Remote Sensing Statistical Inference for Colored Dissolved Organic Matter in Inland Water: Case Study in Qiandao Lake

Weining Zhu

***Abstract*—Compared to traditional remote sensing classification and inversion techniques, remote sensing statistical inference is a novel method for rapidly estimating the statistical properties of ground objects. Despite some initial work, this method has not been thoroughly evaluated for water quality assessment. In this study, using field-measured data from Qiandao Lake, we tested over 240 000 inference models for determining the mean, median, standard deviation, minimum, and maximum of colored dissolved organic matter using a bootstrap approach and various combinations of bands, variables, and functions. The results indicated that all five statistical parameters could be inferred accurately with errors of less than 10%. The best models used two band ratios, three statistical variables, and polynomial functions. The study also demonstrated the importance of redistributing the raw field-measured data for improved performance, as models based on the redistributed data outperformed those based on the raw data.**

***Index Terms*—Colored dissolved organic matter (CDOM), inland water, remote sensing statistical inference.**

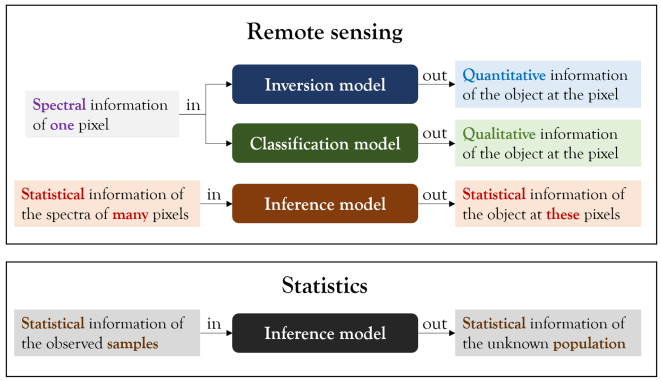

Fig. 1. Input and output of remote sensing inference, classification, inversion, and traditional statistical inference models.

I. Introduction

RECENTLY, a novel remote sensing analytical technique known as remote sensing statistical inference (also referred to as remote sensing inference) has been introduced for the remote sensing of aquatic environments [1], [2], [3], [4]. The theoretical basis for remote sensing inference lies in statistical optics and radiative transfer [5], [6], and it explores the connections between statistical characteristics of both the optical and physical properties of ground objects. For example, when the ground object is a lake, its optical property is the satellite-observed remote sensing reflectance (Rrs), whereas its nonoptical property is the water depth. Based on millions or thousands of lake pixels observed in satellite imagery, remote sensing inference examines the relationship between statistical features (such as the mean, median, min, and max) of Rrs and the statistical features (such as the mean, median, min, and max) of water depth.

Manuscript received 18 April 2023; revised 30 June 2023; accepted 30 July 2023. Date of publication 2 August 2023; date of current version 16 August 2023. This work was supported in part by the Science Foundation of Donghai Laboratory under Grant DH-2022KF01009 and in part by the National Natural Science Foundation of China under Grant 41971371 and Grant 41876031.

The author is with the Institute of Ocean Sensing and Networking, Department of Ocean Informatics, Ocean College, Zhejiang University, Zhoushan 80305, China (e-mail: zhuwn@zju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3301138

The input and output information of remote sensing inference differs from the well-known methods of classification and inversion, as shown in Fig. 1. Unlike classification and inversion, which are based on the spectra themselves, remote sensing inference is based on the statistical features of spectra [7], [8]. For each pixel among millions of lake pixels, it is possible to classify the water as deep or shallow and to determine its depth using inversion algorithms. However, the same work cannot be done by using inference because statistics are only meaningful when applied to a large number of pixels or samples. In most cases, the objects of interest on the ground consist of many pixels, not just one or a few. Using remote sensing inference, the mean, minimum, and maximum depths of a lake can be estimated to be 5 m, 1 m, and 40 m, respectively. Remote sensing statistical inference also differs from the traditional statistical inference used in geoscience and environmental studies [9], [10], [11]. In statistics, statistical inference typically involves estimating a population based on observed samples (see Fig. 1). When applied to remote sensing, it involves using remotely sensed information from sampled pixels to estimate the scenarios of the population pixels [12], [13]. For example, McRoberts and Walters demonstrated the use of remote sensing to estimate net deforestation [14].

