Detection of Suspended-Matter Concentrations in the Shallow Subtropical Lake Taihu, China, Using the SVR Model Based on DSFs

De Yong Sun, Yun Mei Li, Qiao Wang, Heng Lv, Cheng Feng Le, Chang Chun Huang, and Shao Qi Gong

Abstract—Accurate detection of suspended-matter concentrations in water columns is an important task in remotely sensing water color. This letter aims to identify an optimal model for estimating suspended-matter concentration in the optically complex Lake Taihu of China. Remote sensing reflectance \( R_{rs}(\lambda) \), inherent optical properties, and constituent concentrations of the Lake water were synchronously measured in November of 2007. After the effects of water constituents on \( R_{rs}(\lambda) \) were analyzed, the definitive spectral factors were determined, which are indicative primarily of total suspended matter (TSM). Several methods were compared in modeling the relationship between \( R_{rs}(\lambda) \) and TSM. Results show that the support vector regression (SVR) model performs best with a root-mean-square error of 4.7 mg \( \cdot \) L\(^{-1} \) (\( R^2 = 0.968 \)). Its predictive errors in four seasons were also assessed with the mean absolute percentage errors varying in the range of 22.0%–60.0%. Thus, the SVR model can be used to reliably retrieve TSM concentrations in Lake Taihu. This finding offers new insights into the optical signals of in-water constituents in optically complex lakes.

Index Terms—Definitive spectral factors (DSFs), Lake Taihu, support vector regression (SVR), suspended matter.

I. INTRODUCTION

In Case I waters, suspended particulates occur mainly in the form of organic phytoplankton and corresponding detritus [1], and chlorophyll can be easily detected by optical remote sensing images [2]. Nevertheless, this does not always hold true for Case II waters, where terrestrial particulates and resuspended sediments coexist with phytoplankton and colored dissolved organic matter (CDOM) [3]. The suspended particles markedly attenuate light in the water column and interfere with remote chlorophyll estimation [4]. Thus, detection of suspended particulates is crucial to improving estimates of chlorophyll and understanding the variability of water ecological structure and water transparency. Meanwhile, suspended particulates also play an important role in water-quality management since they are related to total primary production and fluxes of heavy metals and micropollutants [5].

Previous studies have made use of visible wavelength ratio algorithms to infer the distribution of suspended matters [6], [7]. These algorithms work well in coastal and estuarine waters whose optical properties are likely dominated by mineral sediments. However, these site-specific algorithms vary widely in their complexity. Their performance is also severely compromised in the presence of other optically active constituents. A novel reflectance ratio model developed by Doxaran et al. [8], which utilizes near-infrared and visible wavelengths, proved successful for waters having an extremely high (e.g., up to 2000 mg \( \cdot \) L\(^{-1} \)) concentration of suspended particulates. At near-infrared wavelengths, the reflectance changes markedly, which differs widely from that of inland productive turbid lakes, where pure water absorption and particulate scattering may exert a similar influence on near-infrared reflectance. A single-band visible reflectance (665 nm) model has been developed to estimate suspended-sediment concentrations in the moderately turbid waters of the Irish Sea [9]. The success of this model, however, is dependent on the absence of interference from phytoplankton absorption at 665 nm in light of relatively low phytoplankton pigment concentrations.

This letter explores the suitability of a new model based on definitive spectral factors (DSFs) in retrieving suspended-matter concentrations in turbid Lake Taihu waters. Meanwhile, the support vector regression (SVR) method is utilized to develop the inverse model, which is regarded as a promising alternative in Case II waters because of its many advantages, such as fewer parameters to be determined, a unique and global minimum solution, and high-generalization ability. This letter also pioneers the retrieval of suspended particulates in a highly turbid lake using the SVR method.

II. MATERIALS AND METHODS

A. Water Sampling

Lake Taihu is one of the largest inland freshwater lakes in China. It has an area of 2338.1 km\(^2\) and an average depth of 1.9 m [10]. Surface water samples were collected in November 8–21, 2007, using Niskin bottles at 74 stations distributed regularly over the entire Lake [3]. They were immediately preserved under low temperature (2 °C to 4 °C). These
water samples were analyzed within the same day in a laboratory. While sampling, climatic conditions (i.e., wind speed, wind direction, and water temperature) were also recorded.

