organisms in aquatic ecosystems [2, 3]. Phytoplankton biomass is one of the critical biologically sensitive elements for assessing the ecological status of water and health risks of aquatic ecosystems [2]. Sufficient data are necessary to be acquired for a lake monitoring system in terms of bloom information because such information is spatially and temporally heterogeneous [3]. Chlorophyll-*a* concentration for quantifying phytoplankton biomass is commonly used as an indicator for assessing the trophic level of lakes [4, 5] and representing the state of water quality [6].

Remote sensing and satellite imaging have been applied to estimate chlorophyll-*a* in inland water in recent years for sustainable water resources management [7]. Numerous researches were reported the use of a band ratio algorithm for estimating coastal and inland water

*Correspondence:
Hone-Jay Chu
honejaychu@geomatics.ncku.edu.tw
[1] Department of Geomatics, National Cheng Kung University, Tainan 701401, Taiwan
[2] Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

characteristics [8, 9]. The algorithms have been developed for chlorophyll-*a* estimation in inland waters; however, the capability to decrease interference on reflectance from a spectral band ratio algorithm is preferable [7]. The chlorophyll-*a* estimation models in waters were developed such as three-band, red–near infrared (NIR), blue–green, green–blue, and red–green models [10–15]. The blue–green band ratio was frequently used for estimating chlorophyll-*a* concentration in clear waters (Case 1 water) [15] because optical properties of clear waters are controlled by phytoplankton. In Case 2 water, turbidity is high and the algorithm experiences difficulty in estimating chlorophyll-*a* in this spectral region due to low correlation of the presence of chromophoric dissolved organic matter and phytoplankton. The red to NIR ratio is used for measuring in turbid waters to avoid non-algal particleabsorption[10]. Moreover, employing machine learning approaches is feasible and effective for modeling water quality parameters in aqueous environments [6, 16]. The application of machine learning algorithms to global ocean and inland water quality retrievals is particularly suited for improving the accuracy of chlorophyll-*a* estimation [16, 17]. Many studies considered machine learning approaches for estimating water quality parameters using satellite imagery [18, 19]. Machine learning provides a promising way for operational use of regional water quality monitoring [18]. Compared to physical models, machine learning algorithms are an approach for handling inversion problems without theoretical analyses of spectral information. Machine learning is the most highly efficient and effective approach to evaluate relations between the water quality parameters and corresponding remote-sensing reflectance [18]. In the current research, machine learning was used to implement the band ratio algorithm and estimate chlorophyll-*a* maps.

The Sentinel-2 and Sentinel-3 are equipped with sensors for measuring and estimating regional chlorophyll-*a* concentrations [20–22]. Sentinel-2 and Sentinel-3 satellite instruments provide a good prospect for chlorophyll-*a* retrieval in complex waters wherein a good spatiotemporal resolution is required [20–22]. Sentinel-2, carries onboard the Multispectral Instrument (MSI), which is a high-resolution satellite instrument that provides orthoimage bottom-of-atmosphere (BoA) corrected reflectance products. Sentinel-2 MSI, which has been on board since June 2015, provides open-access satellite products with the spatial resolution, offering 10, 20, and 60 m through 13 spectral bandwidths. MSI is favorable for advanced chlorophyll-*a* monitoring in inland water because this instrument is qualified for the red-edge and the red bands, which are close to the phytoplankton peak spectral reflectance of 700 nm wavelength. The ability of the MSI band ratio algorithm for chlorophyll-*a* retrieval has been evaluated [7, 20, 23]. Sentinel-3 carries onboard the Ocean and Land Color Instrument (OLCI) sensor, providing broad-coverage images with a spatial resolution of 300 m and rapid revisit time that supports marine applications. The Sentinel-3 OLCI is a sensor for the rapid chlorophyll-*a* retrieval because it has a shorter revisit time than the MSI. The rapid revisiting time of OLCI provided the chances to produce an estimation of chlorophyll-*a* in high temporal resolution. However, the spatial resolution of Sentinel-3 OLCI images is too coarse for small inland reservoirs. In this study, spatiotemporal fusion was applied to refine the spatial resolution of Sentinel-3 OLCI estimations to be the same as those of Sentinel-2 MSI. The data fusion is performed to fuse two satellite image data with high spatial-resolution but low temporal-resolution, and with high temporal-resolution but coarse spatial-resolution [24]. Spatiotemporal fusion aims at fusing sparse fine-resolution images with frequent coarse-resolution images i.e. fusing dense-temporal coarse-spatial-resolution, and sparse-temporal fine-spatial resolution images to create fine spatiotemporal resolution images [24]. The primary categories of spatiotemporal fusion from processing level includes data level, information level, and decision level. Previous studies incorporated fusion at the data level between images that are spatially and spectrally correlated for water quality monitoring [25, 26]. In the current study, machine learning-based spatiotemporal fusion was applied with information level to generate high-spatiotemporal-resolution results by combining chlorophyll-*a* maps from MSI and OLCI data.

High-spatiotemporal-resolution chlorophyll-*a* images in inland waters were estimated by machine learning. Firstly, the various spatiotemporal resolution chlorophyll-*a* maps from MSI and OLCI were estimated mainly using random forest (model A and B). These chlorophyll-*a* estimations were then fused into the high-spatiotemporal-resolution chlorophyll-*a* maps. Random forest identified the relation between MSI and OLCI chlorophyll-*a* (model C). After training, the fusion model simulated high-spatial-resolution chlorophyll-*a* based on OLCI (high-temporal-resolution) chlorophyll-*a*.

2 Materials and study area

2.1 Study sites

The reservoirs exhibit various physical and biogeochemical characteristics, and they are spread in Taiwan (Fig. 1): Northern Taiwan (Shimen and Baoshan Reservoirs), North-west Taiwan (Yongheshan, Mingde, and Liyutan Reservoirs), and Southern Taiwan (Tsengwen, Wushantou, Chingmien, Chengcing, Agongdian, and Fengshan Reservoirs). The water area and depth of