One may wonder about the reasons and circumstances for using remote sensing inference. The information obtained from remote sensing is often applied at different levels of detail. At a basic level, only rough information with minimal detail may be

sufficient, such as determining whether an object on the ground is a lake or a forest, or whether the lake is deep or shallow. This can be achieved by classification. For more quantitative information about a lake, such as its depth distribution or the size of phytoplankton, inversion models can be used. Inference provides an intermediate level of information—statistical information about the object of interest. For example, the average temperature or depth of a lake, which may be of interest to environmental monitors.

If we know the exact depth at each pixel using inversion models, then the average depth of the lake can be easily calculated. However, when we use inversion models to estimate the average depth, we may encounter two major problems: low efficiency and low accuracy. Low efficiency means that we have to run an inversion model millions of times for millions of pixels of a lake. Low accuracy means that some inversion models are location-limited or may be ill-posed, causing their results to deviate from the population truth. Our preliminary work showed that the statistical information estimated from inference models was more accurate than that from the inversion model [15].

The first attempt at remote sensing inference modeling aimed to estimate the mean concentration of colored dissolved organic matter (CDOM) in Qiandao Lake [1]. However, the model was limited as it used a simple linear function ($y = kx + b$) and resulted in low accuracy. In addition, it failed to build a linear model for estimating the standard deviation (std) of CDOM. This study aims to develop and test more advanced inference models for aquatic remote sensing, with a continued focus on CDOM. As one of the three optical components of watercolor, CDOM is crucial for assessing water quality using remote sensing techniques [16], [17], [18]. While CDOM has been inverted from field-measured spectra and satellite imagery with acceptable accuracy in clear open seawater [19], [20], it remains a challenge for inversion models in complex inland or coastal waters [21], [22], [23].

In this study, we aim to infer a wider range of statistical parameters for CDOM, including the mean and std, as well as the median, minimum, and maximum. We will use a combination of linear and nonlinear functions, such as power and exponential functions. Our approach will consider not just one independent variable based on a single band or band ratio, but multiple independent variables based on multiple bands or band ratios. Our work will demonstrate the broad applicability of remote sensing inference models in aquatic environments.

II. Materials and Methods

A. Study Site, Field, and Lab Measurements

Qiandao Lake is a man-made reservoir located in Hangzhou, Zhejiang Province, China. The lake covers an area of 573 km^2 and provides water for nearly 15 million people in Zhejiang. From 2018 to 2019, the water quality and surface spectra of Qiandao Lake were measured five times. The study area map and sampling locations are shown in Fig. 2. At each sampling point, surface water samples were collected and preserved in amber bottles. The above-surface spectra, which included above-surface radiance, sky radiance, and downwelling irradiance,

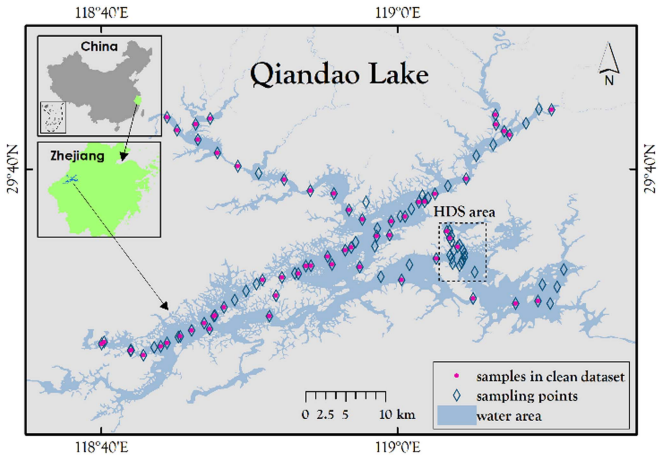

Fig. 2. Study site map and field sampling locations in Qiandao Lake. In the HDS area, samples were densely collected.

were measured using an ASD FieldSpec spectroradiometer (Analytical Spectral Devices, Inc.) and following the NASA measurement protocol [24]. The remote sensing reflectance was calculated from the three measured optical variables. After each field trip, water samples were delivered to our laboratory within 6 h for measurement of water quality parameters. CDOM absorption coefficients (measured in units of m^{-1}), a_{CDOM}, were measured using a Cary-100 spectrophotometer (Agilent Technologies, Inc.), and their values at 440 nm, a_{CDOM} (440), were used as a proxy for CDOM concentrations. A total of 100 samples were collected from the 5 field trips. Further details of the field and laboratory measurements can be found in our previous reports [25].

