Developing New Colored Dissolved Organic Matter Retrieval Algorithms Based on Sparse Learning

Ruihao Zhang, Ruru Deng, Yingfei Liu, Yeheng Liang, Longhai Xiong, Bin Cao, and Wenzhi Zhang

***Abstract*—Colored dissolved organic matter (CDOM) is an important biochemical state indicator of aquatic ecosystems. However, its retrieval from remote sensing datasets remains a challenging task due to its high spatiotemporal variability and interference from other water constituents. In this article, we aim to develop new CDOM inversion algorithms by taking advantage of a representative sparse learning algorithm known as least absolute shrinkage and selection operator (LASSO), which is applied to identify the optimal band arithmetic terms for the CDOM inversion and estimate the model parameters in a more global and robust manner than can statistics-based methods. Moreover, a two-stage inversion framework is presented to further enhance the stability of LASSO in addressing inadequate *in situ* sample circumstances. Within the framework, two different schemes are proposed to handle the band arithmetic terms information propagation between the two stages, for which two new algorithms are proposed. Experimental results obtained from the *in situ* bio-optical dataset and the synthesized dataset under various training sample sizes indicate that both of the proposed algorithms can deliver performance superior to six state-of-the-art CDOM inversion algorithms, and with less sensitivity to the training sample states. In addition, the results offered by the new algorithms also enjoy biochemical interpretability, revealing that the red/blue ratio terms are well suited for inverting the CDOM content on these datasets.**

***Index Terms*—Colored dissolved organic matter (CDOM), remote sensing, sparse learning, water quality inversion.**

I. Introduction

COLORED dissolved organic matter (CDOM), also termed yellow substance or gelbstoff, is an important optically active constituent of natural water that plays a critical role in controlling the radiative transfer process within the water body [1], [2]. The primary impact that CDOM has on the underwater light field is its strong absorption for shortwave radiation ranging from ultraviolet to visible light; meanwhile, its scattering is usually assumed to be negligible [3], [4]. As a result, CDOM serves as a reliable barrier for shielding the aquatic ecosystem from harmful ultraviolet radiation [5], [6]. On the other hand, it has been shown that there exists a significant numerical relationship between the content of CDOM and dissolved organic carbon (DOC); as a result, CDOM is frequently considered a practical optical proxy for inverting the DOC distribution [7], [8]. In this context, it is of ecological sense and practical value to develop effective methods to monitor dynamic changes in the CDOM content [9].

Since the sources and constituents of CDOM are highly complex and may vary spatially and temporally [10]–[12], the CDOM content is commonly characterized by absorption coefficients rather than the traditional concentrations. For simplicity, the absorption coefficients of CDOM at various wavelengths are often assumed to follow an exponentially decreasing function [13] such that they can be described by far fewer parameters. The function contains two critical parameters, the absorption coefficient at a specific wavelength λ_0 (usually set as 412 nm or 443 nm), also called the magnitude parameter, and the spectral slope parameter that describes the decay rate of CDOM absorption with increasing wavelength. In many CDOM inversion studies, the spectral slope parameter is assumed to be a constant during the inversion process [14]–[16]; then, the remaining task of the CDOM retrieval is simplified to estimate the magnitude parameter $\boldsymbol{a}_g(\lambda_0)$. In recent decades, a large number of algorithms have been proposed to invert the CDOM content based on the apparent optical properties (AOPs) of water, e.g., the above-surface remote sensing reflectance or the below-surface remote sensing reflectance. According to whether their derivation procedures rely on the assumption of an underwater radiative transfer process, they can be mainly categorized into analytical algorithms, semianalytical algorithms, and empirical algorithms [17]. Compared to analytical and semianalytical algorithms, empirical inversion methods enjoy the advantages of simplicity and high efficiency, for they aim to straightforwardly determine a linear or nonlinear function that can best reflect the relationships between the AOPs and the magnitude parameters based on the *in situ* samples provided by the users. Once the regression function has been determined, the inversion results can be readily obtained by simply substituting the corresponding

Manuscript received March 16, 2020; revised May 11, 2020; accepted June 8, 2020. Date of publication June 19, 2020; date of current version July 2, 2020. This work was supported in part by the Science and Technology Planning Project of Guangdong Province, China, under Grant 2017B020216001, in part by the Innovation Projects in Water Resource of Guangdong Province, China, under Grant 2016-08, in part by the National Natural Science Foundation of China, under Grant 41071230, in part by the Fundamental Research Funds for the Central Universities, under Grant 19lgpy97, and in part by the Digital River Chiefs Satellite Reote Sensing Monitoring Service Project of Huizhou City, China, under Grant 440000-201903-197019009-0001. *(Corresponding author: Ruru Deng.)*

Ruihao Zhang, Ruru Deng, Yeheng Liang, Longhai Xiong, and Bin Cao are with the School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China, the Guangdong Engineering Research Center of Water Environment Remote Sensing Monitoring, and also with the Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation (e-mail: zhangrh27@mail2.sysu.edu.cn; esdrr@mail.sysu.edu.cn; liangyh28@mail.sysu.edu.cn; xionglh5@mail.sysu.edu.cn; caob6@mail2.sysu.edu.cn).

Yingfei Liu is with the School of Marine Sciences, Sun Yat-sen University, Zhuhai 519000, China, and also with the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519000, China (e-mail: liuyf87@mail.sysu.edu.cn).

Wenzhi Zhang is with the Huizhou branch of Guangdong Hydrology Bureau, Huizhou 516003, China (e-mail: 99045295@qq.com).

Digital Object Identifier 10.1109/JSTARS.2020.3003593

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

AOPs into the function without iterative computation. This characteristic makes them quite efficient and attractive in addressing large datasets, e.g., remote sensing images. Consequently, when *in situ* samples are available, empirical algorithms may be a reasonable choice to invert the CDOM content.

The empirical algorithms can be further divided into nonparametric regression algorithms and parametric regression algorithms according to whether the regression function possesses a predetermined numerical form. Nonparametric algorithms can automatically generate the regression models (including the functional form and parameters) by exploiting the information contained by the training samples. Hence, they are more flexible in characterizing the relationship between the AOPs (hereafter termed as features) and the magnitude parameters (hereafter termed as labels), and have become increasingly popular in CDOM retrieval studies in recent years, such as artificial neural networks (ANNs) [18]–[20], support vector machines (SVMs) [21], and random forests (RFs) [22]. Especially in [22] and [23], the authors systematically compared and analyzed the performance of popular machine learning algorithms for the CDOM inversion. Despite many successful applications, nonparametric inversion algorithms still suffer from certain shortcomings. First and foremost, such algorithms usually require many more training samples to construct a robust inversion model, in contrast to parametric regression algorithms [24], [25]. However, this requirement is difficult to satisfy because collecting and analyzing *in situ* water samples is quite costly and time consuming. Second, the models given by the nonparametric inversion algorithms are commonly complex and inexplicit, and it may be inconvenient to share the inversion model and explain the physical motivation behind it. Parametric inversion algorithms are free from these drawbacks and favor simplicity and interpretability. As a result, in this article, we will concentrate on developing new CDOM inversion algorithms within the scope of parametric inversion.

To develop a parametric inversion algorithm, the first and foremost task is to determine the features used in the regression function. Given the AOPs of the training samples, most algorithms have adopted the band-ratio terms to construct the features. In [26] and [27], the authors proposed using the blue/green ratio terms to develop regression models, while in [28]–[30], the red/blue ratios were adopted. Furthermore, the effectiveness of the red/blue ratio settings was also supported by the experimental results in [31] and [32]. However, as reported in [33], the red/blue ratios may fail in the turbid water scenario in which the particular scattering is significant. In addition, other band ratios, such as the red/green ratio [34] and violet/orange ratio [35], have also been introduced to build the inversion models. In addition to the aforementioned algorithms, some scholars have also attempted to retrieve the CDOM content from the vertical attenuation coefficients (Kd) based on semianalytical relationships [36]. By reviewing the aforementioned studies, we find that it can be difficult to design an optimal band ratio that is well suited for diverse aquatic environments. In this context, it would be better to adjust the features according to the *in situ* samples collected from the research area. To perform this task, many parametric regression algorithms make use of correlation analysis [9], [37] or stepwise regression algorithms [38], [39] to find the best features. However, both such methods have disadvantages. In correlation-based methods, the selection is mainly performed by finding the feature with the highest correlation score (e.g., the absolute value of the Pearson correlation coefficient) between the candidate features and the labels. This method is quite efficient; however, the dependence among the candidate features is not considered. As reported in [40], when addressing optically complex water, it is preferable to use more than one band ratio for mitigating the interference caused by chlorophyll. For the stepwise regression, its forward selection version tends to become easily trapped in local optima when finding the best feature subset [41]. At the same time, its backward elimination version is not capable of addressing the underdetermined regression problem, which restricts its capability in exploiting the rich information within high-dimensional candidate features [42].

To sidestep the aforementioned limitations, in this article, we present two novel parametric inversion algorithms for CDOM inversion based on a representative sparse learning algorithm, called *least absolute shrinkage and selection operator* (LASSO), which can perform feature selection and parameter estimation in a parallel manner. Instead of straightforwardly using the classical LASSO to address the CDOM inversion task, we have made several modifications to further improve its performance, which can be briefly summarized as follows.

1) As LASSO can address the underdetermined regression problem, we are free to explore the best band arithmetic terms from a large number of candidate features. In this case, except for the original reflectance information, all possible band multiplicative terms and band-ratio terms are also considered candidate features.
2) When the number of candidate features becomes extremely high, the solution of LASSO may degenerate because the algorithm will automatically pay greater attention to the feature selection procedure rather than the fitting procedure for maximally reducing the objective function value, thus increasing the risk of underfitting. It is, therefore, desirable to mitigate the interference caused by the large numbers of candidate features to obtain more accurate regression coefficients. At this point, a two-stage inversion framework is introduced to solve this problem.
3) Two different schemes are presented to facilitate the passing of feature importance information within the framework. Moreover, the results given by the correlation-based selection algorithm and the stepwise regression algorithm are also incorporated during the passing procedure to enhance the accuracy and robustness of the algorithms.

The remainder of this article is organized as follows. Section II introduces the main principles of LASSO. In Section III, the detailed implementation of the two-stage inversion framework is provided. Experiments conducted using a simulated dataset and an *in situ* dataset are described in Section IV. Finally, Section V concludes this article with possible future research directions.