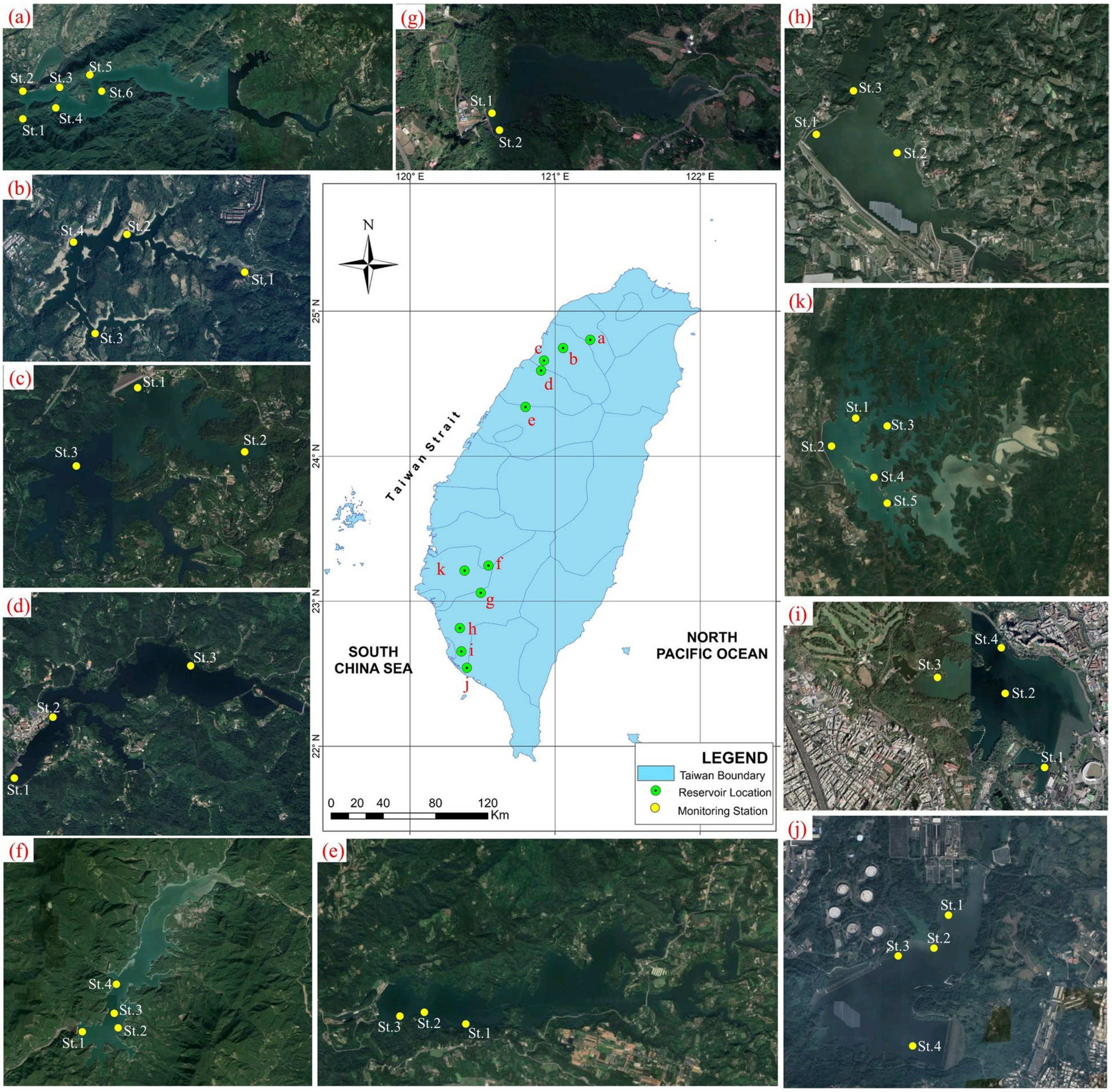

Fig. 1 Spatial locations of Taiwan's reservoirs as study sites. Distribution of monitoring stations: **a** Shimen Reservoir, **b** Baoshan Reservoir, **c** Yongheshan Reservoir, **d** Mingde Reservoir, **e** Liyutan Reservoir, **f** Tsengwen Reservoir, **g** Chingmien Reservoir, **h** Agongdian Reservoir, **i** Chengcing Reservoir, **j** Fengshan Reservoir, and **k** Wushantou Reservoir

reservoirs are provided in Supplementary materials, Table S1. Residential and industrial areas near reservoirs also influence anthropogenic pollution in Taiwan's major reservoirs [27]. Approximately 80% of the study sites (9 of 11 reservoirs) are categorized under mesotrophic lake condition, while 20% of the study sites are categorized under eutrophic lake level. With regard to the turbidity data from Taiwan's Environmental Protection Administration (EPA), most of the study sites are categorized as fresh water, with a turbidity level of < 10 nephelometric turbidity units.

2.2 Chlorophyll-*a* Observation

EPA in Taiwan provides environmental information through an open data platform (https://data.epa.gov.tw/en/) that includes spatially and temporally resolved chlorophyll-*a* observation. Figure 1 shows the locations of monitoring stations (sampling sites) across reservoirs.

The data collected from 11 reservoirs are used to expand the observation data to obtain a wide range of chlorophyll-*a* concentrations for model development. The observation data of chlorophyll-*a* were collected from 2019 to 2021 at a sample depth of 0.5 m. The dataset is divided into two categories: Datasets A and B. Dataset A comprises chlorophyll-*a* observation data that correspond to Sentinel-2 satellite images, and is used to develop a chlorophyll-*a* estimation (Chla_S2) model for MSI (Model A). Dataset B consists of chlorophyll-*a* observation data that correspond to Sentinel-3 satellite images, and is used to develop an estimation model under OLCI environment (Model B) because the spatial resolution of OLCI (300 m) limits the utilization of small reservoirs. A single pixel in OLCI that is covered with more than one chlorophyll-*a* monitoring station is excluded for training. Consequently, 433 and 218 chlorophyll-*a* observation data from study sites were collected for Dataset A and B, respectively. Descriptive statistic of dataset A and B are shown in Tables S2 and S3.

2.3 Remote sensing images

Sentinel-2 Level-2A satellite imagery is retrieved as orthoimage BoA reflectance (atmospheric corrections included). Sentinel-2 Level-2A imagery is available under open access in the Google Earth Engine (GEE) catalog [28]. GEE consists of a multi-petabyte analysis of large geospatial remotely sensed data, and it is considered a high-performance cloud computing platform. The quality of the Sentinel-3A OLCI dataset in GEE is not preferable for scientific applications. The Sentinel-3 dataset in the GEE repository lacks atmospheric data and observation geometry, which are requirements for atmospheric correction and further image processing. However, GEE transforms the original data, resulting in the deviation of pixel values from the original data. Sentinel-3A OLCI Level-1 data are derived from an open-source platform provided by NASA (https://ladsweb.modaps.eosdis.nasa.gov/). OLCI has an atmospheric issue; hence, radiometric correction is necessary before it can be used [29]. Image correction for atmospheric effects (iCOR) is a method for the atmospheric correction instrument of satellite data over land and coastal and inland waters [30, 31]. The iCOR atmosphere correction module involves Sentinel 3A-OLCI top-of-atmosphere radiance to obtain atmospheric-corrected data. iCOR's functional applications, continuous development, and dexterity to different processing infrastructure provide a stable, efficient, and high-quality processing performance, making iCOR qualified for various further implementations.

Cloud-free days should be preferably designated for the observation sampling of chlorophyll-*a* and satellite overpass. However, many conditions and limitations cause the observation sampling data to not precisely correspond with cloud-free days. Temporal matching is allowed to acquire sufficient data when no matchup point exists between observation and satellite overpass [32]. If no corresponding image is available during the sampling date of water quality observation, then the closest temporal image within a time window of observation data is obtained, i.e., 3 days. Ensuring an adequate number of matchup points is a strategy for developing a model. The low synchronization of the matchup between chlorophyll-*a* observation and satellite images significantly affects the degree of correlation [33]. The number of observations within the time window is shown in Tables S4 and S5. In addition, the rapid change in water quality may be unsuitable for supporting model construction [32]. Satellite image acquisition and observation date are determined under a 3-d time window. A total of 119 MSI images and 55 OLCI images were used for generating Model A and B.

3 Methods

This research is conducted using machine learning models: (A) chlorophyll-*a* estimation model for MSI, (B) chlorophyll-*a* estimation model for OLCI, and data fusion used for high spatiotemporal chlorophyll-*a* mapping (Fig. 2). Under Sentinel-2 MSI imagery and chlorophyll-*a* field data, Model A transfers from the Sentinel-2 band ratio (BR_S2) to the concentration of chlorophyll-*a*. The proposed technique is also performed under Sentinel-3 OLCI environment and considering chlorophyll-*a* observation for control points (estimation model B). Under the specified band ratios, the model B uses the Sentinel-3 band ratio to estimate chlorophyll-*a*. The current research focuses on using band combinations that are critical to chlorophyll-*a* estimation in water. Six developed and frequently used band ratios, including green–blue and red–NIR in terms of two- and three-band ratios (Table 1), are selected for evaluating the performance and degree of correlation toward chlorophyll-*a* estimation. Eventually, the high spatial resolution of MSI and high temporal resolution of OLCI are integrated to obtain a higher spatiotemporal resolution estimation. In data fusion, the random forest model identifies the relation between Sentinel-3 and Sentinel-2-based chlorophyll-*a*. The data fusion can estimate chlorophyll-*a* of high-spatiotemporal resolution with Sentinel-2 spatial resolution and Sentinel-3 temporal resolution (Fig. 2). Before training the model, the outlier detection strategies are performed. Consequently, a significant error exerts a relatively greater effect on total square error. Root-mean-square error (RMSE) or standard error residual, coefficient of determination (R^2), and mean absolute