B. Measurement of All Parameters

Water reflectance spectra were measured with a portable ASD FieldSpec spectroradiometer, which we can refer to Sun et al. [3]. Suspended-particle concentrations were measured according to the investigation criteria for lakes in China [11], and the detailed description also can be seen in Sun et al. [12]. Measurement of phytoplankton pigment absorption coefficient \(a_{ph}(\lambda)\), nonalgal particle absorption \(a_{NAP}(\lambda)\), and CDOM absorption \(a_{CDOM}(\lambda)\) were carried out using the quantitative filter technique [13]. The particulate scattering coefficient \(b_p\) was measured with a WETLabs AC-S [14], and the particulate backscattering coefficient \(b_{bpg}\) was measured using a WETLabs ECO-BB9 [15]. This was also described in detail by Sun et al. [12].

C. Data Selection and Mathematical Methods

Some of the reflectance spectra were abnormal. A few were very similar to that of vegetation, an outcome resulting from algal blooms. Other abnormal spectra were attributed to the presence of submerged plants at the lake floor in shallower parts. These abnormal spectra were discarded to ensure the highest quality of the data set, leaving 47 groups of data useful. These samples were randomly divided into two parts: a training data set of 32 samples, and a validation data set of 15 samples. The mean absolute percentage error (MAPE) and root-mean-square error (rmse) were calculated to indicate the accuracy of retrieval models. These accuracy criteria are calculated as

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - y'_i}{y_i} \right| \quad (1)
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - y'_i)^2} \quad (2)
\]

where \(n\) is the number of samples, \(y_i\) is the measured value, and \(y'_i\) is the estimated value.

The neural network (NN) and SVR methods were described simply here. The NN method was originally developed to model the functioning of the human brain. In theory, it has been proven that any function, no matter how complex, can be represented by an NN, which is known as the Kolmogorov representation theorem [16]. However, NN models may suffer an important problem, i.e., the NN model is based on the empirical risk minimization principle, which can easily lead to overfitting of the training data set and thereby, poor generalization. Support vector machine is a universal learning method proposed by Vapnik [17], which has been gradually applied to water color of remote sensing field [3], [18]. The basic concept is to map the generic input pattern \(x\) into a high-dimensional feature space via nonlinear mapping and to solve the linear regression problem in this feature space. The SVR function is implemented on the basis of specific kernel functions with corresponding parameters. The radial basis function kernel is selected because of its less numerical difficulties than other kernels [19]. Meanwhile, the \(\varepsilon\)-SVR algorithm was chosen for its significant advantage over \(\varepsilon\)-SVR in automatically adjusting the width of the \(\varepsilon\)-tube around the function being approximated, which was implemented based on a MATLAB software.

III. Results

A. Variability of Suspended-Particle Concentrations and Optical Parameters

Suspended-particle concentrations and optical parameters show a great variability in Lake Taihu. Total suspended matter (TSM) concentration ranges widely from 8.5 to 116.6 mg \(\cdot\) \(1^{-1}\) with a mean of 32.7 mg \(\cdot\) \(1^{-1}\) and a standard deviation of 24.0 mg \(\cdot\) \(1^{-1}\). The organic suspended matter (OSM) has a mean concentration of 8.8 mg \(\cdot\) \(1^{-1}\). The mean value is 23.9 mg \(\cdot\) \(1^{-1}\) for inorganic suspended matter (ISM) concentration. Thus, ISM concentration is higher than OSM concentration in Lake Taihu waters as a whole. Moreover, TSM concentration is correlated closely with ISM concentration in the form of ISM = 0.9484 TSM \(−\) 7.189 (\(R^2 = 0.9881, p < 0.001\)). In contrast, TSM is not correlated with OSM, with a regression relationship of OSM = 0.0516 TSM \(+\) 7.189 (\(R^2 = 0.1968, p < 0.002\)). This indicates that ISM dominates the suspended-particle concentration variability. If the TSM concentration is retrieved successfully, ISM and OSM concentrations can also be estimated.

Nonalgal particulates have a mean absorption coefficient of 1.75 m\(^{-1}\) (standard deviation, SD = 1.27 m\(^{-1}\)) at 440 nm, the largest among all water components. This value is much higher than the mean values of phytoplankton pigments and CDOM at 0.92 and 0.47 m\(^{-1}\), respectively. The particulate-backscattering coefficient ranges from 0.09 to 0.25 m\(^{-1}\) with a mean value of 0.21 m\(^{-1}\) (SD = 0.03 m\(^{-1}\)) at 532 nm. Particulate scattering has a large range of variation at 4.84\(−\)74.04 m\(^{-1}\) with the mean being 25.75 m\(^{-1}\). The maximal particulate-scattering coefficient of 74.04 m\(^{-1}\) is expected, given that the Lake is laden with suspended particles. This value, however, is lower than the \(b_p(555)\) value of about 100 m\(^{-1}\) reported by Doxaran et al. [20]. The backscattering ratio \((b_{bpg}/b_p)\) can denote precisely the nature of particulate assemblage [21], which ranged from 0.003 to 0.026 with a mean value of 0.012 (SD = 0.006). The \(R_{rs}(\lambda)\) spectra show a large variability due to highly variable water compositions [3]. These spectra are also very similar to those found in other turbid waters [22]. The reflectance has obvious peaks in the vicinity of 570, 700, and 815 nm. There is a reflectance trough in the vicinity of 675 nm caused by chlorophyll \(a\) absorption. These spectral characteristics form the foundation for retrieving in-water constituents.