II. Basic Algorithm

In this section, the original LASSO algorithm, which acts as the critical algorithm foundation for the following sections, is introduced. Let $\mathbf{y} \in \mathbb{R}^N$ be a column vector that stacks the

magnitudes of the water samples (hereafter termed as labels), i.e., $\mathbf{y} = [\mathbf{a}_g^1(\lambda_0), \mathbf{a}_g^2(\lambda_0), \ldots, \mathbf{a}_g^N(\lambda_0)]^{\mathrm{T}}$, $\mathbf{A} \in \mathbb{R}^{N\times M}$ denotes the feature matrix that collects all the band arithmetic terms (hereafter termed as features), and $\mathbf{x} \in \mathbb{R}^M$ represents the regression coefficient vector. Then, LASSO can be expressed as the following constrained optimization problem [43]:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \qquad \text{subject to} \quad \|\mathbf{x}\|_1 \le k \tag{1}$$

where N is the number of training samples, M is the number of candidate features, k is the sparsity of $\mathbf{x}$, which equals the number of features being selected, $\|\cdot\|_1$ is the ℓ_1-norm operator, and $(\cdot)^{\mathrm{T}}$ denotes the matrix transpose operator. By imposing the ℓ_1-norm constraint on the regression coefficients, the optimal solution of (1) will be a sparse vector, that is, most of the elements in the vector $\mathbf{x}$ will equal zero. As the objective function is consistently constructed by the contribution of each feature during the solution process, its solution enjoys more global properties compared with correlation-based algorithms and stepwise algorithms.

Intuitively, there are two key steps for us in applying the LASSO algorithm: constructing the feature matrix and solving the optimization problem.

A. Feature Matrix Construction

Assuming that $\mathbf{r}_n \in \mathbb{R}^L$ denotes the remote sensing reflectance vector of the nth training sample, then the spectral matrix $\mathbf{A}_0$ of all the training samples can be defined as $\mathbf{A}_0 = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N]^{\mathrm{T}}$. In addition to the original reflectance information, the band multiplicative terms [44] and the band-ratio terms are introduced to augment the feature matrix, which are, respectively, defined as $\mathbf{A}_1$ and $\mathbf{A}_2$.

$$\begin{aligned}\mathbf{A}_1 = [&\mathbf{A}_0(:,1)\odot\mathbf{A}_0(:,1), \mathbf{A}_0(:,1)\odot\mathbf{A}_0(:,2),\\ &\ldots, \mathbf{A}_0(:,L-1)\odot\mathbf{A}_0(:,L), \mathbf{A}_0(:,L)\odot\mathbf{A}_0(:,L)]\end{aligned} \tag{2}$$

$$\begin{aligned}\mathbf{A}_2 = [&\mathbf{A}_0(:,1)\oslash\mathbf{A}_0(:,2), \mathbf{A}_0(:,1)\oslash\mathbf{A}_0(:,3),\\ &\ldots, \mathbf{A}_0(:,L)\oslash\mathbf{A}_0(:,L-2), \mathbf{A}_0(:,L)\oslash\mathbf{A}_0(:,L-1)]\end{aligned} \tag{3}$$

where $\odot$ is the elementwise product operator, $\oslash$ denotes the elementwise division operator, and $\mathbf{A}_0(:,\ell)$ denotes the ℓth column of $\mathbf{A}_0$. Through the aforementioned manipulation, the final feature matrix $\mathbf{A}$ can be constructed by the following matrix expansion:

$$\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2]. \tag{4}$$

Although the single band terms and the multiplicative terms are not as popular or effective as the band-ratio terms among the current studies, we still consider them in this article for two reasons. On the one hand, we should verify their effectiveness in the CDOM inversion and in complementing the band ratio features if possible. On the other hand, even if they do not contribute to improving the accuracy of the CDOM inversion and may even make the retrieval problem trickier, it can still serve as a way to examine the performance of the proposed algorithms in handling such extremely high-dimensional problems. It should be noted that any other form of band arithmetic terms can also be introduced into the feature matrix; hence, it provides great flexibility for exploring the optimal features.

B. ℓ_1-Regularized Optimization

Once the feature matrix has been constructed, the remaining procedure of LASSO is to obtain the regression coefficients $\mathbf{x}$ by solving the constrained optimization problem shown in (1). First, for the sake of reducing the computational complexity, it is better to transform the objective function into an unconstrained problem, which can be achieved by the well-known Lagrangian multiplier strategy as follows:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{5}$$

where λ is the regularization coefficient, which has to be positive. A larger λ will lead to a sparser $\mathbf{x}$, and thus, yield a more concise model. However, it may also increase the risk of underfitting. In particular, when λ is set to zero, the solution of LASSO is identical to the least-squares estimators. Hence, λ should be carefully tuned in practical applications. In this article, we apply a widely used strategy, termed k-fold cross validation [45], to determine the optimal λ.

Given λ, numerous methods have been proposed to address this optimization problem in recent decades. In this article, we adopt a simple yet efficient algorithm, called cyclic coordinate descent (CCD), to find the solution. The main principle of CCD is to update the solution in an elementwise manner within one cyclic period. More specifically, for each the element $\mathbf{x}_m$ in $\mathbf{x}$, its updating rule is given by the following equation [43]:

$$\mathbf{x}_m = \operatorname{sign}\left(\frac{\mathbf{A}_m^{\mathrm{T}}\mathbf{r}_m}{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m}\right)\max\left(\left|\frac{\mathbf{A}_m^{\mathrm{T}}\mathbf{r}_m}{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m}\right| - \lambda, 0\right) \tag{6}$$

where

$$\mathbf{r}_m = \mathbf{y} - \mathbf{A}_{-m}\mathbf{x}_{-m} \tag{7}$$

$$\operatorname{sign}(\mathrm{a}) = \begin{cases} 1, & \text{if } a > 0\\ 0, & \text{if } a = 0\\ -1, & \text{if } a < 0. \end{cases} \tag{8}$$

During the iteration, the regression coefficients corresponding to the features that are considered insignificant will automatically shrink to 0.

III. Proposed Inversion Framework

To further enhance the robustness of LASSO in addressing high-dimensional and underdetermined inversion problems that may be frequently encountered when hyperspectral information is available, in this section, we introduce a two-stage inversion framework. The first stage concentrates more on exploiting the relative importance of the band arithmetic terms, while the second stage aims at refining the solution, i.e., correcting the bias in the solution caused by numerous redundant features, based on the feature prior knowledge acquired in the first stage. A

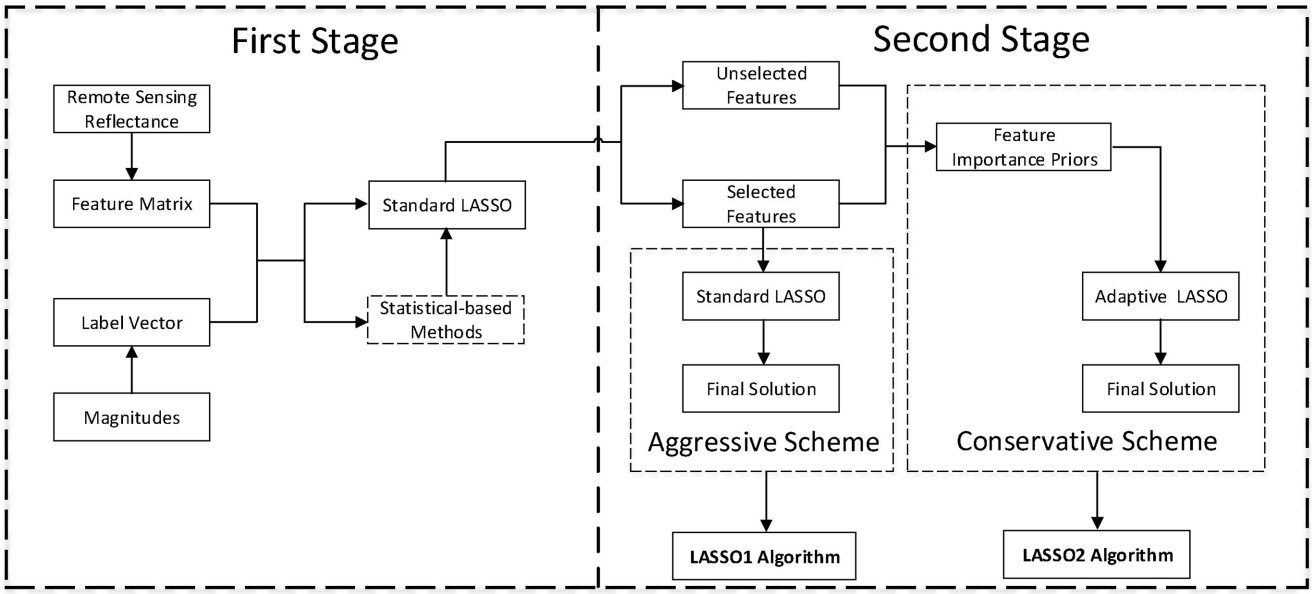


Fig. 1. Flowchart of the two-stage inversion framework.

critical task of implementing this framework refers to the rule of information communication between the two stages. To address this issue, an aggressive scheme and a conservative scheme are developed, respectively. The detailed implementation of this framework is shown by the flowchart in Fig. 1.

A. Aggressive Scheme

The first scheme is aggressive because only the features selected in the first stage will be considered in the refinement procedure, and the nonselected features will be directly discarded. More specifically, in the first stage, a standard LASSO procedure is performed to filter the insensitive features, and thus, yield a much lower dimensional inversion problem. Then, the LASSO algorithm is again applied to the reduced feature space in the second stage. In addition, to further reduce the risk of missing important features, the results delivered by the stepwise regression algorithm and the correlation-based method are also utilized to supplement the LASSO solution obtained in the first stage. Let $\mathbf{I}_0, \mathbf{I}_1,$ and $\mathbf{I}_2$ be a nonempty set recording the indices of the selected feature from the three algorithms; then, a union operation is used to determine the features entering the second stage

$$\mathbf{I} = \mathbf{I}_0 \cup \mathbf{I}_1 \cup \mathbf{I}_2. \tag{9}$$

Defining an indicator vector $\mathbf{u} \in \mathbb{R}^M$ based on the index set $\mathbf{I}$

$$\mathbf{u}_m = \begin{cases} 1, & \text{if } m \in \mathbf{I} \\ 0, & \text{if } m \notin \mathbf{I}. \end{cases} \tag{10}$$

By eliminating the redundant features, the feature matrix can be simplified as follows:

$$\mathbf{A} = \mathbf{A}\mathrm{diag}(\mathbf{u}) \tag{11}$$

where $\mathrm{diag}(\cdot)$ is the diagonal operator. Finally, the simplified feature matrix is used to obtain the final LASSO solution.