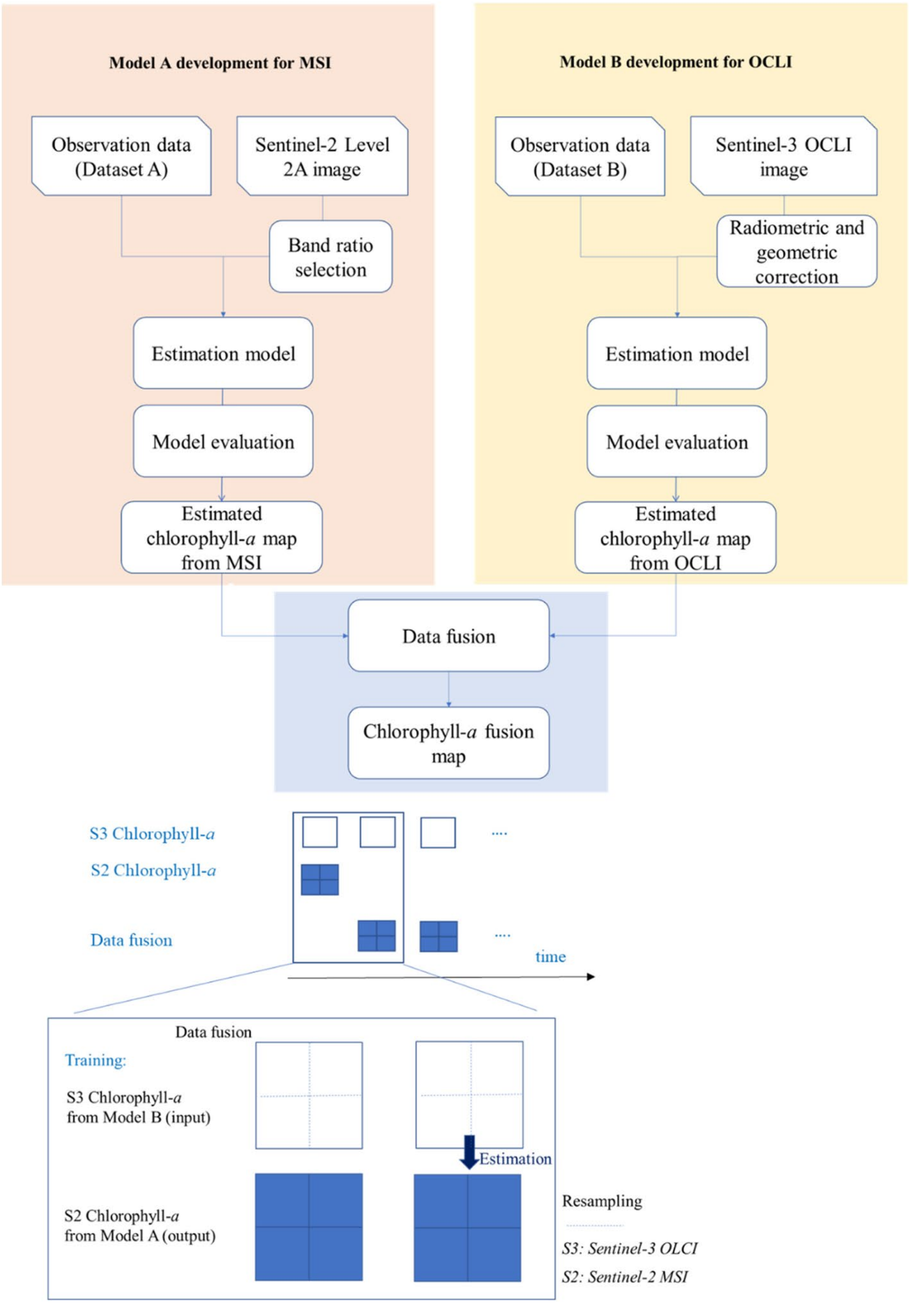
Model A development for MSI
Observation data (Dataset A)
Sentinel-2 Level 2A image
Band ratio selection
Estimation model
Model evaluation
Estimated chlorophyll-*a* map from MSI
Model B development for OCLI
Observation data (Dataset B)
Sentinel-3 OCLI image
Radiometric and geometric correction
Estimation model
Model evaluation
Estimated chlorophyll-*a* map from OCLI
Data fusion
Chlorophyll-*a* fusion map
S3 Chlorophyll-*a*
S2 Chlorophyll-*a*
Data fusion
....
....
time
Data fusion
Training:
S3 Chlorophyll-*a* from Model B (input)
S2 Chlorophyll-*a* from Model A (output)
Estimation
Resampling
S3: Sentinel-3 OLCI
S2: Sentinel-2 MSI

Fig. 2 Research workflow (Estimation model A: chlorophyll-*a* estimated from MSI; Estimation model B: chlorophyll-*a* estimated from OLCI; Data fusion: high-spatiotemporal-resolution chlorophyll-*a* estimated from the fusion of chlorophyll-*a* Models A and B)

error (MAE) are used to identify the performance of the generated model. The dataset is divided into 80 and 20% for training and testing in model A and B, respectively. At least one representative location is picked as testing data in most lakes. In addition, the waterbody is detected by utilizing the normalized difference water index using green and NIR wavelengths.

3.1 Estimation model A development for MSI

The model transfers from BR_S2 to the concentration of chlorophyll-*a*. The model input is used by BR_S2 from each observation. After the selection of band ratios, the red–NIR and green–blue band ratios are used. The model output represents the Chla_S2. The function f from BR_S2 to Chla_S2 in Model A is expressed as follows.

Table 1 Developed band ratios of chlorophyll-*a* estimation in inland water

Index	Band Ratio (BR)	Reference
BR 1	$R(560)/R(443)$	[7, 32]
BR 2	$R(560)/R(492)$	[7]
BR 3	$R(708)/R(665)$	[11]
BR 4	$R(753)/R(665)$	[11]
BR 5	$[R(665)^{-1} - R(708)^{-1}] \times R(753)$	[11]
BR 6	$[R(708) + R(753)] \times R\ (665)^{-1}$	[7, 33]

R: surface reflectance

$$\text{Model A, f} : \text{BR\_S2} \in \mathbb{R}^2 \rightarrow \text{Chla\_S2} \in \mathbb{R}^1 \quad (1)$$

where $\mathbb{R}^n$: n-dimensional vector. The developed machine learning approach, including random forest or decision tree, is an established model for identifying the function from BR_S2 to Chla_S2. Two band ratios are used as the inputs in a linear regression with the intercept. Random forest and decision tree are nonparametric supervised learning methods with moderately convenient algorithms that are capable of handling large datasets [34]. The chlorophyll-*a* estimation model is generated under scikit-learn machine learning in Python. Three models i.e. decision tree, random forest, and linear regression are used for test and comparison. In terms of generating a machine learning model for estimating chlorophyll-*a*, the observation dataset is randomly divided into training (80%, $N=327$) and testing (20%, $N=82$) data, employing the same input for all the models. The estimation model uses a regression technique that is processed under scikit-learn 0.21 supported Python 3.7. To obtain the best hyperparameter scheme, iterative tuning is utilized to determine the optimal value for Model A. The model minimizes the L2 loss using the mean of each terminal. To obtain the best hyperparameters scheme, iterative tuning is utilized to determine the optimal parameters for random forest and decision trees (parameters in Table S6).

3.2 Estimation model B development for OLCI

The model transfers from Sentinel-3 band ratios (BR_S3) to chlorophyll-*a* concentration. The band ratios are selected, i.e. red–NIR and green–blue band ratios are used here. The model inputs are BR_S3 from each observation, and the output is the chlorophyll-*a* concentration vector (Chla_S3).

$$\text{Model B, f} : \text{BR\_S3} \in \mathbb{R}^2 \rightarrow \text{Chla\_S3} \in \mathbb{R}^1 \quad (2)$$

Here, random forest is applied as the model (parameters in Table S6). With regard to generating a random forest model for estimating chlorophyll-*a* under OLCI environment, the observation dataset B is divided into training (80%, $N=175$) and testing (20%, $N=43$) data. The data used in the further development of Model B are Sentinel-3A OLCI satellite images and the corresponding chlorophyll-*a* observation data (Dataset B). The utilization of OLCI is due to the short revisit time to improve the low temporal resolution in MSI.