B. DSFs for TSM Retrieval

Shown in Fig. 1 is the fractional contribution of pure water and nonwater substances to total absorption over 400\(−\)900 nm. While nonwater materials dominate the total absorption in short wavelengths, pure water accounts for most of the total absorption in long wavelength bands. At 680 nm, the mean absorption contribution ratio of pure water is approximately
the same as that of nonwater materials, both being 50%, with a standard deviation of about 10% (Fig. 1). This contribution ratio increases gradually from 680 nm to longer wavelengths. At 720 nm, the mean contribution ratio reaches about 95% with a standard deviation of 2.6%. Beyond 750 nm, absorption of nonwater materials is usually assumed to be zero, hence, the total absorption is attributed entirely to pure water. The effect of nonwater materials on water absorption after 720 nm can be safely ignored.

In order to determine backscattering, the mean contribution ratio of pure water to total backscattering was calculated from the collected samples, and the curves at the two sides of the middle curve represent contribution ratios of the mean ± SD, respectively.

Based on the previous analysis, the widely used model proposed by Gordon et al. [23] [3] for retrieving in-water constituents can be approximated as (4) for spectral wavelengths beyond 720 nm

\[
R_{rs}(\lambda) \propto -\frac{b_{bt}(\lambda)}{a_t(\lambda) + b_{bt}(\lambda)}, \quad (\lambda \geq 720 \text{ nm})
\]

where \(b_{bt}\) is the total backscattering coefficient of water column, \(a_t\) is the total absorption coefficient of water column, \(b_{bp}\) is the particulate backscattering coefficient, \(a_w\) is the absorption coefficient of pure water, \(\gamma\) is the corresponding parameter, and \(\lambda\) is the wavelength. Backscattering must be isolated, a key step in retrieving suspended-particulate concentrations from remotely sensed reflectance, because only \(b_{bp}(\lambda)\) in the previous model carries information on suspended particulates.

The \(R_{rs}(\lambda)\) spectra exhibit a reflectance valley near 750 nm and a reflectance peak at around 815 nm, where pure water has a corresponding absorption peak and valley (Fig. 2). As \(R_{rs}(\lambda)\) gradually decreases from 815 nm to longer wavelengths, the absorption coefficient of pure water shows an opposite tendency of rapid increase. These spectral variations in \(R_{rs}(\lambda)\) depend closely on pure-water absorption, suggesting that pure water dominates the shape of \(R_{rs}(\lambda)\) in near-infrared spectral bands. Variation of particulate backscattering has a smooth trend that has not been observed in this spectral range before.

Therefore, spectral wavelengths with minimal pure-water absorption in the near-infrared wavebands (> 720 nm) should be selected in order to suppress the effect of pure water on \(R_{rs}(\lambda)\). The reflectance at six wavelengths of 725, 730, 735, 810, 815, and 820 nm was determined to be the inputs to the DSFs. These wavelengths can also be verified through analysis of the correlation between \(R_{rs}(\lambda)\) and suspended-particulate concentrations. Both TSM and ISM concentrations correlate well with \(R_{rs}(\lambda)\) in the near-infrared spectrum. There are two distinctly sensitive spectral peaks in the vicinity of 730 and 815 nm (not shown in the figure here), with the correlation coefficient between the six \(R_{rs}(\lambda)\)'s and TSM concentration all over 0.86.

C. TSM Retrieval Models

In this study, three original models were developed and compared to identify the best in retrieving TSM concentrations: a linear regression model, a back-propagation NN model, and an SVR model. The NN method has a critical limitation of overfitting in that an NN model trained with one data set may not perform well if data from another site are used. Thus, it is necessary to explore more stable and accurate models in monitoring suspended particulates.