B. Bootstrap-Based Method for Inference Modeling and Validation

The primary objective of remote sensing statistical inference is to determine the relationship between the statistical features of the observed optical properties (X) of a ground object and the statistical features of other nonoptical properties (Y), such as its geophysical, ecological, and environmental properties. Since X is determined by Y, the forward relationship between X and Y is expressed as $X = f(Y)$. However, remote sensing inference focuses more on the backward relationship, which takes the form of $Y = f^{-1}(X) = g(X)$. In this study, we employed a bootstrap-based method to develop the model $Y = g(X)$ for the following cases of X, Y, and g.

1) *Statistical parameter Y of CDOM:* We examined five statistical parameters, i.e., mean, median, std, minimum, and maximum of CDOM's a_{CDOM}(440), to provide a rough description of its statistical distribution in Qiandao Lake. In this study, the mean, median, and std were calculated using their well-known formulas, whereas the minimum and maximum were obtained from several minimal or maximal values. For example, if the dataset consisted of 80 samples, the minimum/maximum a_{CDOM} (440) values were calculated from the mean value of the 5 minimal

TABLE I
STATISTICAL VARIABLES x OF SPECTRA USED FOR INFERRING STATISTICAL PARAMETER Y OF CDOM

CDOM	Spectra		
Y	x_1	x_2	x_3
Mean	Mean	Median	Std
Median	Median	Mean	Std
Std	Std	Mean	Median
Min	Min	Std	Max
Max	Max	Std	Min

or maximal $a_{CDOM}(440)$ values among the 80 samples. This was done because, in practical applications, some extremely high or low field-measured single values may not reflect the true minimum or maximum of the study water.

2) *Statistical variables X of spectra:* In our previous studies, Y and X in inference models were, respectively, the same statistical features of optical and nonoptical properties of the ground object. For example, the statistical features were both the mean values, that is, the mean a_{CDOM} (440) was inferred from the mean Rrs. In this study, we added more statistical variables X into an inference model, for example, the mean a_{CDOM} (440) was inferred from the mean, median, and std values of the Rrs, that is, one statistical parameter Y was inferred by other three statistical variables x_1, x_2, and x_3, see Table I.
3) *Bands or band ratios of X:* We examined four single bands (B1 443 nm, B2 483 nm, B3 551 nm, and B4 655 nm) of Landsat-8 and six single band ratios (B1/B2, B1/B3, B1/B4, B2/B3, B2/B4, and B3/B4) generated from the four bands. We also explored the use of two bands or two band ratios in one model. For the 4 single bands, there are 6 cases of two-band combinations, and for the 6 single band ratios, there are 15 cases of two-band-ratio combinations. A total of 31 cases were therefore examined, each using different bands or band ratios.
4) *Datasets for modeling and validation:* In this study, we used two datasets for inference modeling and validation. The first dataset D_{raw} contains all 100 samples collected during field trips, whereas the second dataset D_{clean} is a filtered version of D_{raw}. Fig. 3 shows the histogram of the 100 raw data points, which exhibit an approximately exponential statistical distribution. We excluded 40 samples from the raw data, and the remaining 60 samples comprise the clean dataset. D_{clean} represents a better exponential distribution of CDOM than D_{raw} since the samples in D_{clean} were chosen based on their distribution pattern. The 40 removed samples were mostly taken from densely collected samples within a small area (HDS area in Fig. 2) due to the slow speed of the cruise boat. These removed samples, called "statistically dirty data," could potentially distort the true statistical distribution of CDOM in Qiandao Lake.
5) *Model functions g:* In our study, we examined three types of model functions commonly used in remote sensing of CDOM: linear, power, and exponential functions. The formulas for these functions are presented in (1) to (8).

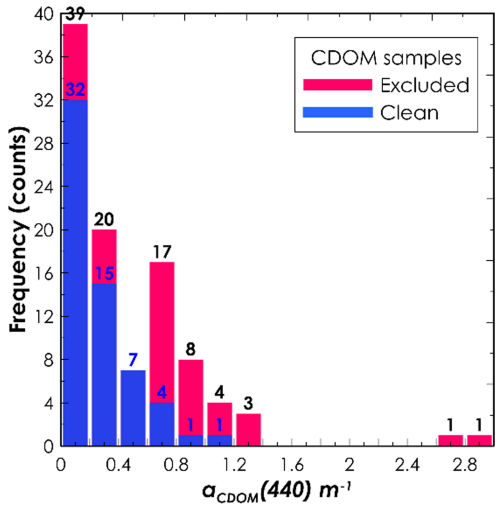

Fig. 3. Resampling the raw data (100 samples) to make a clean dataset (60 samples). Some data (40 samples) were excluded from the raw dataset, making the rest 60 samples present an exponential distribution.