B. Conservative Scheme

Compared to the aggressive scheme, the second scheme tends to be more conservative in addressing the LASSO results obtained in the first stage. The nonselected features from the first stage will be combined with the selected features to guide the refinement stage based on their relative importance rather than removing them directly. First, we use the absolute value of the regression coefficients to represent the importance of the features. It should be emphasized that this method only holds when the feature matrix and the label vector have been standardized. Similar to the first scheme, the results achieved by the stepwise regression and correlation-based method are also introduced to assist in determining the LASSO solution. Before integrating the results, a normalization process must first be applied to the results to alleviate scale mismatches. Assuming that their solutions are denoted by $\mathbf{x}, \mathbf{x}_s,$ and $\mathbf{x}_c$, the normalization is implemented as follows:

$$\mathbf{x} = \frac{|\mathbf{x}|}{\max(|\mathbf{x}|)}, \quad \mathbf{x}_s = \frac{|\mathbf{x}_s|}{\max(|\mathbf{x}_s|)}, \quad \mathbf{x}_c = \frac{|\mathbf{x}_c|}{\max(|\mathbf{x}_c|)}. \tag{12}$$

Following normalization, the magnitudes of each solution are scaled into the interval of $[0, 1]$, and then, the normalized solutions are integrated to obtain the final importance vector via a weighted sum manner:

$$\begin{aligned} \mathbf{w} &= g_1\mathbf{x} + g_2\mathbf{x}_s + g_3\mathbf{x}_c \\ &= [\mathbf{x}, \mathbf{x}_s, \mathbf{x}_c]\,\mathbf{G} \end{aligned} \tag{13}$$

where $\mathbf{G} = [g_1, g_2, g_3]^{\mathrm{T}}$ contains the weights of the three solutions, which are defined according to their overall validation

errors in the training procedure. Let $\mathbf{E}$ be the vector that collects the errors of the three algorithms, i.e., $\mathbf{E} = [e, e_s, e_c]^{\mathrm{T}}$; then, the weight vector $\mathbf{G}$ is derived based on the inverse distance weighting rule as follows:

$$\mathbf{G} = \left(\mathbf{1}_{3\times 1} \oslash \mathbf{E}^d\right) / (\mathbf{1}_{1\times 3}(\mathbf{1}_{3\times 1} \oslash \mathbf{E}^d)) \tag{14}$$

where $(\cdot)^d$ is the Hadamard power operator, and d represents the distance order, which controls the sensitivity of the weights in responding to the difference in the errors and is set to $d = 1$ in this article. Evidently, through the aforementioned manipulation, a solution with a smaller cross-validation error will contribute more to the final importance vector. The nonselected features in the first stage have not obtained the weights so far because their corresponding magnitudes in the solution are equal to zero. To ensure that they will be considered in the second stage, we assign them the same nonzero weight $\mathbf{w}_u$. Assuming that the weights of the selected features are $\mathbf{w}_s$, we set the weight for the nonselected features as $\mathbf{w}_u = \min(\mathbf{w}_s)/\beta$. The β is a positive constant, and its value reflects the uncertainties of the solutions. More specifically, a larger β means that the feature selection results in the first stage are reliable such that the importance of the nonselected features should be much less than the selected features. Finally, once the weight vector has been determined, it is used as prior knowledge to guide the refinement procedure using a well-known LASSO variant, called adaptive LASSO, which can be expressed by the following objective function:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2 + \lambda \left\|\mathbf{1}_{M\times 1} \oslash \mathbf{w} \odot \mathbf{x}\right\|_1. \tag{15}$$

The sole difference between (15) and the standard LASSO problem lies in the regularization function, and each element of the regression coefficients is multiplied by a preset constant that is determined by the provided importance vector. When solving this optimization problem, the features with larger weights will have higher probabilities to be selected. Moreover, the introduction of $\mathbf{w}$ also enhances the flexibility of the adaptive LASSO. On the one hand, when $\mathbf{w} = \mathbf{1}$, the adaptive LASSO reduces to the standard LASSO and to ordinary least squares in the case of $\mathbf{w} = \mathbf{0}$. On the other hand, any other band arithmetic terms that have been proposed in past studies could be easily incorporated into this algorithm by imposing the corresponding features with proper weights. In addition, the introduction of the importance vector also endows the standard LASSO algorithm with oracle properties [41].

Similar to LASSO, the adaptive LASSO problem can also be efficiently solved by the CCD algorithm. As the derivation process is almost identical to the LASSO algorithm, we would like to directly present the updating rule for each element of the solution vector

$$\mathbf{x}_m = \operatorname{sign}\left(\frac{\mathbf{A}_m^{\mathrm{T}}\mathbf{r}_m}{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m}\right)\max\left(\left|\frac{\mathbf{A}_m^{\mathrm{T}}\mathbf{r}_m}{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m}\right| - \lambda/\mathbf{w}_m, 0\right). \tag{16}$$

IV. Experiments and Discussions

In this section, CDOM inversion experiments with two public datasets are carried out to examine the performance of the proposed algorithms in MATLAB 2017b, and the experimental results are compared with several classical CDOM inversion algorithms using different performance metrics. In the experiments, the absorption coefficients at 443 nm are used as the labels. All the available samples are randomly split into a training set and a testing set, which are used to train the model and determine the inversion accuracy, respectively. To evaluate the effects of the training sample size in detail, the sample splitting procedure is implemented under various training sample proportions (varying from 0.1 to 0.9 with an interval of 0.1). Moreover, to alleviate the randomness associated with the splitting procedure, for each training sample proportion, we repeat the experimental steps for 40 Monte Carlo runs, and then, the final results are obtained by averaging the performance metrics from these runs. In addition, for the sake of brevity, the algorithms adopting the aggressive scheme and conservative scheme are, respectively, termed the LASSO1 algorithm and LASSO2 algorithm in the following text.

As the proposed algorithms are empirical inversion algorithms, the compared algorithms will also generally be of this type of algorithm. In this article, two categories of empirical algorithms are considered. The first category includes parametric inversion algorithms, including the methods proposed by Mannino *et al.* [27], Tiwari and Shanmugam [30], and Menken *et al.* [46]. As there is no formal appellation for these algorithms, we have to call them by their primary developers' names; this naming rule was also adopted in [32]. The second group of algorithms used for comparison is the popular nonparametric inversion methods, among which the aforementioned SVM algorithm [21], RF algorithm [22], and ANN algorithm [23] are considered. The detailed descriptions of the compared algorithms are given in Table I.

Note that to ensure a fair comparison, all the compared algorithms are also implemented in MATLAB R2017b and refitted using the same training samples in each experiment as the proposed algorithms. In addition, the hyperparameters associated with the compared algorithms are determined following the setting given by the original articles if available; otherwise, the default settings offered by MATLAB are used.

A. Evaluation Metrics

To quantitatively evaluate the results in the experiments, three commonly adopted performance metrics, i.e., the root mean square error (RMSE), the bias, and the linear correlation coefficients (r), are used and are calculated as follows:

1) RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{\left\|\mathbf{y}_{\mathrm{true}} - \mathbf{y}_{\mathrm{est}}\right\|_2^2}{N}}. \tag{17}$$

2) Bias:

$$\mathrm{Bias} = \mathbf{1}_N^{\mathrm{T}}(\mathbf{y}_{\mathrm{true}} - \mathbf{y}_{\mathrm{est}})/N. \tag{18}$$

3) Linear correlation coefficient (r):

$$r = \frac{(\mathbf{y}_{\mathrm{true}} - \overline{\mathbf{y}}_{\mathrm{true}})^{\mathrm{T}}(\mathbf{y}_{\mathrm{est}} - \overline{\mathbf{y}}_{\mathrm{est}})}{\left\|\mathbf{y}_{\mathrm{true}} - \overline{\mathbf{y}}_{\mathrm{true}}\right\|_2 \left\|\mathbf{y}_{\mathrm{est}} - \overline{\mathbf{y}}_{\mathrm{est}}\right\|_2} \tag{19}$$

TABLE I
COMPARED ALGORITHMS

Parametric Regression Methods	Descriptions	Nonparametric Regression Methods	Descriptions
Mannino	$\mathbf{y} = \log((\mathbf{r}(490) \oslash \mathbf{r}(555) + p_1)/p_2)/p_3$	Support Vector Machine	[21]
Tiwari	$\mathbf{y} = p_1 \times \mathbf{r}(670) \oslash \mathbf{r}(490) + p_2$	Random Forest	[22]
Menken	$\mathbf{y} = p_1 \times (\mathbf{r}(670) \oslash \mathbf{r}(571))^{p_2}$	Artificial Neural Network	[23]

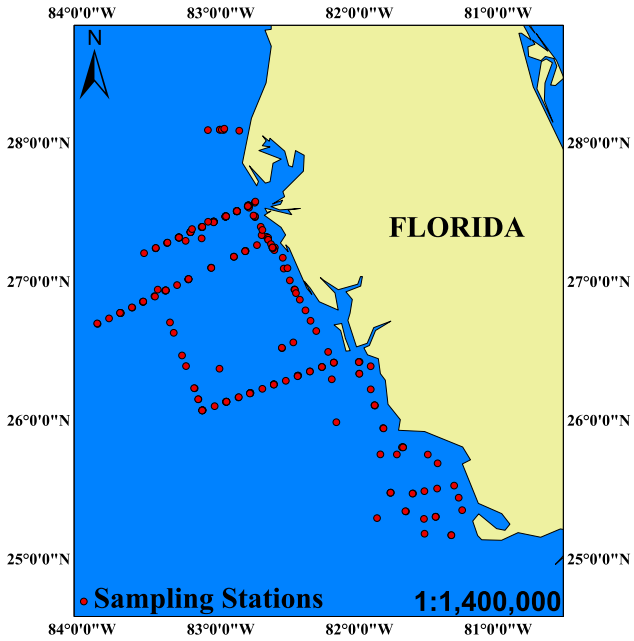


Fig. 2. Map of the sampling locations of the extracted samples from the NOMAD dataset.