3.3 Data fusion

Spatiotemporal fusion is developed to obtain an estimation model for evaluating variability of rapid changes in a small reservoir. Sentinel-3 OLCI exhibits a limitation in result interpretation because of its lower spatial resolution. In data fusion, OLCI is combined with MSI and complies with the information under MSI environment. The concept is related to the major technique of data fusion that injects the detailed information selected from low-resolution images with high spatial resolution [35]. The fusion transfers from resampled Sentinel-3 chlorophyll-*a* (Chla_S3′ as input) to Chla_S2 as output for data fusion.

$$\text{Data fusion, f} : \text{Chla\_S3}' \in \mathbb{R}^1 \rightarrow \text{Chla\_S2} \in \mathbb{R}^1 \quad (3)$$

First, Sentinel-3-based chlorophyll-*a* is resampled to the same resolution as Sentinel-2 MSI by using bicubic interpolation (Chla_S3'). For data integration of OLCI and MSI, the random forest model identifies the relation between Sentinel-3 and Sentinel-2-based chlorophyll-*a*. After training, the fusion model can simulate high-spatial-resolution chlorophyll-*a* based on Sentinel-3 chlorophyll-*a*. The best model is obtained from 10 tree estimators, and the model parameters are shown in the Fig. 6.

4 Results

4.1 Chlorophyll-*a* estimation from MSI

Figure 3 presents the chlorophyll-*a* estimation performance of multiple combinations of band ratios by using the three approaches. The model with $[R(665)^{-1} - R(708)^{-1}] \times R(753)$, (BR 5) and $R(560)/R(492)$ (BR 2) produces an excellent result by using the random forest method. The model achieves an RMSE of 6.99 μg L^{-1} and a high correlation of $R^2=0.873$ (Fig. 4, Table 2), outperforming decision tree ($R^2=0.807$) and multiple linear regression ($R^2=0.343$). This proposed model consists of the BR 2 and 5 combination of chlorophyll-*a* retrieval algorithms (Fig. 3). A combination of band ratios will produce a reliable result due to the sensitivity levels and complexity of each band. Using multiple-band ratio can boost the performance metric (R^2) and produce a robust model; the greatest contribution to model improvement is the combination of red–NIR and green–blue band ratios [36]. As a red–NIR three-band model, BR 5 has been proven to be effective in improving

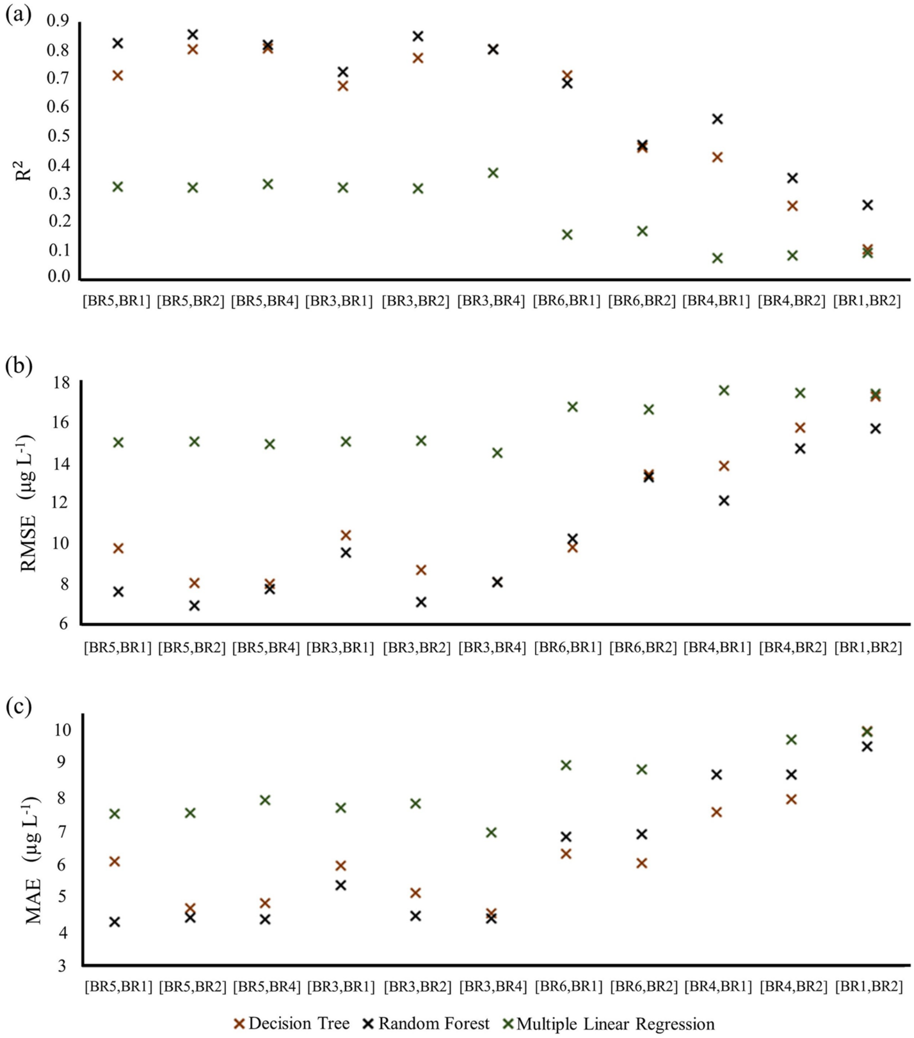

Fig. 3 Performance of model A from linear regression, decision tree, and random forest in chlorophyll-*a* retrieval by using multiple band ratios assessed by **a** R^2, **b** RMSE, and **c** MAE

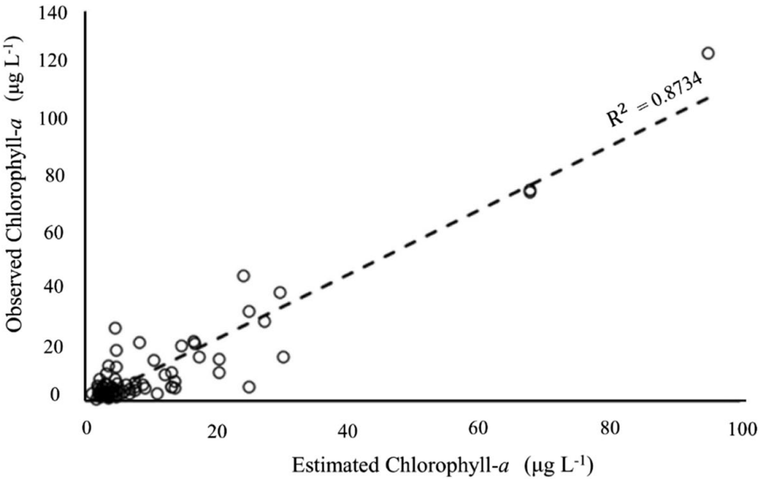

Fig. 4 Validation of the best model (random forest) by using 20% of the dataset of the observed and estimated chlorophyll-*a* produced by the [BR 5, BR 2] algorithm

Table 2 RF model performance for training and testing in model A and B

Model	R^2		RMSE (μgL^{-1})	
	Training	Testing	Training	Testing
A	0.882	0.873	5.670	6.994
B	0.801	0.822	1.385	1.185

model accuracy. The most important factor is described by$[R_{665}^{-1} - R_{708}^{-1}]$ for BR 5 because the fittest estimation model is obtained from the combination that consists of BR 5 (RMSE) = 6.99–7.82 μg L^{-1}) input variable through the random forest method. The existence of reflectance at 753 nm combined with BR 5 is highly correlated with chlorophyll-*a* because of its ability to decrease the scattering effect of inorganic particles [37].

The multiple-band ratio random forest model is adopted to process Sentinel-2 MSI data and map inland water. As the largest reservoir in Taiwan, the Tsengwen Reservoir is selected for demo purposes. Figure 5 illustrates the capability of random forest to map and present chlorophyll-*a* distribution on February 2019, April 2019, February 2020, and May 2020 with RMSEs of 1.21, 1.49, 1.10 and 1.23 μg L^{-1}, respectively. The Tsengwen Reservoir reached the maximum capacity during the wet season (April–July), while sedimentation or other self-sinking mechanisms dominate the spatial distribution pattern during the dry season (November to February). The concentration of chlorophyll-*a* is high in the upstream area, but it decreases downstream; this condition is related to the occurrence of dilution.