The SVR model accepts hyperspectral data as the input. The previously proposed DSFs were fed to the SVR model, and TSM concentration was treated as the only output. In this way,
TSM estimation was changed into a regression problem of SVR with six inputs and one output. It can be expressed as

\[
X = (R_{rs}(725), R_{rs}(730), R_{rs}(735), R_{rs}(810), R_{rs}(815), R_{rs}(820))
\]

\[
Y = \text{TSM}.
\]

Two key parameters, penalty of estimation error \(C\) and kernel parameter \(\sigma\), largely govern the performance of the SVR model [17]. Thus, the establishment of the SVR model relies critically upon finding the optimal \((C, \sigma)\) pair. They were determined via repeatedly training the model with different combinations of \((C, \sigma)\) values. It was found that the \((10^5, 0.21)\) pair resulted in the highest accuracy, and the corresponding \(R^2\) and rmse were, respectively, 0.968 and 4.7 mg \(\cdot \) \(\text{l}^{-1}\).

Two other models for estimating suspended-particulate concentrations were also tried with the \textit{in situ} measured data set. The first is the Tassan [6] model for the Mediterranean Sea. This model has the following form:

\[
\text{lg(TSM)} = a + b \text{lg} \left[ \left( R_{rs}(555) + R_{rs}(670) \right) \times \left( R_{rs}(490)/R_{rs}(555) \right) \right]
\]

where \(a\), \(b\), and \(c\) are regional empirical coefficients. They were determined to be 3.334, 1.536, and \(-0.316\), respectively, using the 32 training samples. The corresponding \(R^2\) is 0.585, and the rmse is 18.5 mg \(\cdot \) \(\text{l}^{-1}\) at a significance level of \(p < 0.005\). The second is the Binding \textit{et al.} [9] model. Essentially, this model is a quadratic function between spectral reflectance at 665 nm and suspended-particulate concentrations. This model achieved an \(R^2\) value of 0.609 and an rmse value of 16.2 mg \(\cdot \) \(\text{l}^{-1}\) with a corresponding significance level of \(p < 0.001\). Additionally, the \(R_{rs}(850)/R_{rs}(550)\) model proposed by Doxaran \textit{et al.} [8] was also tested, but with a low accuracy \((R^2 = 0.40)\).

All the five models were validated for their performance using the 15 samples. Comparatively, the regression model performs better than the Tassan and Binding models. Nevertheless, this model is less accurate than the SVR and NN models. Although they have a similar MAPE, their rmse differs widely (e.g., 5.7 mg \(\cdot \) \(\text{l}^{-1}\) for the SVR model versus 11.6 mg \(\cdot \) \(\text{l}^{-1}\) for the NN model) (Table I). Meanwhile, although the NN model can produce relatively high accuracy sometimes, its performance depends strongly on samples. In contrast, the performance of the SVR model is quite stable with the highest modeling accuracy and the lowest predictive errors. Therefore, the SVR model is deemed the best at estimating TSM concentrations.

### Table I

<table>
<thead>
<tr>
<th>Model type</th>
<th>MAPE(%)</th>
<th>RMSE (mg l(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tassan model</td>
<td>45.9</td>
<td>22.9</td>
</tr>
<tr>
<td>Binding model</td>
<td>42.3</td>
<td>18.8</td>
</tr>
<tr>
<td>Regression model</td>
<td>29.6</td>
<td>14.9</td>
</tr>
<tr>
<td>NN model</td>
<td>17.1</td>
<td>11.6</td>
</tr>
<tr>
<td>SVR model</td>
<td>18.3</td>
<td>5.7</td>
</tr>
</tbody>
</table>

**D. Model Applicability and Sensitivity Analysis**

The established SVR model is based on the \textit{in situ} data set measured in November 2007. Because inherent optical properties (IOPs) of water components may vary with sampling time, its applicability to and performance in other seasons were explored using data sets collected at different times. Over the years, several data sets have accumulated. They were dated back to August 2006, October 2006, March 2007, and November 2008, covering nearly all four seasons.

Total suspended particulates in Lake Taihu waters varied with seasons. While the largest mean TSM concentration of 60.0 mg \(\cdot \) \(\text{l}^{-1}\) (SD of 18.5 mg \(\cdot \) \(\text{l}^{-1}\)) occurred in summer, the widest variation range was found to be from 6.5 to 143.5 mg \(\cdot \) \(\text{l}^{-1}\) in autumn (Table II). The mean values of the two seasons are both much higher than those of the samples \((32.7 \pm 24.0 \text{ mg} \cdot \text{l}^{-1})\) in this study. The performance of the SVR model is not affected by high TSM values. The MAPEs in the two seasons are 22.0% and 25.5%, and the rmses are 18.8 and 15.1 mg \(\cdot \) \(\text{l}^{-1}\), respectively. This also testifies to the advantage of the SVR model in providing global results in spite of the presence of large variability of suspended-particulate concentrations. In spring, suspended-particulate concentration displays a relatively small variability range of 12.9–mg \(\cdot \) \(\text{l}^{-1}\). The SVR model has a larger predictive error (MAPE, 60.0%) than in other seasons.