Models using one statistical variable x_1 at one band or band ratio λ.

$$Y = c_1\, x_1\,(\lambda) + c_2 \tag{1}$$

$$Y = c_1\, x_1(\lambda)^{c_2} \tag{2}$$

$$Y = c_1 \exp\left(c_2 x_1\,(\lambda)\right). \tag{3}$$

Models using one statistical variable x_1 at two bands or band ratios λ_1 and λ_2.

$$Y = c_1\, x_1\,(\lambda_1) + c_2 x_1\,(\lambda_2) + c_3 \tag{4}$$

$$Y = c_1\, x_1(\lambda_1)^{c_2} + c_3 x_1(\lambda_2)^{c_4} \tag{5}$$

$$Y = c_1 \exp\left(c_2 x_1\,(\lambda_1)\right) + c_3 \exp\left(c_4 x_1\,(\lambda_2)\right). \tag{6}$$

Models using three statistical variables x_1, x_2, and x_3 at one band or band-ratio λ.

$$Y = c_1\, x_1(\lambda)^3 + c_2 x_2(\lambda)^2 + c_3 x_3\,(\lambda) + c_4. \tag{7}$$

Models using three statistical variables x_1, x_2, and x_3 at two bands or band ratios λ_1 and λ_2.

$$Y = c_1\, x_1(\lambda_1)^3 + c_2 x_1(\lambda_2)^3 + c_3 x_2(\lambda_1)^2 + c_4 x_2(\lambda_2)^2 + c_5 x_3\,(\lambda_1) + c_6 x_3\,(\lambda_2) + c_7. \tag{8}$$

Note that we consider both polynomial functions, i.e., (7) and (8), as linear functions in this study because their variable powers are fixed, and only the coefficients need to be adjusted. Power and exponential functions with three statistical variables are too complex to be explored in this study.

To evaluate the above inference models with different cases, we adopt a bootstrap-based inference modeling and validation procedure of the following seven steps.

Step 1: Load two modeling datasets, namely the raw dataset D_{raw} and the clean dataset D_{clean}, which contain 100 and 60 samples, respectively.

Step 2: Randomly select $n = 67$ samples from D_{raw} and $n = 40$ samples from D_{clean} to create two 1-boot modeling datasets D_{raw_mb} and D_{clean_mb}.

Step 3: Randomly select $n = 67$ samples from D_{raw} and $n = 40$ samples from D_{clean} to create two 1-boot validation datasets D_{raw_vb} and D_{clean_vb}, which are independent of D_{raw_mb} and D_{clean_mb}.

Step 4: Calculate the five statistical parameters (mean, median, std, min, and max) of a_{CDOM} (440) and Rrs at single bands, band ratios, and their combinations, based on the selected 67 or 40 samples in D_{raw_mb}, D_{clean_mb}, D_{raw_vb}, and D_{clean_vb}.

Step 5: Repeat Steps 2–4 five hundred times and record the results together to obtain a 500-boot modeling dataset $D_{raw\_modeling}$ and $D_{raw\_validation}$, $D_{clean\_modeling}$, and $D_{clean\_validation}$ for the clean dataset and validation.

Step 6: For the five statistical parameters Y we want to infer and fit model functions listed in (1)–(8) using datasets $D_{raw\_modeling}$ and $D_{clean\_modeling}$, and evaluate their performance by the indicators shown in the next section. Validate the fitted models by using the independent datasets, $D_{raw\_validation}$ and $D_{clean\_validation}$. Record the best model with respect to the given statistical indicator, based on the model with the best performance among all models [e.g., the highest R^2 or the lowest root mean squared error (RMSE)].

Step 7: Repeat Steps 2–6 two hundred times, counting the number of times each model performed as the best model. For each statistical parameter, the best model with the largest count is selected as the final inference model [26].

The flowchart of the above seven steps is presented in [1]. The bootstrap parameters, such as the resampling number and repetition times, have been tested within wide ranges to assure that the modeling results have no significant changes.