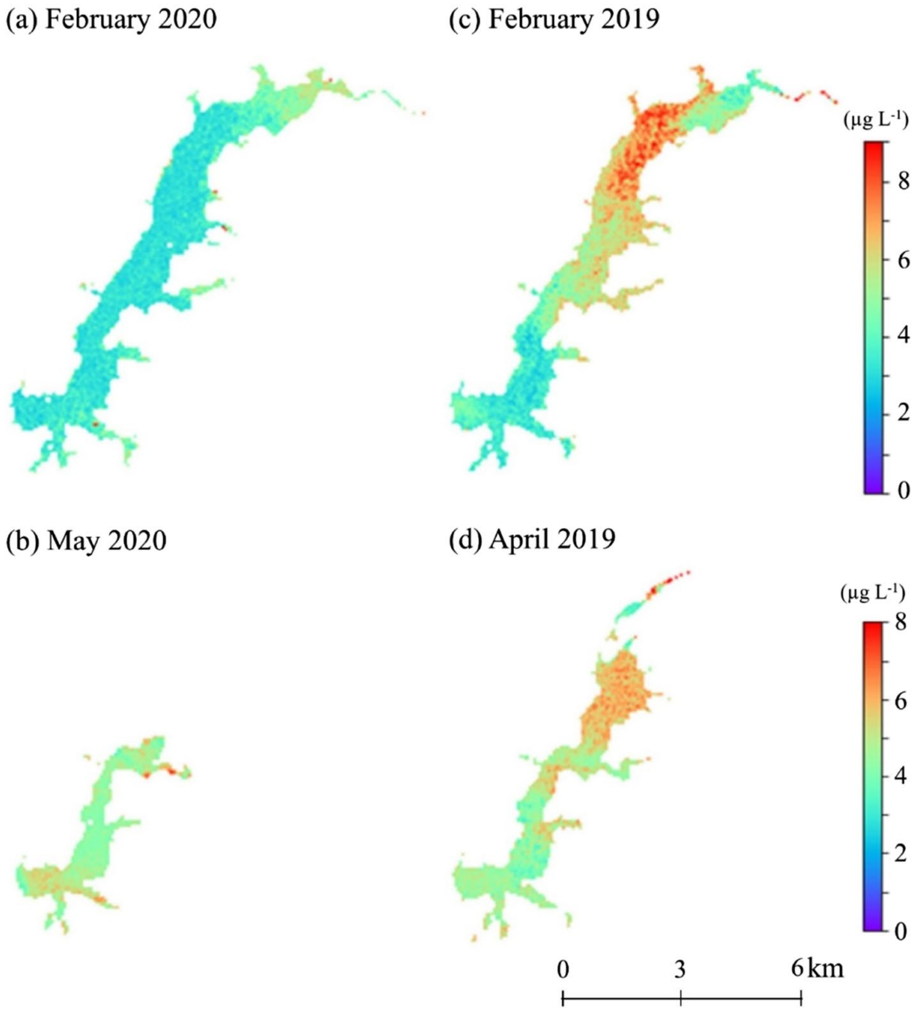

Fig. 5 Spatial map of estimated chlorophyll-*a* applied to Sentinel-2 Level-2A in the Tsengwen Reservoir generated by Model A, during the wet season (a,c) and dry season (b,d)

4.2 Chlorophyll-*a* estimation from OLCI

Considering the rapid revisit time and multiple match-ups of Sentinel-3 OLCI with MSI, it is considered an MSI surrogate for rapid chlorophyll-*a* retrieval assessment. The [BR5, BR2] random forest exhibits robust performance on Sentinel-3 OLCI, wherein the model performs satisfactorily in chlorophyll-*a* retrieval. Random forest presents the best fit in machine learning and quantified on the basis of the mean square phase. Figure 6 depicts the further validation of the estimated chlorophyll-*a* with 20% observed chlorophyll-*a* data. The model acquires RMSE = 1.19 μg L^{-1} and $R^2 = 0.822$ (Table 2). In terms of mapping purpose, the poor spatial resolution of OLCI has limited its application to smaller water bodies and estuaries.

Figure 7 presents the estimation map of the Tsengwen Reservoir generated using Model B OLCI at February 2019 (RMSE = 1.32 μg L^{-1}), December 2019 (RMSE = 1.53 μg L^{-1}), and February 2020 (RMSE = 1.37 μg L^{-1}). Sentinel-3 OLCI has been a practical instrument for estimating chlorophyll-*a* and indicating and monitoring the spatiotemporal dynamic process of water quality. However, the spatial resolution of OLCI is insufficient for inland water quality usage, especially regional reservoirs. The 300 m spatial resolution of Sentinel-3 OLCI is slightly coarse. Thus, describing the details of the spatial distribution pattern is difficult in the Tsengwen Reservoir.

4.3 Data fusion from chlorophyll-*a* estimations

Random forest in the fusion model produces a fused map since OLCI and the nearly synchronous MSI images are selected. Figure 8 presents the estimations through the closest time of the satellite constellation MSI (February 16, 2019), OLCI (February 15, 2019), and fine-resolution fusion map on February 18, 2019, obtaining an estimated RMSE of 1.21, 1.32, and 1.35 μg L^{-1}, respectively. Result

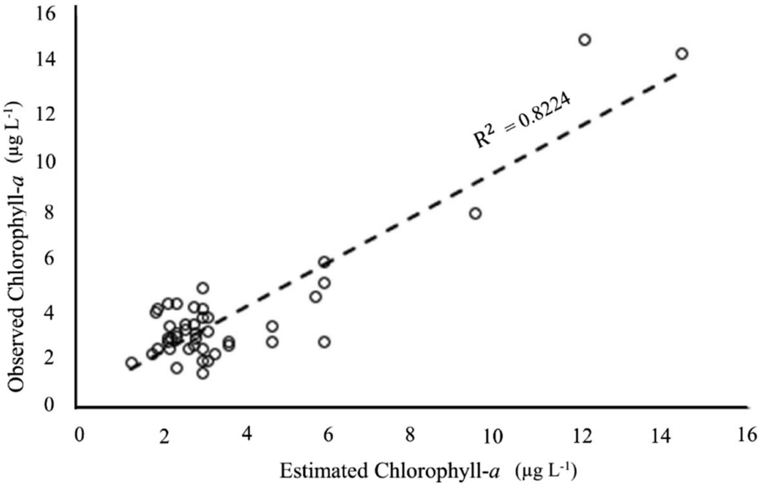

Fig. 6 Validation of Model B by using the observed and estimated chlorophyll-*a* by using 20% of the dataset

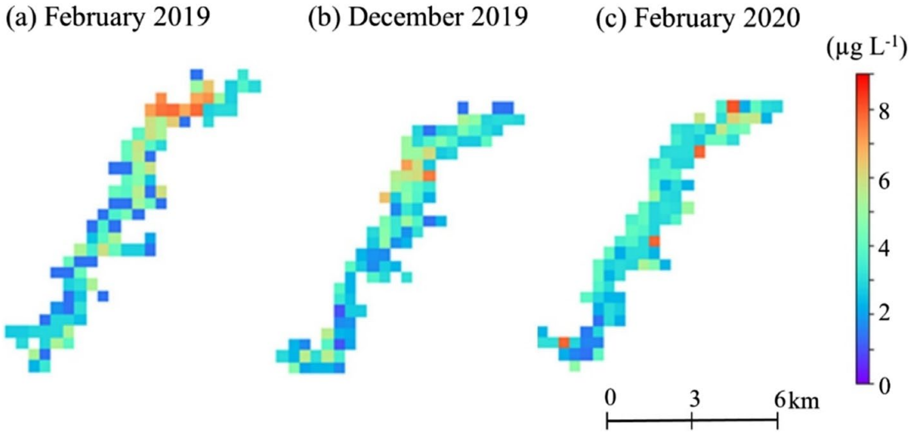

Fig. 7 Estimation maps of the Tsengwen Reservoir by using Model B from Sentinel-3 OLCI