In winter, most samples \((N = 40)\) were selected to test the SVR model, and the obtained MAPE and rmse are 38.9% and 12.6 mg \(\cdot \) \(\text{l}^{-1}\), respectively. Although the predictive errors of the SVR model in four seasons are all higher than that by the synchronous validation data set and vary among different seasons, the SVR model as a whole can estimate TSM concentrations with relatively high accuracies in Lake Taihu waters.

The sensitivity of a model reflects its dependence on samples, which should be included in the evaluation of its performance. The developed SVR model requires six input parameters, namely, \(R_{rs}\) at 725, 730, 735, 810, 815, and 820 nm. The sensitivity \((\Delta\text{TSM})\) of the model to each input reflectance was analyzed through the following function:

\[
\Delta\text{TSM} = |\text{SVM}(x + \Delta x) - \text{SVM}(x)|
\]

where \(x\) represents the input parameter and \(\Delta x\) represents its uncertainty \((\Delta x \times 1, \Delta x \times 2, \Delta x \times 3, \Delta x \times 4, \Delta x \times 5, \text{and } \Delta x \times 6\) are the uncertainty of the input reflectance corresponding to six wavelengths). According to the variation ranges of remotely sensed reflectance at the six wavelengths, \(\Delta x\) was derived by dividing the reflectance span by 2, 5, 10, 20, 50, and 100, respectively.
\(\Delta\text{TSM}\) at 735 nm is always the biggest among the six wavelengths irrespective of the uncertainty of \(R_{rs}\). \(\Delta\text{TSM}\) increases from 12.6 to 618.2 mg \(\cdot \text{m}^{-1}\) as \(\Delta \times 2\) changes from 0.0003 sr\(^{-1}\) to 0.0146 sr\(^{-1}\). This indicates that the SVR model is most sensitive to the \(R_{rs}(735)\) variability. Conversely, the model is least sensitive to the \(R_{rs}(815)\) variability because \(\Delta\text{TSM}\) has the minimal change as \(\Delta \times 2\) changes from 1.1 sr\(^{-1}\) to 12.5 sr\(^{-1}\). \(\Delta\text{TSM}\) also rises considerably in tandem with the \(R_{rs}(730)\) and \(R_{rs}(810)\) variability, even though the values are lower than that of \(R_{rs}(735)\). It is concluded that the developed SVR model has varying dependences on input reflectance at the six wavelengths, and its sensitivity has an order of \(R_{rs}(735) > R_{rs}(730) > R_{rs}(810) > R_{rs}(725) > R_{rs}(820) > R_{rs}(815)\).

IV. CONCLUSION

Several approaches, including linear regression, NN training, and SVR, have been used to model the relationship between remotely sensed reflectance and total suspended-particle concentrations. The results show that the developed SVR model has the highest accuracy and most stable performance. Based on DSFs, this model for remotely estimating TSM concentrations in turbid Lake Taihu waters has an \(R^2\) value of 0.968 and an rmse of 4.7 mg \(\cdot \text{m}^{-1}\). Overall, the model also has relatively low predictive errors, even though they vary with seasons. For instance, the MAPE has a range of 22.0\%–60.0\% in different seasons, while the rmse ranges from 12.6 to 18.8 mg \(\cdot \text{m}^{-1}\). Thus, it can serve as a reliable technical framework for retrieving TSM concentrations in the study area. This high accuracy level also demonstrates that the DSFs have been properly determined and are an excellent mathematical method. The close statistical relationship between ISM and TSM in Lake Taihu waters signifies that TSM concentrations derived from the SVR model can further deduce ISM and OSM concentrations. It is concluded that the explored SVR model is the best in accurately monitoring suspended particulate concentration distribution of Lake Taihu waters.

ACKNOWLEDGMENT

The authors would like to thank K. Shi, Y. Yang, R. Xia, X. Jin, Y. Wang, B. Yin, H. Zhang, Y. Xu, Z. Liu, and X. Xu for their participation in the field experiment, also the three anonymous reviewers for their useful comments and suggestions.

REFERENCES