Taking into account all the functions from (1) to (8), we tested 124 models for each statistical parameter and dataset. As there were 5 statistical parameters and 2 datasets, a total of 248 models were tested for each statistical parameter, resulting in 1240 models tested for each modeling and validation process. As the processes were repeated 200 times, we tested a total of 248 000 models for CDOM inference in Qiandao Lake.

Model fits were performed using the MATLAB function fitnlm. It should be noted that in some cases, the measured Rrs and CDOM concentrations were too small to achieve convergence in the regression of some of the nonlinear models. To overcome this problem, Rrs and CDOM values had to be expanded by a factor of 10 before being used to fit the models.

C. Evaluation of Model Performance

We evaluated the inference models of CDOM and their performance using five statistical indicators: the coefficient of determination (R^2), RMSE, and the mean absolute percentage error (MAPE) for validation data. For modeling data, we calculated R^2 and RMSE. To compute the RMSE and MAPE, we used the following formula:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{\mathrm{estimated}\_i} - Y_{\mathrm{observed}\_i}\right)^2} \tag{9}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{Y_{\mathrm{observed}\_i} - Y_{\mathrm{estimated}\_i}}{Y_{\mathrm{observed}\_i}}\right| \tag{10}$$

where n is the number of samples or fitted points, $Y_{\mathrm{estimated}\_i}$ is the inferred result of the ith model, and Y_{observed_i} is the ith true value measured in the field. The unit of RMSE is m^{-1}, which is the same as the unit of a_{CDOM} (440) in this study.

III. Results

The modeling and validation results demonstrate that the statistical parameters (mean, median, std, min, and max) of a_{CDOM} (440) can be precisely inferred with minimal errors (MAPE<10%). The model's performance is influenced by various factors, such as the modeling datasets, the function forms used, the number of variables included in the functions, and the bands or band ratios selected for each variable. Further information and detailed results can be found in the following sections.

A. Results of Using Different Datasets

The results indicate that inference models developed from the clean dataset (referred to as clean models) outperformed those developed from the raw dataset (referred to as raw models). The clean models had a mean model R^2 of 0.084, a model RMSE of 0.033, a validation R^2 of 0.057, a validation RMSE of 0.038, and a validation MAPE of 12.012%. In contrast, the raw models showed poorer statistical indicators, with a mean model R^2 of 0.048, a model RMSE of 0.067, a validation R^2 of 0.046, a validation RMSE of 0.194, and a validation MAPE of 53.1%.

Fig. 4 presents a comparison between each pair of the same models, with the only difference being that they were modeled and validated using raw and clean datasets, respectively. Regardless of which statistical indicator was used to evaluate their performance, we observe that almost every clean model outperforms the raw model. The superiority of clean models is particularly evident in inferring the parameters mean, std, and max (as shown in Table II). For raw models, the MAPEs of the three parameters were 65.671%, 84.187%, and 62.111%, but these values were significantly reduced to 10.687%, 8.815%, and 9.911%, respectively, when using clean models. Furthermore, the accuracy of the median also improved from MAPE of 44.384% to 11.836%. However, the inference models for the minimum value did not show a significant difference; their R^2 values slightly improved, but RMSE and MAPE slightly worsened.

B. Results of Using Different Model Functions and the Number of Statistical Variables

Fig. 5 shows, among the 200 tests, the number of best models based on different functions (using linear, power, or exponential functions), using different numbers of bands or band ratios

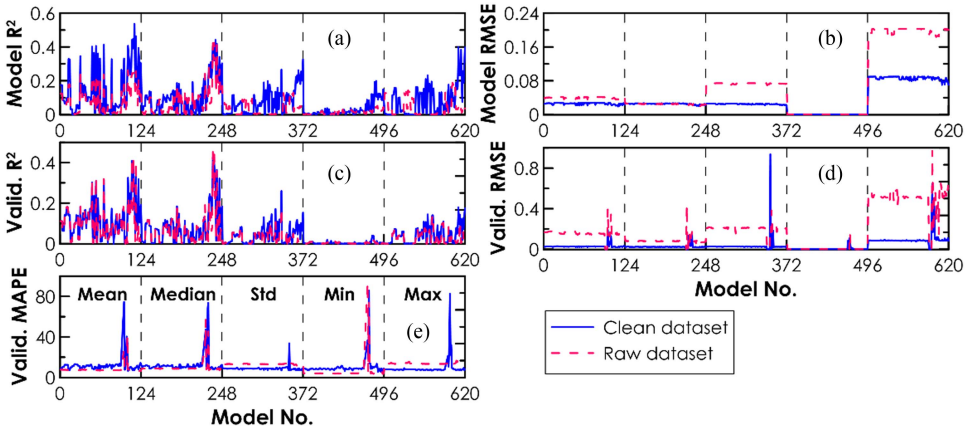

Fig. 4. Comparing model performance for different datasets (raw and clean), using five statistical indicators. (a) Model R^2. (b) Model RMSE. (c) Validation R^2. (d) Validation RMSE. (e) Validation MAPE.