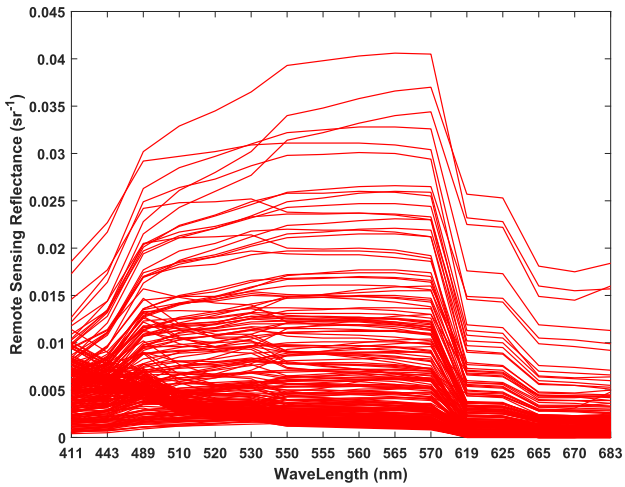


Fig. 3. Remote sensing reflectance of the extracted samples from the NOMAD dataset.

in which $\mathbf{y}_{\text{est}}$ denotes the inversion results of the test samples, $\mathbf{y}_{\text{true}}$ stands for their corresponding observed values, $\overline{\mathbf{y}}_{\text{true}}, \overline{\mathbf{y}}_{\text{est}}$ represent their mean values, and N is the size of the test set. For the RMSE and Bias, the closer they are to zero, the more likely they are to be accurate. However, r should be close to unity for good results.

B. NOMAD Dataset Experiments

The first dataset is a collection of *in situ* samples. The samples are extracted from the NASA bio-optical marine algorithm dataset (NOMAD) [47], which has been widely used to develop and validate ocean color inversion algorithms. In the experiments, 253 samples are utilized, each of which consists of 16 bands that lie within 411 nm $\sim$ 681 nm, while the magnitudes of the samples vary from 0.0044 m^{-1} to 0.8213 m^{-1}. The locations of the sampling stations and the remote sensing reflectance of the selected samples at various wavelengths are, respectively, shown in Figs. 2 and 3.

The RMSE statistics for different algorithms on this dataset are plotted in Fig. 4. The performance of the proposed algorithms improves as the training sample size increases, with their RMSEs dropping from $0.07 \sim 0.08$ to $0.05 \sim 0.06$ when the training sample proportion changes from 0.1 to 0.9. The accuracy of the LASSO2 algorithm is fairly consistently better than that of LASSO1 in the various experiments; moreover, the advantage in general tends to be larger when more training samples are available. One possible explanation for this behavior is that the high dimensionality burden that plagues the refinement procedure of the LASSO2 algorithms will be greatly alleviated in this case; thus, the merit of the LASSO2 algorithm in identifying the global-optimal feature subset can be best achieved, and thus, provide more precise results.

When compared to parametric regression algorithms, both of the new algorithms deliver superior results in most cases. The only exceptional case occurred when the training sample proportion was $P = 0.1$; the Tiwari model outperformed the proposed algorithms and yielded competitive performance under other proportions. This event seems to be reasonable since the Tiwari model was developed using the NOMAD dataset, that is, its model form and features had been elaborately designed. Then, the only task for this model in the experiments was to refine the model parameters regardless of the feature selection and model form determination; therefore, its inversion accuracy will not be considerably affected by the limitation of the training sample size. In contrast to the Tiwari model, both the Mannino model and Menken model produce much worse results compared to the proposed algorithms. The reasons are twofold. On the one hand, the refitting procedures of these two algorithms are prone

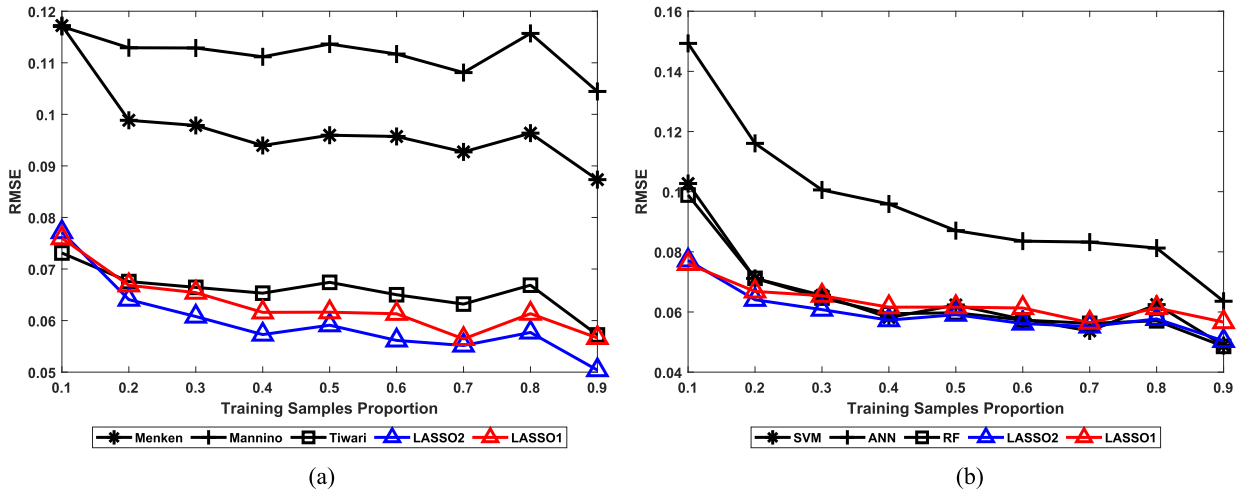


Fig. 4. RMSE comparison for the NOMAD dataset as a function of the algorithm and of the training set proportion. (a) RMSE results obtained by the new algorithms and the parametric inversion algorithms. (b) RMSE results obtained by the new algorithms and the nonparametric inversion algorithms.

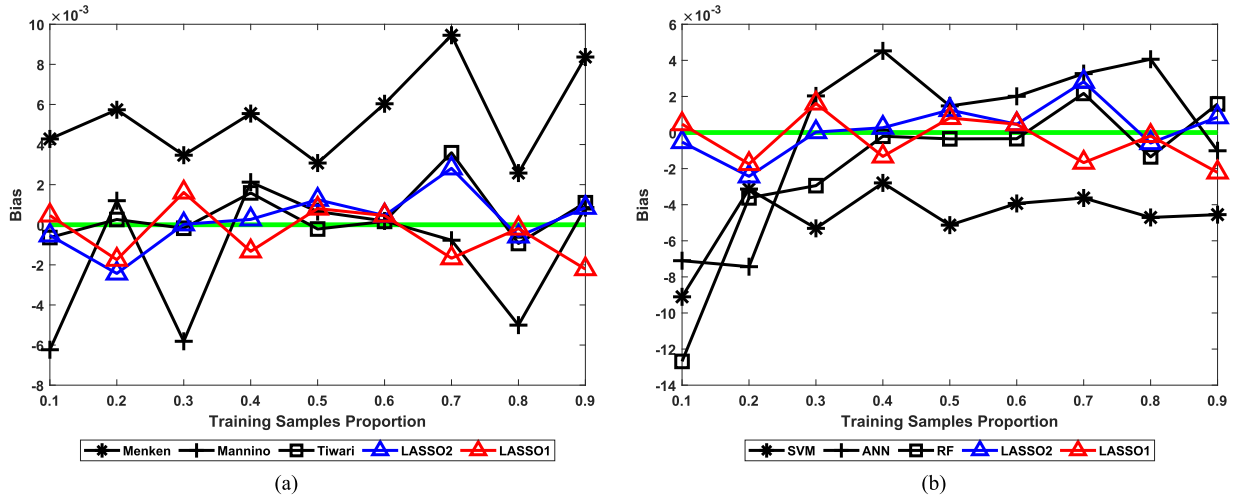


Fig. 5. Bias comparison for the NOMAD dataset as a function of the algorithm and of the training set proportion. (a) Bias results obtained by the new algorithms and the parametric inversion algorithms. (b) Bias results obtained by the new algorithms and the nonparametric inversion algorithms.

to becoming trapped in local optima due to their model forms, therefore leading to underfitting. On the other hand, the band-ratio term used in these models may not be sensitive enough to the labels in this dataset. These facts have also highlighted the need to perform feature selection when retrieving the CDOM content in a new research area.

When compared with nonparametric algorithms, we can observe that the LASSO2 algorithm also achieves predominant results in most cases, especially when the training sample proportion is relatively low ($P \leq 0.3$). The LASSO1 algorithm yields competitive performance in the case of small sample sizes; however, its performance has been slightly exceeded by the RF algorithm and the SVM algorithm as the training sample proportion increases. By comparison, both the RF and SVM algorithms produce relatively good results in comparison with the parametric algorithms and outperform the proposed algorithms in some experiments, particularly when the training sample proportion is higher ($P > 0.3$). This scenario is not unexpected because these algorithms are highly nonlinear and can discover more complex potential relationships between the features and the labels than the proposed algorithms at the cost of collecting more *in situ* samples. On the other hand, their superior results in terms of high training sample proportions have also motivated us to consider more types of nonlinear models and band arithmetic terms in future research to achieve more accurate results. The ANN algorithm does not perform as well as the RF and SVM algorithms, which may be caused by the fact that its model structure is more complicated and has stricter requirements for training sample size.