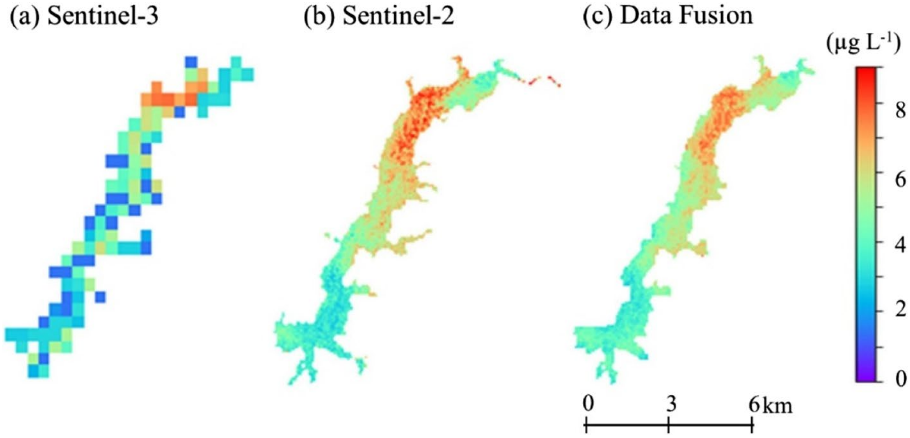

Fig. 8 Estimated chlorophyll-*a* by using Sentinel-3 OLCI, Sentinel-2 MSI, and fine-resolution estimation map from data fusion in Tsengwen Reservoir on February 2019

implies that the fusion model can refine the low-resolution concentration from OCLI to the fine-resolution estimated one. These maps consistently show high chlorophyll-*a* concentration in the upstream and midstream (Fig. 8).

Figure 9 presents the data-fusion estimated maps during two months. The estimation RMSE of 1.37, 1.47, and 1.25 μg L^{-1} are obtained using the integrated estimation of the control points on December 5, 2019, January 7, 2020, and February 4, 2020, respectively. Spatial maps exhibit

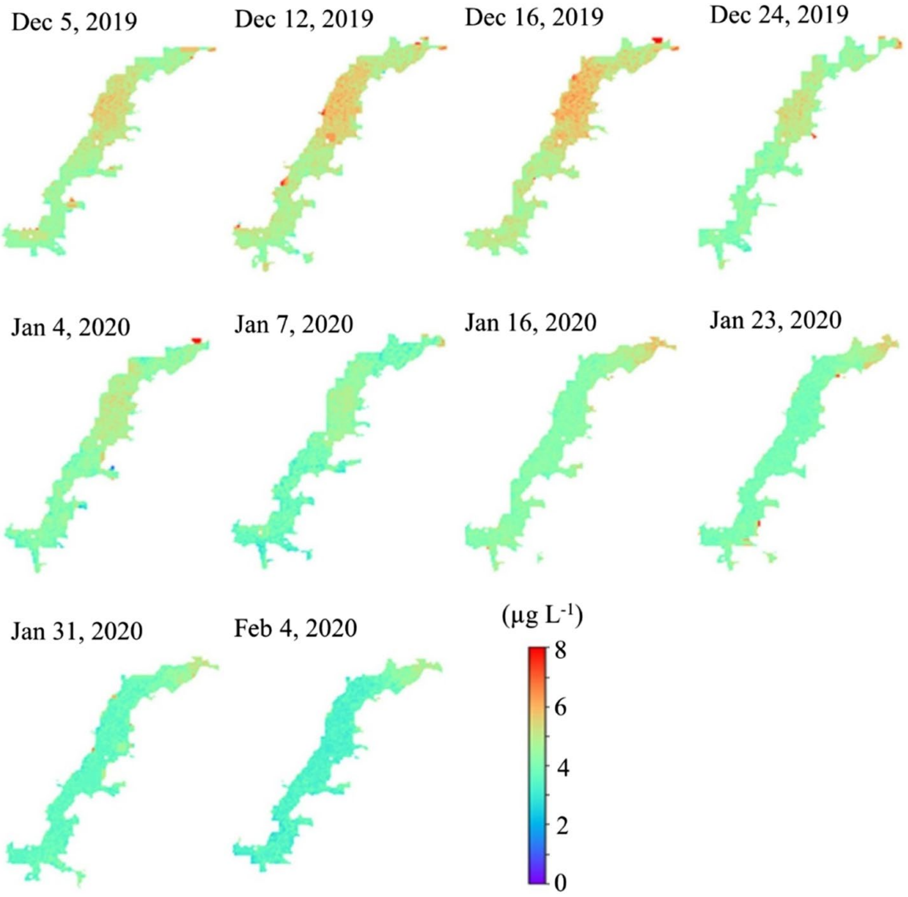

Fig. 9 Fine spatiotemporal resolution chlorophyll-*a* concentration in the Tsengwen Reservoir from December 2019 to February 2020 (field observation dates: December 5, 2019, January 7, 2020, and February 4, 2020)

similar continuous concentration patterns in the midstream area from December 5, 2019 to January 4, 2020. The higher concentration in the upstream occurred from January 16 to 31, 2020. The patterns show that the chlorophyll-*a* hotspot varies with time. The high spatiotemporal resolution of chlorophyll-*a* is estimated within the reservoir when considering spatiotemporal fusion. Figure 10 presents the temporal variability of the predicted chlorophyll-*a* in six monitoring stations of the Tsengwen Reservoir within two months. The field monitoring is insufficient because monthly chlorophyll-*a* observation data from EPA's monitoring station are available on December 5, 2019, January 7, 2020, and February 4, 2020. The concentration of predicted chlorophyll-*a* decreased from December 5, 2019 to February 4, 2020, and aligned with the observed chlorophyll-*a*. The utilization of our model proves to be a good option for producing spatiotemporal variation of chlorophyll-*a* from data fusion. The model is applied to time-series OLCI images for estimating dense-temporal chlorophyll-*a* concentration and analyzing the varied pattern. The spatiotemporal estimated chlorophyll-*a* between the observation dates can be derived from Sentinel-3 OLCI through the fusion model. The spatial distribution of fused chlorophyll-*a* varies with time. From the data fusion with few field observations, we can clearly understand the changes of the chlorophyll-*a* in time and space.

5 Discussion

5.1 Machine learning for water quality monitoring

Although the complexity and nonlinearity of the relationship between water properties and factor parameters require an advanced analysis [19, 38, 39], the machine-learning based models are implemented well [6, 18], especially image-based estimation [36] However, few studies considered multiple satellite image datasets to provide high spatio-temporal resolution water quality mapping products. Based on the machine learning approaches, this study contains great potentials for timely environmental monitoring and assessment using water quality inversion and fusion. Chlorophyll-*a* estimation with two spatiotemporal resolutions (model A and B), and spatiotemporal fusion of chlorophyll-*a* (model C) are developed.

Clear water usually adopts the blue and green spectral regions, while turbid water adopts the red–NIR spectral bands [12, 34]. In Taiwan, the states of water quality in reservoirs are heterogenous. Considering the multiple band ratios e.g. blue–green and red-NIR band ratios are the best selections for chlorophyll-*a* estimation in this study. Consequently, blue reflectance is not a reliable predictor because of the overlapping absorption across its spectral region [37]. In turbid inland waters, NIR and red-edge ratio can evaluate chlorophyll-*a* concentration with significant accuracy. The wavelength peak of chlorophyll-*a* typically depends on two factors: chlorophyll fluorescence and minimum absorption coefficient. Significant chlorophyll-*a* absorption in low reflectance between 400 and 500 nm results in a broad reflectance while a low absorption of algae is produced at approximately 560 nm. Furthermore, the combination of NIR and red bands while considering critical reflectance around 675 nm is proposed [11].