TABLE II
COMPARISON OF MODEL PERFORMANCE FOR FIVE STATISTICAL PARAMETERS USING THE RAW AND CLEAN DATASETS, RESPECTIVELY

	Using raw dataset					Using clean dataset				
Stat. Para.	Model R^2	Model RMSE	Valid. R^2	Valid. RMSE	Valid. MAPE	Model R^2	Model RMSE	Valid. R^2	Valid. RMSE	Valid. MAPE
Mean	0.081	0.040	0.103	0.161	65.671	0.140	0.026	0.118	0.030	10.687
Median	0.080	0.025	0.069	0.084	44.384	0.103	0.025	0.079	0.026	11.836
Std	0.031	0.073	0.025	0.210	84.187	0.071	0.025	0.042	0.034	8.815
Min	0.014	0.0002	0.003	0.002	10.130	0.030	0.0004	0.004	0.003	18.904
Max	0.034	0.200	0.028	0.523	62.111	0.077	0.086	0.042	0.099	9.911

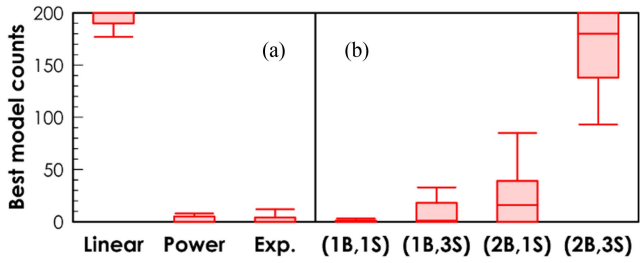

Fig. 5. Among 200 modeling and validation tests, the counts of the best models with different inference functions, see sub-part (a), the number of bands or band ratios (1B: one band or band ratio, 2B: two bands or band ratios), or the number of statistical variables (1S: one statistical variable, 3S: three statistical variables), see sub-part (b).

(using 1B, one band/band ratio or 2B, two bands/band ratios), or using different numbers of statistical variables (using 1S, one variable, or 3S, three variables), regardless of which statistical parameters were inferred and which statistical indicators were used to evaluate the model performance. The results show that linear functions outperformed power and exponential functions, with about 190 of the 200 best models using linear functions. Only in a few cases (∼10 out of 200) did power or exponential functions outperform linear functions [see Fig. 5(a)].

In addition, Fig. 5(b) shows that the use of more bands/band ratios and statistical variables improved model performance. Approximately 165 out of the 200 best models used 2B and 3S, i.e., model (8). The average occurrences of (1B, 1S), (1B, 3S), and (2B, 1S) in 200 tests were only 1, 9, and 25, respectively, indicating that using only one band/band ratio and one variable was not a good modeling method. Furthermore, the fact that (2B, 1S) performed better than (1B, 3S) suggests that using more spectral information was better than using more statistical information.

C. Results of Using Different Bands and Band Ratios

Fig. 6 shows the results of the best bands or band ratios used, without taking into account the model functions, the number of variables, the datasets used, and the statistical parameters inferred. For example, Fig. 6(a) shows that of the 200 inference modeling and validation runs, the best models for inferring mean CDOM were those that used (B1/B2, B2/B3) and were counted 80 times for the highest modeling R^2 and lowest modeling RMSE, 57 times for the highest validation R^2, 52 times for the lowest validation RMSE, and 50 times for the lowest validation MAPE. The second best models, which used bands (B1, B2) to infer the mean CDOM, did not perform as well in modeling but outperformed them in validation, with 57, 52, and 50 times being the best models for the best validation R^2, RMSE, and MAPE, respectively.