The Bias results shown in Fig. 5 reveal that the two proposed algorithms neither severely overestimate nor underestimate the CDOM content across various experiments when compared

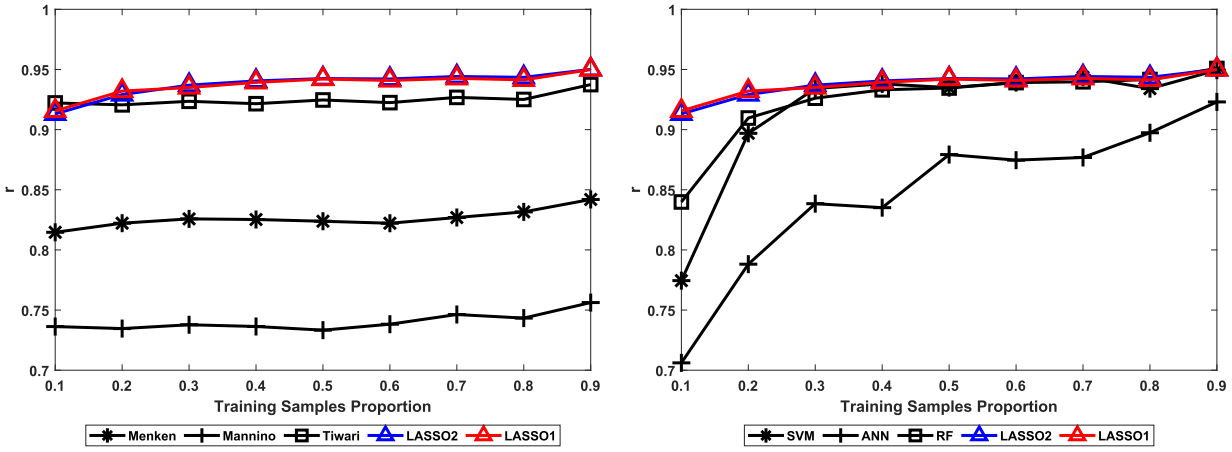


Fig. 6. r comparison for the NOMAD dataset as a function of the algorithm and of the training set proportion. (a) r results obtained by the new algorithms and the parametric inversion algorithms. (b) r results obtained by the new algorithms and the nonparametric inversion algorithms.

to other algorithms. Moreover, their changes do not exhibit large oscillations, which means that their results distribute more systematically around the true values. Regarding the compared algorithms, only the Tiwari algorithm provides relatively stable results. In Fig. 6, we further depict the average linear correlation results between the predicted labels and the true labels for different algorithms. Both of the proposed algorithms achieve excellent performance, with their statistics consistently higher than 0.9, specifically growing from approximately 0.91 to approximately 0.95 as the training sample proportions increase from 0.1 to 0.9. Moreover, the results reported by them are quite close across all the experiments. Concerning the compared algorithms, a notable trend is that the nonparametric algorithms are more sensitive to the training sample sizes than the parametric algorithms, with their results exhibiting dramatic improvements when the training sample proportions change from 0.1 to 0.3. The SVM algorithm and the RF algorithm in particular provide comparable results when the training sample proportion has reached or exceeded 0.3. In addition, the Tiwari algorithm still offers stable and competitive performance in these experiments. As emphasized before, a notable advantage of the proposed algorithms is their capabilities in offering intuitive and interpretable models in comparison with nonparametric regression algorithms. Hence, we would like to present the feature selection results obtained by the proposed algorithms. To alleviate the randomness originating from sample splitting imposed on the solutions, we apply the proposed algorithms in 40 Monte Carlo runs and calculate the frequency of the features being selected and their magnitudes in the model. The specific regression function will not be given, as there may be so many features being selected that it may be inconvenient to display. Note that the magnitudes recorded here are used to represent the relative importance of the features, and such implementation only holds in case whereby the feature matrix and the label vector have been standardized using their means and standard deviations before the training procedure. For filtering the accidental results, only the features with frequencies greater than 10 (25% of the total) will be considered significant features, as shown in Fig. 7.

The two new algorithms present nearly the same feature selection results, and the only slight difference lies in their frequency and magnitudes. Combining the feature selection results obtained by the algorithms, we can easily find that most of the selected features are the red/blue band-ratio terms, and no band multiplicative terms are chosen. Among these terms, the 665/443 term is recognized as the most sensitive feature, which has the highest frequency and the largest magnitudes. Note that the feature 670/489 used by the Tiwari model is also indicated; however, the results reveal that its importance is lower than that of the 665/443 and 665/489 ratio terms. In addition, the single band term at 411 nm is also selected.

The selection results have clear biooptical interpretations. First, CDOM exhibits strong absorption in the blue region of the visible light spectrum; hence, it is reasonable to set the blue region reflectance as the denominator of the band-ratio terms so that a decrease in reflectance in this band can reflect an increasing content of CDOM and vice versa [48]. Second, the reason for using 665 nm or its adjacent wavelength 670 nm to normalize the blue region spectrum can be divided into two aspects. First, the absorption of CDOM in this spectrum region is negligible; thus, it will not overtake the setting of the denominator. In addition, the reflectance in this spectrum region is considered less variable because the absorption of chlorophyll is approximately offset by its backscattering [49]. Concerning the only selected single-band term, the remote sensing reflectance at 411 nm, the possible reason for its low selection frequency is that LASSO inclines to select only one feature from a group of highly correlated candidate features [50], [51]. Therefore, once the terms using the remote sensing reflectance at 443 nm have been identified as the optimal features, the terms constructed by the remote sensing reflectance at 411 nm are more likely to be filtered out because the linear correlation coefficient between the remote sensing reflectance at these two wavelengths exceeds 0.9 in this dataset.

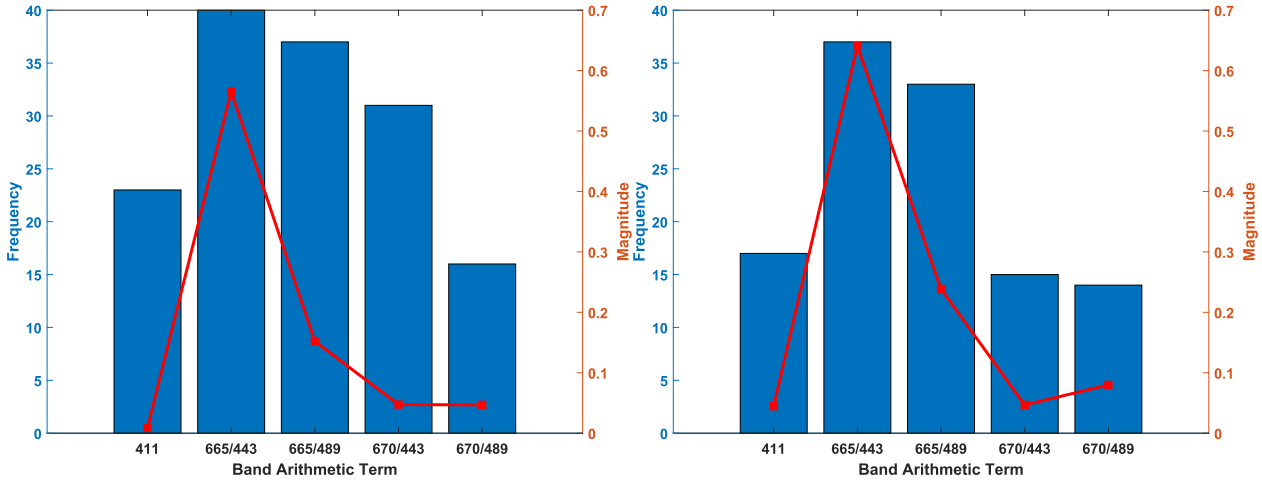


Fig. 7. Feature selection results obtained by the proposed algorithms. (a) LASSO1. (b) LASSO2.

This is also the cause for the significant selection frequency differences among the ratios using close wavelengths, e.g., 665 and 670 nm.

Combining all the experimental results from this dataset, we can conclude that both of the proposed algorithms can deliver superior or at least competitive results under different training sample sizes in comparison to the classical CDOM inversion algorithms while providing explicit and physically meaningful models.

C. IOCCG Dataset Experiments

The second dataset is a synthesized dataset acquired from IOCCG Report 5 [52]. As stated in the report, this dataset is generated from a well-known water body radiative transfer numerical simulation tool called Hydrolight, which can readily simulate AOPs under different environmental factors and inherent optical properties (IOPs). In this article, a total of 500 samples are selected to conduct the experiments. The wavelengths of the samples range from 400 to 800 nm, with a spectral resolution of 10 nm, and the magnitudes range from 0.0025 to 2.3677 m^{-1}. The remote sensing reflectances of the experimental samples at various wavelengths are shown in Fig. 8. As the available bands of this dataset are larger than those of the NOMAD dataset, the dimensions of the inversion problem will also be much higher, and consequently, be more difficult to solve. In this context, it is preferable to transform the final solution into the least-square form or the relaxed form [53].

Fig. 9 plots the RMSE results of different algorithms under various training sample proportions on this dataset. Similar to the NOMAD dataset experiments, both the LASSO1 and LASSO2 algorithms outperform the comparison algorithms in most cases, and their accuracy steadily improves as the training sample proportion increases, with their RMSEs dropping from approximately 0.1 to approximately 0.05. Meanwhile, the LASSO2 algorithm performs slightly better than the LASSO1 algorithm when the training sample proportion is less than or

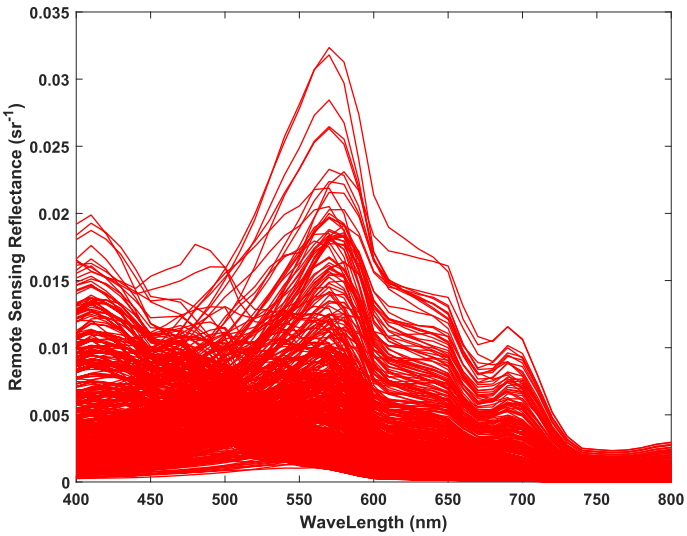


Fig. 8. Remote sensing reflectance of the extracted samples from the IOCCG dataset.

equal to 0.3, and in other cases, their RMSE results are nearly the same. Concerning the compared algorithms, all the parametric regression algorithms provide poor results, even though the training sample sizes have been increased. This may be because their features are not sensitive to the labels on this dataset; this behavior again proves the fact that it is necessary to redetermine the optimal features for the parametric regression methods when the data source has been changed. In contrast to parametric regression methods, nonparametric algorithms produce more reliable results, especially the SVM algorithm, and its inversion accuracy is close to that of our proposed algorithms. Finally, it should be noted that the ANN algorithm is not implemented on this dataset because the space required for storing the temporary variables exceeded the memory limits of our computers.