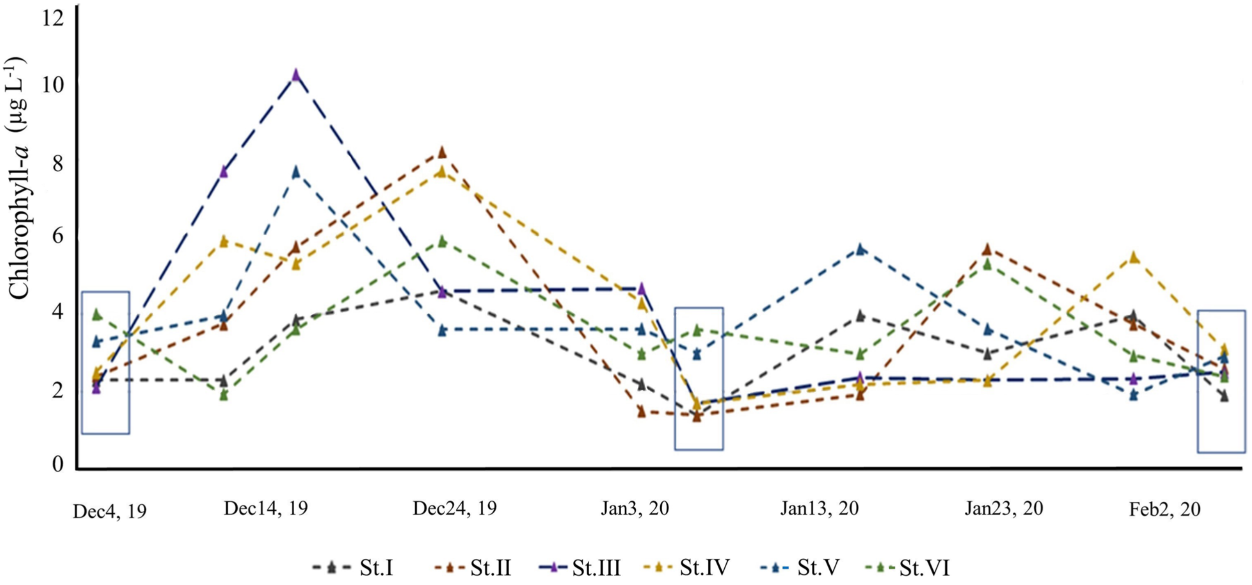

Fig. 10 Chlorophyll-*a* estimated concentration in six monitoring stations of the Tsengwen Reservoir from December 2019 to February 2020 (squared: monthly field observation)

Basing on the range of chlorophyll-*a* in the study lakes, most reservoirs in Taiwan are marked as mesotrophic waters such as Tsengwen Reservoir. In addition, not all algae are rich in chlorophyll-*a*, such as diatoms, brown algae, etc. The machine learning model will be feasible for other water quality parameters since in-situ water quality monitoring can be considered more parameters. The model with the effective band ratios can provide an accurate and appropriate water quality estimation. Machine learning models can estimate chlorophyll-*a* concentrations well within the range of the training data [18]. In this demo case, chlorophyll-*a* concentrations were stable in the field monitoring periods, but were varied between the monitoring periods (<6 μg L^{-1} at most time). The model is still available because its variations within the range of the training data. Moreover, some extreme points are not performed well because the only one global estimation model is trained. In the future, multiple local models will be developed after datasets are applied firstly in cluster analysis due to lake-specific. Furthermore, few observations cannot allow us to examine the structure of estimation discrepancy with respect to space and time for detailed validation. The UAV will be applied to collect actual chlorophyll-*a* concentration data in the future.

5.2 Spatiotemporal fusion

Spatiotemporal fusion is originally developed for blending reflectance from Landsat and MODIS data, and is potential for interdisciplinary applications such as land surface temperature [40], air quality mapping [41] and etc. In this study, spatiotemporal fusion is applied to fuse Sentinel-2 with Sentinel-3 images to estimate dense-temporal 10 m spatial resolution water quality parameters. Chlorophyll-*a* estimation from Sentinel-3 OLCI is refined into fine resolution as Sentinel-2 MSI. The fusion model is estimated from images alone without using any measurements at all. The difficulty of OLCI in capturing fine spatial-resolution variations of chlorophyll-*a* at a small reservoir is presented as a result of its sensor technical limitation. In accordance with temporal resolution, Sentinel-2 MSI cannot achieve high-frequency chlorophyll-*a* monitoring. Therefore, combining with Sentinel-3 OLCI (rapid revisiting time: 2 days) can significantly improve the frequency of water quality monitoring [22].

Here, we consider the fusion of multisource satellite data to obtain water quality information. Chlorophyll-*a* concentration and characteristics, instead of raw sensor data fusion, are used in the fusion process. The data fusion at chlorophyll-*a* information from MSI and OLCI rather than sensor data level can provide enhancement to the chlorophyll-*a* estimation processes and patterns for solving practical problems. Amplifying this approach through utilizing multiple spatiotemporal resolutions of images in data fusion can greatly expand the high spatiotemporal resolutions of chlorophyll-*a* concentration in monitoring and assessment. Moreover, the spatiotemporal fusion can be extended for mapping other water quality parameters if the training dataset is prepared. The fusion model can automatically and accurately identify the temporal variation and spatial distribution of chlorophyll-*a* concentration in near real-time mapping. Satellite imagery is used to determine regional water quality parameters in the aquatic system and plays an immense role in sustainable water resources management. Not only Sentinel images, long-sequence Landsat images are suitable to use for monitoring inland water quality dynamics including inter-annual, seasonal, and abrupt changes. Furthermore, the fusion model depends on spectral consistency of spectral to chlorophyll-*a* transformation. The fusion model may have some limitations if the water components seriously change due to nature or human interference.

6 Conclusions

This study aims to generate a rapid machine learning approach of chlorophyll-*a* estimation in inland water by using band ratio algorithms based on Sentinel-2A MSI and Sentinel-3 OLCI images. The spatiotemporal fusion technique is effective methods to integrate multiple chlorophyll-*a* concentration images for water quality monitoring. Result shows that the most robust model is multiple-band ratio random forest ($R^2=0.873$), i.e., the combination of green–blue (BR 2) and red–NIR (BR 5) ratios from Sentinel-2A MSI. Moreover, the developed model also robustly performs for chlorophyll-*a* retrieval under a Sentinel-3 OLCI environment ($R^2=0.822$). In accordance with the estimation of Tsengwen Reservoir, the model performs well using the MSI model (RMSE = 1.1–1.61 μg L^{-1}), and the OLCI model (RMSE = 1.32–1.53 μg L^{-1}).

The spatiotemporal fusion can estimate dense-temporal 10 m spatial resolution chlorophyll-*a* by utilizing only the time-varying Sentinel-3 OLCI after training under MSI and OLCI environments. The random forest model fused the sparse fine-resolution chlorophyll-*a* images with frequent coarse-resolution chlorophyll-*a* images to create the high spatiotemporal resolution ones. This study validates the potential of high-spatiotemporal-resolution chlorophyll-*a* estimation from the spatiotemporal fusion of the MSI and OLCI (RMSE = 1.25–1.47 μg L^{-1}) in Tsengwen Reservoir. The high spatiotemporal resolution concentration of chlorophyll-*a* will be used as an indicator for quantifying phytoplankton, which contribute to the primary productivity of inland waters. In the future, Carlson trophic state index values for the trophic state of the reservoir will be considered, and correlations or models among chlorophyll-*a*, total phosphorus concentration, and Secchi depth will be identified firstly. This approach will be applied to estimate other water quality parameters such as Secchi depth.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42834-023-00170-1.

Additional file 1:

Acknowledgements

We thank the European Space Agency and Environmental Protection Agency, Taiwan for data support.

Authors'contributions

conceptualization, H.J. Chu; methodology, Wachidatin;validation, Wachidatin; formal analysis, Wachidatin and Tatas; writing—originaldraft preparation, Wachidatin; writing—review and editing, H.J Chu;visualization, Wachidatin; supervision, H.J. Chu and Lalu Muhamad Jaelani. The author(s) read and approved the final manuscript.

Funding

The study was supported by Ministry of Science and Technology, Taiwan (grant number: 110–2621-M-006 -005) and SATU *JRS* projects, NCKU.