According to the Bias results displayed in Fig. 10, the results obtained by the proposed algorithms are closer to zero and show

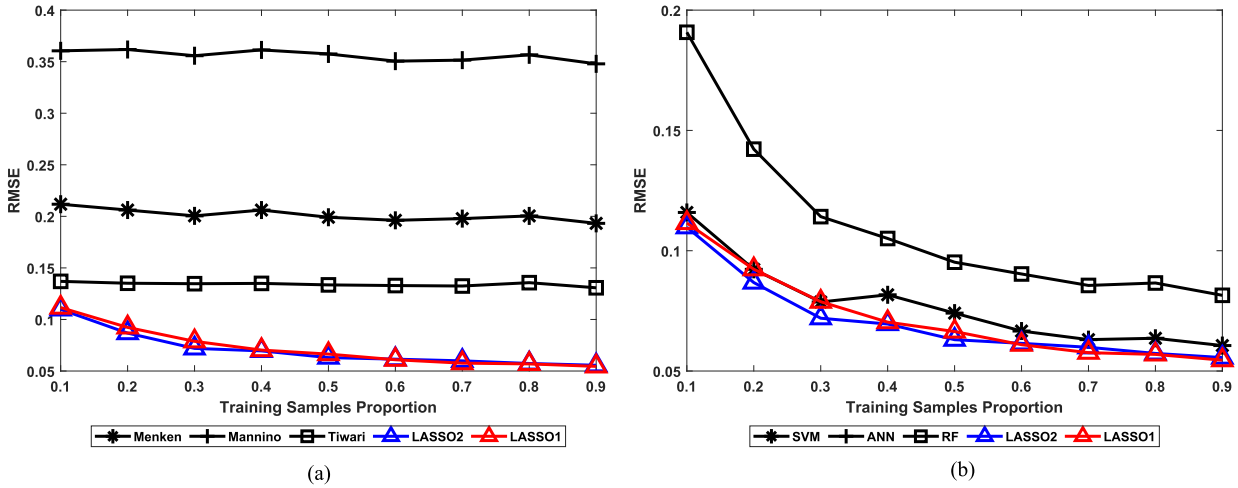


Fig. 9. RMSE comparison for the IOCCG dataset as a function of the algorithm and of the training set proportion. (a) RMSE results obtained by the new algorithms and the numerical regression methods. (b) RMSE results obtained by the new algorithms and the machine learning methods.

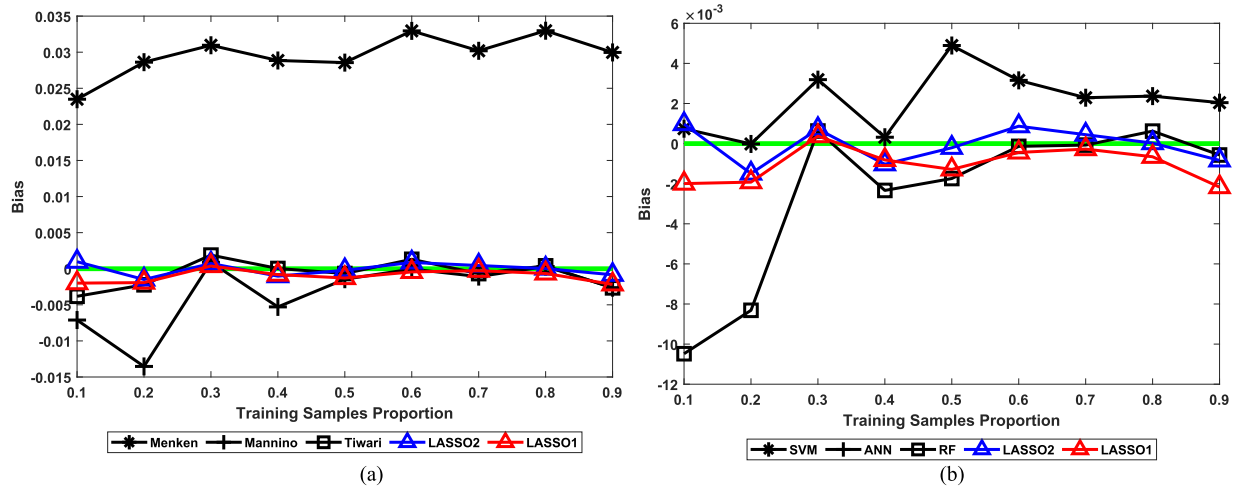


Fig. 10. Bias comparison for the IOCCG dataset as a function of the algorithm and of the training set proportion. (a) Bias results obtained by the new algorithms and the parametric inversion algorithms. (b) Bias results obtained by the new algorithms and the nonparametric inversion algorithms.

less perturbation in most experiments, which indicates that their overall systematic errors are lower than those of the comparison algorithms. On the other hand, it can also be observed that the LASSO1 algorithm tends to slightly underestimate the true values since nearly all of its biased results are negative. Regarding the comparison algorithms, the Menken algorithm and the SVM algorithm consistently deliver overestimated results, while the Mannino algorithm and the RF algorithm generally do the opposite. The Tiwari algorithm also yields low systematic error results; however, its results exhibit changes when the training sample proportion is relatively low ($P \leq 0.3$).

The average linear correlation results between the true labels and the predicted labels are depicted in Fig. 11. The results offered by the two presented algorithms are generally correlated with the true labels; even in the trickiest case, in which the training sample proportion is 0.1, the r statistics both exceed 0.97. In addition, LASSO2 tends to perform slightly better than LASSO1 when the training sample proportion is $P \leq 0.5$, but in general, the superiority is not significant when compared with the performance of other algorithms. Regarding the comparison algorithms, the Tiwari algorithm and the SVM algorithm provide satisfactory results and are not far behind the proposed algorithms. However, according to the RMSE statistics, there are large inversion errors in the Tiwari algorithm results. This fact also reveals the limitation of these metrics in reflecting the model errors.

The feature selection results are shown in Fig. 12. It should be noted that as the features being selected are far greater in number than on the NOMAD dataset, for illustrative convenience, we have to adjust the proportion of selected features to be displayed so that only the features with frequencies higher than 0.8 for the LASSO1 algorithm and 0.5 for the LASSO2 algorithm are

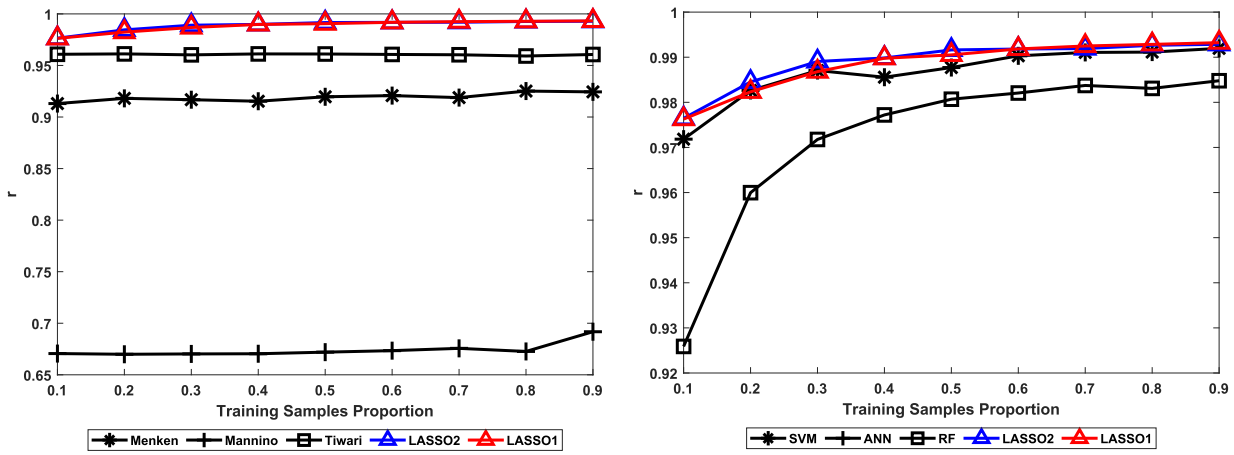


Fig. 11. r comparison for the IOCCG dataset as a function of the algorithm and of the training set proportion. (a) r results obtained by the new algorithms and the parametric inversion algorithms. (b) r results obtained by the new algorithms and the nonparametric inversion algorithms.

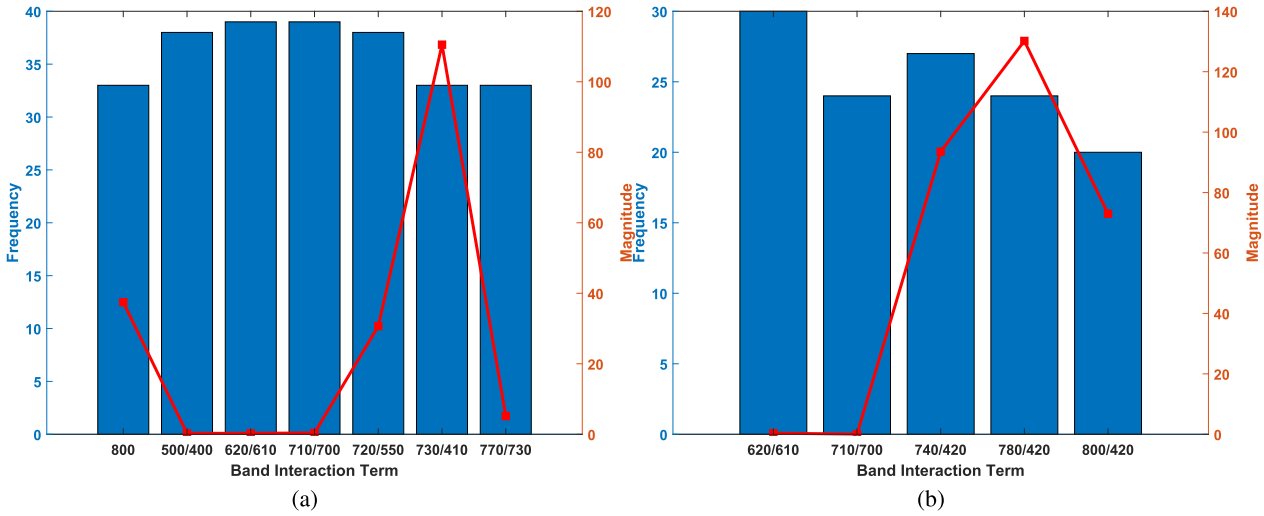


Fig. 12. Feature selection results obtained by the proposed algorithms. (a) LASSO1. (b) LASSO2.

displayed. We can see that the features selected by the proposed algorithms are quite different in terms of the wavelength being used. Despite these differences, there is a notable tendency of the selection results, that is, most features with large magnitudes favor red/blue or near-infrared/blue constructions. This finding is similar to the NOMAD dataset experiment, although the specific wavelengths are somewhat diverse. However, as the chlorophyll and suspended particular matter content vary significantly across the samples and given that the contribution of the inelastic scattering of the water constituents is also excluded from the AOPs, it may be difficult to further explain the biochemical principles associated with the results.

In summary, the two proposed algorithms consistently provide outstanding performance under various training sample proportion settings on this dataset, and the major trends of the feature selection results are also similar to the NOMAD dataset experiments. Nonetheless, the effect of hyperspectral information on the CDOM inversion will require further examination with more *in situ* hyperspectral datasets.

D. *Effects of Imbalanced Sampling*

In the previous experiments, the training samples were uniformly sampled from the whole dataset, which ensured that the trained model can reliably represent the relationship between the features and labels over different CDOM concentration ranges. However, such conditions may be difficult to attain and would thus considerably deteriorate the stabilities of the algorithms, especially when inverting large-scale and complex water bodies. This issue will become more difficult when only a few *in situ* samples are available. Hence, with the aim of evaluating the effectiveness of the algorithms toward such challenging scenarios, it is necessary to examine their generalization abilities under imbalanced sampling. For this purpose, we conduct three

TABLE II
EXPERIMENTAL RESULTS UNDER VARIOUS IMBALANCED SAMPLING STATES

Experiment 1								
Metrics	**LASSO1**	**LASSO2**	**Tiwari**	**Mannino**	**Menken**	**ANN**	**RF**	**SVM**
RMSE	**0.0123**	0.0155	0.0162	0.1595	0.0569	0.0237	0.0476	0.0267
Bias	0.0108	-0.0130	-0.0146	0.1496	-0.0511	-0.0232	**-0.0045**	-0.0188
r	**0.3751**	0.2140	0.3148	0.3059	0.3748	0.0905	0.0169	0.0383
Experiment 2								
Metrics	**LASSO1**	**LASSO2**	**Tiwari**	**Mannino**	**Menken**	**ANN**	**RF**	**SVM**
RMSE	**0.0273**	0.0394	0.0338	0.1323	0.0406	0.0645	0.0535	0.0320
Bias	**0.0019**	-0.0134	-0.0137	-0.1145	-0.0230	-0.0424	-0.0172	-0.0064
r	0.7360	0.7483	0.7198	0.7764	0.7311	0.7289	0.6463	**0.7611**
Experiment 3								
Metrics	**LASSO1**	**LASSO2**	**Tiwari**	**Mannino**	**Menken**	**ANN**	**RF**	**SVM**
RMSE	0.1639	**0.1372**	0.1467	0.2591	0.1996	0.2611	0.2670	0.2523
Bias	0.1229	**0.0902**	0.1023	0.1997	0.1417	0.1957	0.1758	0.1881
r	0.8070	0.8081	**0.8308**	0.7947	0.7734	0.2697	-0.2541	0.3850

The bold entities denote the best results.

inversion experiments on the NOMAD dataset by splitting the samples into three groups based on their CDOM concentrations. The experiments are implemented as follows:

1) *Experiment 1:* Training using the medium and high concentration sample groups, testing using the low concentration sample group.
2) *Experiment 2:* Training using the low and high concentration sample groups, testing using the medium concentration sample group.
3) *Experiment 3:* Training using the low and medium concentration sample groups, testing using the high concentration sample group.

It is worth noting that, these experimental settings correspond to the trickiest imbalanced sampling scenarios, because in practical applications, the CDOM concentration range of training samples should cover a wide, or at least part of the whole concentration range of the research area. The results from the three experiments are shown in Table II. According to the experimental results, we can observe that both of the proposed algorithms present good performance in comparison with other algorithms; meanwhile, the LASSO1 algorithm yields the best results in Experiments 1 and 2, while LASSO2 obtains the highest overall accuracy in Experiment 3. This behavior is slightly different from the results acquired in the aforementioned experiments in which the LASSO2 algorithm consistently outperforms the LASSO1 algorithm. One possible reason for this is that, regarding the LASSO1 algorithm, only some of the features will enter the second stage, thus reducing the uncertainties within the refinement procedure under such extreme circumstances and leading to more robust results. At the same time, the advantage of LASSO2 in identifying the global solution has significantly weakened due to the dual influence of the high dimensionality and imbalanced sampling. Concerning the comparison algorithm, the performance of Tiwari approaches that of the proposed algorithms because its feature has been determined based on the experimental dataset, which covers the full CDOM concentration range. In addition, benefiting from the good linear relationship between its feature and the labels, its linear model form favors decreasing the risk of underfitting and better endows it with more stable characteristics compared to other nonlinear models in this case.

Generally, compared with the optimal results given by the classical algorithms, the proposed algorithms can deliver better performance in terms of RMSE and Bias, and comparable accuracy with regard to r even in the trickiest cases of imbalanced sampling.

V. CONCLUSION

In this article, we have presented two new empirical CDOM retrieval algorithms based on a two-stage LASSO inversion framework, which enjoy great flexibility for addressing the

high-dimensional inversion task. The experiments conducted on the *in situ* bio-optical dataset and the synthesized dataset demonstrate that both of the proposed algorithms deliver superior results in small-sample cases and competitive results when more training samples are available, in contrast to current CDOM inversion algorithms. Furthermore, we can find that the features selected by the proposed algorithms enjoy great biochemical interpretability and are supported by many existing studies.

Finally, it should be noted that this article not only intends to propose new CDOM retrieval algorithms but also presents a framework that is of great extensibility. On the one hand, the two-stage inversion framework can be easily adapted to invert other water quality parameters. On the other hand, the band arithmetic terms used by existing algorithms can be readily embedded into the proposed algorithms by stacking them onto the feature matrix. In future work, we will further extend the proposed algorithms into nonlinear versions by replacing the regression function with a nonlinear function or incorporating other nonlinear band arithmetic terms because the superiorities of nonparametric regression algorithms in large-training-proportion scenarios have highlighted the fact that the relationship between the features and the labels is more complex than we have considered. Additionally, the variability of feature selection results under different CDOM concentration ranges will also be considered.

Acknowledgment

The authors would like to thank the NASA Ocean Biology Processing Group, the SeaBASS, and the International Ocean-Colour Coordinating Group for compiling high-quality bio-optical datasets for them to develop and validate the new algorithms. The authors would also like to thank Prof. K.L. Carder for sharing the precious *in situ* biooptical dataset in the NOMAD dataset.

References

[1] F. E. Hoge, A. Vodacek, R. N. Swift, J. K. Yungel, and N. V. Blough, "Inherent optical properties of the ocean: Retrieval of the absorption coefficient of chromophoric dissolved organic matter from airborne laser spectral fluorescence measurements," *Appl. Opt.*, vol. 34, no. 30, pp. 7032–7038, Oct. 1995. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-34-30-7032

[2] J. Chen, W.-N. Zhu, Y. Q. Tian, and Q. Yu, "Estimation of colored dissolved organic matter from Landsat-8 imagery for complex inland water: Case study of Lake Huron," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2201–2212, Apr. 2017.

[3] S. A. Green and N. V. Blough, "Optical absorption and fluorescence properties of chromophoric dissolved organic matter in natural waters," *Limnol. Oceanogr.*, vol. 39, no. 8, pp. 1903–1916, 1994.

[4] M. J. Sayers *et al.*, "Spatial and temporal variability of inherent and apparent optical properties in western Lake Erie: Implications for water quality remote sensing," *J. Great Lakes Res.*, 2019, vol. 45, no. 3, pp. 490–507, 2019.

[5] J. Wei, Z. Lee, M. Ondrusek, A. Mannino, M. Tzortziou, and R. Armstrong, "Spectral slopes of the absorption coefficient of colored dissolved and detrital material inverted from UV-visible remote sensing reflectance," *J. Geophys. Res., Oceans*, vol. 121, no. 3, pp. 1953–1969, 2016.

[6] D.-P. Häder, E. Helbling, C. Williamson, and R. Worrest, "Effects of UV radiation on aquatic ecosystems and interactions with climate change," *Photochem. Photobiol. Sci.*, vol. 10, no. 2, pp. 242–260, 2011.

[7] C. G. Fichot and R. Benner, "A novel method to estimate DOC concentrations from CDOM absorption coefficients in coastal waters," *Geophys. Res. Lett.*, vol. 38, no. 3, 2011, doi: 10.1029/2010GL046152.

[8] E. T. Harvey, S. Kratzer, and A. Andersson, "Relationships between colored dissolved organic matter and dissolved organic carbon in different coastal gradients of the Baltic Sea," *Ambio*, vol. 44, no. 3, pp. 392–401, 2015.

[9] J. Xu *et al.*, "Optical models for remote sensing of chromophoric dissolved organic matter (CDOM) absorption in Poyang Lake," *ISPRS J. Photogrammetry Remote Sens.*, vol. 142, pp. 124–136, 2018.

[10] Z. Wen, K. Song, Y. Zhao, J. Du, and J. Ma, "Influence of environmental factors on spectral characteristics of chromophoric dissolved organic matter (CDOM) in Inner Mongolia Plateau, China," *Hydrol. Earth Syst. Sci.*, vol. 20, no. 2, pp. 787–801, 2016.

[11] J. R. Helms, A. Stubbins, J. D. Ritchie, E. C. Minor, D. J. Kieber, and K. Mopper, "Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter," *Limnol. Oceanogr.*, vol. 53, no. 3, pp. 955–969, 2008.

[12] Y. Zhang, X. Liu, M. Wang, and B. Qin, "Compositional differences of chromophoric dissolved organic matter derived from phytoplankton and macrophytes," *Organic Geochem.*, vol. 55, pp. 26–37, 2013.

[13] A. Bricaud, A. Morel, and L. Prieur, "Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domains 1," *Limnology Oceanogr.*, vol. 26, no. 1, pp. 43–53, 1981.

[14] F. E. Hoge and P. E. Lyon, "Satellite retrieval of inherent optical properties by linear matrix inversion of oceanic radiance models: An analysis of model and radiance measurement errors," *J. Geophys. Res., Oceans*, vol. 101, no. C7, pp. 16 631–16 648, 1996.

[15] V. E. Brando and A. G. Dekker, "Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1378–1387, Jun. 2003.

[16] B. B. Barnes, R. Garcia, C. Hu, and Z. Lee, "Multi-band spectral matching inversion algorithm to derive water column properties in optically shallow waters: An optimization of parameterization," *Remote Sens. Environ.*, vol. 204, pp. 424–438, 2018.

[17] A. Y. Morel and H. R. Gordon, "Report of the working group on water color," *Boundary-Layer Meteorol.*, vol. 18, no. 3, pp. 343–355, 1980.