Availability of data and materials

The data that support the findings of this study are openly available in https://ladsweb.modaps.eosdis.nasa.gov/; https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR; https://wq.epa.gov.tw/EWQP/zh/ConService/DownLoad/HistoryData.aspx

Declarations

Competing interests

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

Received: 31 October 2022 Accepted: 20 February 2023
Published online: 17 March 2023

References

1. Paerl HW, Otten TG. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol. 2013;65:995–1010.
2. Hunter PD, Tyler AN, Gilvear DJ, Willby NJ. Using remote sensing to aid the assessment of human health risks from blooms of potentially toxic cyanobacteria. Environ Sci Technol. 2009;43:2627–33.
3. Kutser T. Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters. Int J Remote Sens. 2009;30:4401–25.
4. Baban SMJ. Trophic classification and ecosystem checking of lakes using remotely sensed information. Hydrolog Sci J. 1996;41:939–57.
5. Kasprzak P, Padisak J, Koschel R, Krienitz L, Gervais F. Chlorophyll *a* concentration across a trophic gradient of lakes: An estimator of phytoplankton biomass? Limnologica. 2008;38:327–38.
6. Cao ZG, Ma RH, Duan HT, Pahlevan N, Melack J, Shen M, et al. A machine learning approach to estimate chlorophyll-*a* from Landsat-8 measurements in inland lakes. Remote Sens Environ. 2020;248:111974.
7. Ha NTT, Thao NTP, Koike K, Nhuan MT. Selecting the best band ratio to estimate chlorophyll-*a* concentration in a tropical freshwater lake using sentinel 2A images from a case study of Lake Ba Be (Northern Vietnam). Isprs Int J Geo-Inf. 2017;6:290.
8. Gitelson A, Garbuzov G, Szilagyi F, Mittenzwey KH, Karnieli A, Kaiser A. Quantitative remote-sensing methods for real-time monitoring of inland waters quality. Int J Remote Sens. 1993;14:1269–95.
9. Kallio K, Kutser T, Hannonen T, Koponen S, Pulliainen J, Vepsalainen J, Pyhalahti T. Retrieval of water quality from airborne imaging spectrometry of various lake types in different seasons. Sci Total Environ. 2001;268:59–77.
10. Gilerson AA, Gitelson AA, Zhou J, Gurlin D, Moses W, Ioannou I, et al. Algorithms for remote estimation of chlorophyll-*a* in coastal and inland waters using red and near infrared bands. Opt Express. 2010;18:24109–25.
11. Le CF, Hu CM, Cannizzaro J, Duan HT. Long-term distribution patterns of remotely sensed water quality parameters in Chesapeake Bay. Estuar Coast Shelf S. 2013;128:93–103.
12. Gurlin D, Gitelson AA, Moses WJ. Remote estimation of chl-*a* concentration in turbid productive waters – Return to a simple two-band NIR-red model? Remote Sens Environ. 2011;115:3479–90.
13. Yang Z, Anderson Y. Estimating chlorophyll-*a* concentration in a freshwater lake using Landsat 8 imagery. J Environ Earth Sci. 2016;6:134–42.
14. Chu HJ, He YC, Chusnah WN, Jaelani LM, Chang CH. Multi-reservoir water quality mapping from remote sensing using spatial regression. Sustainability. 2021;13:6416.
15. Smith ME, Lain LR, Bernard S. An optimized chlorophyll *a* switching algorithm for MERIS and OLCI in phytoplankton-dominated waters. Remote Sens Environ 2018;215:217–27.
16. Hu C, Feng L, Guan Q. A Machine learning approach to estimate surface chlorophyll *a* concentrations in global oceans from satellite measurements. IEEE T Geosci Remote. 2021;59:4590–607.
17. Kwon YS, Baek SH, Lim YK, Pyo JC, Ligaray M, Park Y, et al. Monitoring coastal chlorophyll-*a* concentrations in coastal areas using machine learning models. Water. 2018;10:1020.
18. Sagan V, Peterson KT, Maimaitijiang M, Sidike P, Sloan J, Greeling BA, et al. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth-Sci Rev. 2020;205:103187.
19. Chen JY, Chen SS, Fu R, Li D, Jiang H, Wang CY, Peng YS, Jia K, Hicks BJ. Remote sensing big data for water environment monitoring: current status, challenges, and future prospects. Earth's Future. 2022;10:e2021EF002289.
20. Toming K, Kutser T, Laas A, Sepp M, Paavel B, Noges T. First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery. Remote Sens-Basel. 2016;8:640.
21. Gernez P, Doxaran D, Barille L. Shellfish aquaculture from space: potential of Sentinel2 to monitor tide-driven changes in turbidity, chlorophyll concentration and oyster physiological response at the scale of an oyster farm. Front Mar Sci. 2017;4:137.
22. Cazzaniga I, Bresciani M, Colombo R, Della Bella V, Padula R, Giardino C. A comparison of Sentinel-3-OLCI and Sentinel-2-MSI-derived chlorophyll-*a* maps for two large Italian lakes. Remote Sens Lett. 2019;10:978–87.
23. Bramich J, Bolch CJS, Fischer A. Improved red-edge chlorophyll-*a* detection for Sentinel 2. Ecol Indic. 2021;120:106876.
24. Gevaert CM, Garcia-Haro FJ. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sens Environ. 2015;156:34–44.
25. Dona C, Chang NB, Caselles V, Sanchez JM, Camacho A, Delegido J, et al. Integrated satellite data fusion and mining for monitoring lake water quality status of the Albufera de Valencia in Spain. J Environ Manage. 2015;151:416–26.
26. Kremezi M, Karathanassi V. Data fusion for increasing monitoring capabilities of Sentinel optical data in marine environment. IEEE J-Stars. 2020;13:4809–15.
27. Putri MSA, Lin JL, Chiang Hsieh LH, Zafirah Y, Andhikaputra G, Wang YC. Influencing factors analysis of Taiwan eutrophicated reservoirs. Water. 2020;12:1325.
28. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017;202:18–27.
29. Su H, Lu XM, Chen ZQ, Zhang HS, Lu WF, Wu WT. Estimating coastal chlorophyll-*a* concentration from time-series OLCI data based on machine learning. Remote Sens-Basel. 2021;13:576.

30. De Keukelaere L, Sterckx S, Adriaensen S, Knaeps E, Reusen I, Giardino C, et al. Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. Eur J Remote Sens. 2018;51:525–42.
31. Nurgiantoro, Muliddin, Kurniadin N, Putra AYSI, Azharuddin M, Hasan J, et al. Assessment of atmospheric correction results by iCOR for MSI and OLI data on TSS concentration. IOP Conf Ser: Earth Environ Sci. 2019;389:012001.
32. Wang L, Xu M, Liu Y, Liu HX, Beck R, Reif M, et al. Mapping freshwater chlorophyll-*a* concentrations at a regional scale integrating multi-sensor satellite observations with Google Earth Engine. Remote Sens-Basel. 2020;12:3278.
33. Boucher J, Weathers KC, Norouzi H, Steele B. Assessing the effectiveness of Landsat 8 chlorophyll *a* retrieval algorithms for regional freshwater monitoring. Ecol Appl. 2018;28:1044–54.
34. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comp Sci. 2003;43:1947–58.
35. Wady SMA, Bentoutou Y, Bengermikh A, Bounoua A, Taleb N. A new IHS and wavelet based pansharpening algorithm for high spatial resolution satellite imagery. Adv Space Res. 2020;66:1507–21.
36. Chusnah WN, Chu HJ. Estimating chlorophyll-*a* concentrations in tropical reservoirs from band-ratio machine learning models. Remote Sens Appl. 2022;25:100678.
37. Yacobi YZ, Moses WJ, Kaganovsky S, Sulimani B, Leavitt BC, Gitelson AA. NIR-red reflectance-based algorithms for chlorophyll-*a* estimation in mesotrophic inland and coastal waters: Lake Kinneret case study. Water Res. 2011;45:2428–36.
38. Chen GQ, Long TY, Xiong JG, Bai Y. Multiple random forests modelling for urban water consumption forecasting. Water Resour Manag. 2017;31:4715–29.
39. Shin Y, Kim T, Hong S, Lee S, Lee E, Hong S, et al. Prediction of chlorophyll-*a* concentrations in the Nakdong river using machine learning methods. Water. 2020;12:1822.
40. Huang B, Wang J, Song HH, Fu DJ, Wong K. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. IEEE Geosci Remote S. 2013;10:1011–5.
41. Wu JA, Li TW, Zhang CY, Cheng Q, Shen HF. Hourly $PM_{2.5}$ concentration monitoring with spatiotemporal continuity by the fusion of satellite and station observations. IEEE J-Stars. 2021;14:8019–32.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.