[18] D. Y. Sun *et al.*, "A neural-network model to retrieve CDOM absorption from *in situ* measured hyperspectral data in an optically complex lake: Lake Taihu case study," *Int. J. Remote Sens.*, vol. 32, no. 14, pp. 4005–4022, 2011.

[19] S. Heddam, "Generalized regression neural network (GRNN)-based approach for colored dissolved organic matter (CDOM) retrieval: Case study of Connecticut river at middle Haddam Station, USA," *Environmental Monit. Assessment*, vol. 186, no. 11, pp. 7837–7848, 2014.

[20] M. Hieronymi, D. Müller, and R. Doerffer, "The OLCI neural network swarm (ONNS): A bio-geo-optical algorithm for open ocean and coastal waters," *Frontiers Marine Sci.*, vol. 4, 2017, Art. no. 140, doi: 10.3389/fmars.2017.00140.

[21] J. Zhao *et al.*, "Estimating CDOM concentration in highly turbid estuarine coastal waters," *J. Geophys. Res., Oceans*, vol. 123, no. 8, pp. 5856–5873, 2018.

[22] A. Ruescas, M. Hieronymi, G. Mateo-Garcia, S. Koponen, K. Kallio, and G. Camps-Valls, "Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 786, doi: 10.3390/rs10050786.

[23] S. Keller *et al.*, "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity," *Int. J. Environmental Res. Public Health*, vol. 15, no. 9, 2018, Art. no. 1881, doi: 10.3390/ijerph15091881.

[24] K. Takezawa, *Introduction to Nonparametric Regression*, vol. 606. Hoboken, NJ, USA: Wiley, 2005.

[25] J. Shao, *Mathematical Statistics*. New York, NY, USA: Springer-Verlag, 2003.

[26] E. J. D'Sa, "Colored dissolved organic matter in coastal waters influenced by the Atchafalaya River, USA: Effects of an algal bloom," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 023502.

[27] A. Mannino, M. E. Russ, and S. B. Hooker, "Algorithm development and validation for satellite-derived distributions of DOC and CDOM in the US Middle Atlantic Bight," *J. Geophys. Res., Oceans*, vol. 113, no. C7, 2008, doi: 10.1029/2007JC004493.

[28] T. Kutser, D. C. Pierson, K. Y. Kallio, A. Reinart, and S. Sobek, "Mapping lake CDOM by satellite remote sensing," *Remote Sens. Environ.*, vol. 94, no. 4, pp. 535–540, 2005.

[29] S. Koponen *et al.*, "A case study of airborne and satellite remote sensing of a spring bloom event in the Gulf of Finland," *Continental Shelf Res.*, vol. 27, no. 2, pp. 228–244, 2007.

[30] S. Tiwari and P. Shanmugam, "An optical model for the remote sensing of coloured dissolved organic matter in coastal/ocean waters," *Estuarine, Coastal Shelf Sci.*, vol. 93, no. 4, pp. 396–402, 2011.

[31] K. Kallio *et al.*, "Landsat ETM+ images in the estimation of seasonal lake water quality in Boreal River basins," *Environmental Manage.*, vol. 42, no. 3, pp. 511–522, 2008.

[32] W. Zhu, Q. Yu, Y. Q. Tian, B. L. Becker, T. Zheng, and H. J. Carrick, "An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments," *Remote Sens. Environ.*, vol. 140, pp. 766–778, 2014.

[33] M. W. Matthews, "A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters," *Int. J. Remote Sens.*, vol. 32, no. 21, pp. 6855–6899, 2011.

[34] P. Ammenberg, P. Flink, T. Lindell, D. Pierson, and N. Strombeck, "Bio-optical modelling combined with remote sensing to assess water quality," *Int. J. Remote Sens.*, vol. 23, no. 8, pp. 1621–1638, 2002.

[35] D. Doxaran, R. Cherukuru, and S. Lavender, "Use of reflectance band ratios to estimate suspended and dissolved matter concentrations in estuarine waters," *Int. J. Remote Sens.*, vol. 26, no. 8, pp. 1763–1769, 2005.

[36] H. Loisel, V. Vantrepotte, D. Dessailly, and X. Mriaux, "Assessment of the colored dissolved organic matter in coastal waters from ocean color remote sensing," *Opt. Exp.*, vol. 22, no. 11, 2014, Art. no. 13109.

[37] A. Campanelli *et al.*, "An empirical ocean colour algorithm for estimating the contribution of coloured dissolved organic matter in north-central western Adriatic Sea," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 180, doi: 10.3390/rs9020180.

[38] A. Mannino, M. G. Novak, S. B. Hooker, K. Hyde, and D. Aurin, "Algorithm development and validation of CDOM properties for estuarine and continental shelf waters along the northeastern US coast," *Remote Sens. Environ.*, vol. 152, pp. 576–602, 2014.

[39] F. Cao *et al.*, "Remote sensing retrievals of colored dissolved organic matter and dissolved organic carbon dynamics in North American estuaries and their margins," *Remote Sens. Environ.*, vol. 205, pp. 151–165, 2018.

[40] A. Morel and B. Gentili, "A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data," *Remote Sens. Environm.*, vol. 113, no. 5, pp. 998–1011, 2009.

[41] H. Zou, "The adaptive LASSO and its oracle properties," *J. Amer. Statistical Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[42] A. K. Chattopadhyay and T. Chattopadhyay, *Statistical Methods for Astronomical Data Analysis*, vol. 3. Berlin, Germany: Springer, 2014.

[43] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statistical Soc.. B*, pp. 267–288, 1996.

[44] P. A. Herrault, L. Gandois, S. Gascoin, N. Tananaev, T. L. Dantec, and R. Teisserenc, "Using high spatio-temporal optical remote sensing to monitor dissolved organic carbon in the Arctic River Yenisei," *Remote Sens.*, vol. 8, no. 10, 2016, Art. no. 803, doi: 10.3390/rs8100803.

[45] T. T. Wu *et al.*, "Coordinate descent algorithms for lasso penalized regression," *Ann. Appl. Statist.*, vol. 2, no. 1, pp. 224–244, 2008.

[46] K. D. Menken, P. L. Brezonik, and M. E. Bauer, "Influence of chlorophyll and colored dissolved organic matter (CDOM) on lake reflectance spectra: Implications for measuring lake properties by remote sensing," *Lake Reservoir Manag.*, vol. 22, no. 3, pp. 179–190, 2006.

[47] P. J. Werdell and S. W. Bailey, "An improved *in-situ* bio-optical data set for ocean color algorithm development and satellite data product validation," *Remote Sens. Environ.*, vol. 98, no. 1, pp. 122–140, 2005.

[48] D. Bowers, G. Harker, P. Smith, and P. Tett, "Optical properties of a region of freshwater influence (the Clyde Sea)," *Estuarine, Coastal Shelf Sci.*, vol. 50, no. 5, pp. 717–726, 2000.

[49] K. Kallio, J. Pulliainen, and P. Ylöstalo, "MERIS, MODIS and ETM+ channel configurations in the estimation of lake water quality from subsurface reflectance using semianalytical and empirical algorithms," *Geophysica*, vol. 41, no. 1–2, pp. 31–55, 2005.

[50] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statistical Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.

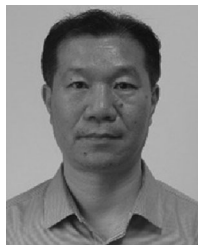
[51] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statistical Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[52] Z.-P. Lee *et al.*, *Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications.* Dartmouth, NS, Canada: International Ocean Colour Coordinating Group, 2006.

[53] T. Hastie, R. Tibshirani, and R. J. Tibshirani, "Extended comparisons of best subset selection, forward stepwise selection, and the LASSO," 2017.

Ruihao Zhang received the B.Sc. and M.Sc. degrees in geographic information system and human geography from South China Normal University, Guangzhou, China, in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree with the School of Geography and Planning, Sun Yat-sen University, Guangzhou.

His research interests include machine/statistical learning, nonconvex optimization and their applications in hyperspectral unmixing and ocean color remote sensing.



Ruru Deng received the B.S. degree in geology from the Wuhan College of Geology, Wuhan, China, in 1984, the M.E. degree in remote sensing geology from the China University of Geosciences, Wuhan, in 1989, and the Ph.D. degree in microwave remote sensing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2002.

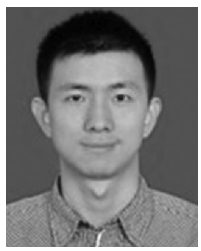
He is currently a Professor with Sun Yat-sen University, Guangzhou, China. His research interests include atmospheric correction of optical remote sensing data, quantitative remote sensing of inland waters and coral reefs, and application development to microwave remote sensing.

Yingfei Liu received the B.S. and Ph.D. degrees from the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China, in 2012 and 2018, respectively.

He has been a Postdoctoral Scholar with the School of Marine Sciences, Sun Yat-sen University, Zhuhai, China, since February 2019. His research interests include intelligent target detection and information retrieval from satellite imagery.

Yeheng Liang received the B.S., M.S., and Ph.D. degrees in geographic information system from Sun Yat-sen University, Guangzhou, China, in 2010, 2012, and 2016, respectively.

He is currently an Associate Research Fellow with the School of Geography and Planning, Sun Yat-sen University. His research interests include inversion of heavy metals in water from satellite images.



Longhai Xiong received the B.S. degree in geographic information system from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2011, the M.S. degree in marine geography, and the Ph.D. degree in cartography and geographic information system from Sun Yat-sen University, Guangzhou, China, in 2013 and from Sun Yat-sen University, 2018, respectively.

He has been a Postdoctoral Scholar with the School of Geography and Planning, Sun Yat-sen University, since December 2018. His research interests include surface water extraction, estimation of river runoff, soil moisture, and groundwater depth using remote sensing.

Bin Cao received the bachelor's degree in electronic and information engineering from Zhengzhou University, Zhengzhou, China, in 2015, and the master's degree in environmental science and engineering from Shanghai Ocean University, Shanghai, China, in 2018. He is currently working toward Ph.D. degree with the School of Geography and Planning, Sun Yat-sen University, China.

His research interests include optical remote sensing and bathymetry.

Wenzhi Zhang received the B.S. degree from Wuhan University, Wuhan, China, in 2005, and the M.S. degree from Hohai University, Nanjing, China, in 2015.

He is currently a Senior Engineer with the Huizhou branch of Guangdong Hydrology Bureau, Huizhou, China. His research interests include hydrology, water resources, and aquatic ecosystem assessments.