For inferring the median CDOM, the best bands were also (B1/B2, B1/B3), followed by (B1/B2, B2/B3), as these two models were significantly better than the others. To infer the CDOM std, the best bands were (B2/B4, B3/B4), which were much better and were counted as the best models more than 100 times compared to other bands, whereas (B1, B2), (B1/B3,

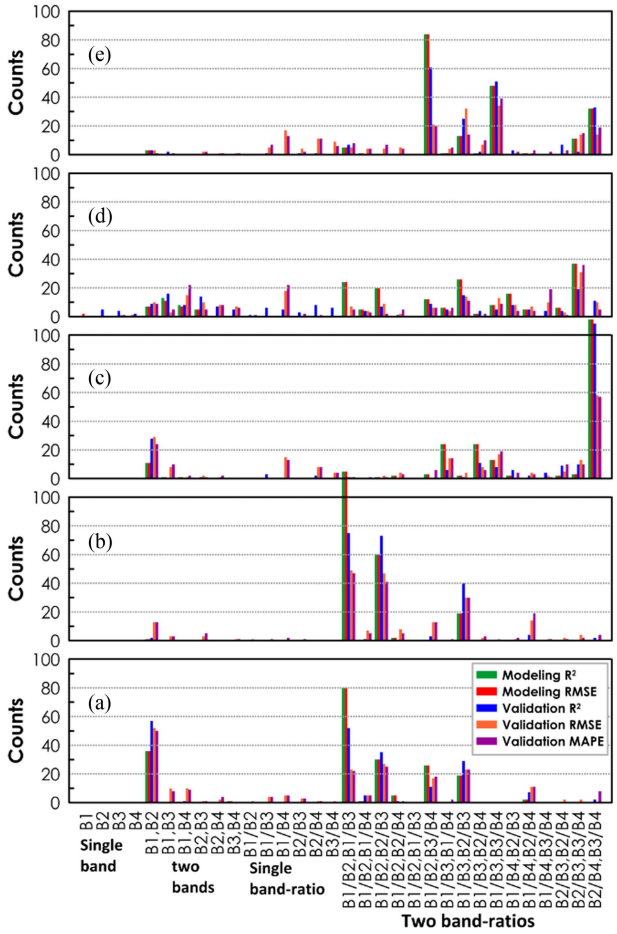

Fig. 6. Seven steps (Steps 1– 7 listed in Section II-B, but only using the clean dataset) of inference modeling and validation were repeated 200 times, and at each time, the bands, band ratios, or their combinations used by the best models (with the highest R^2, or the lowest RMSE and MAPE) were recorded, and their numbers were counted at last. For example, in (a), the counts of the models using two bands B1 and B2 and evaluated by modeling R^2 is 38, meaning that among the 200 times tests, there are 38 times that B1 and B2 played as the best band combination which gave the highest modeling R^2. The subfigures show the results of inferring different statistical parameters. (a) Mean. (b) Median. (c) Std. (d) Min. (e) Max.

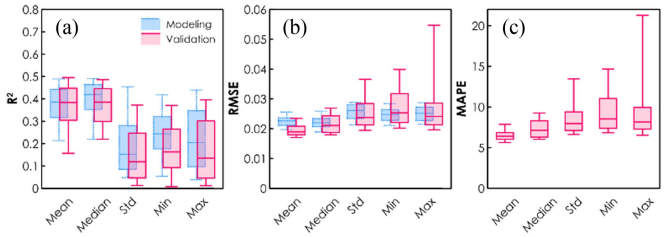

Fig. 7. Model performance for inferring different statistical parameters, shown by the indicators. (a) Modeling and validation R^2. (b) Modeling and validation RMSE. (c) Validation MAPE. The whisker lines indicate the percentiles of 10%, 25%, 50%, 75, and 90%.

B2/B4), and (B1/B3, B1/B4) are alternative options. However, when it comes to deriving the minimum CDOM, no band or band ratio stood out significantly from the others. Nevertheless, (B2/B3, B3/B4) (∼30 counts) was superior to (B1/B3, B2/B3), (B1/B2, B1/B3), and (B1, B4), which each achieved ∼20 counts as the best models. Regarding the inference of the maximum CDOM, two models using (B1/B2, B3/B4) and (B1/B3, B3/B4) were much better than the other models, with the former and the latter being the best models ∼80 and ∼50 times, respectively, out of 200 tests.

D. Results of Using Different Statistical Parameters

Fig. 7 illustrates the model performance for inferring five statistical parameters (mean, median, std, minimum, and maximum) using the same model forms, but with coefficients fitted and validated separately for each model. The results show that the mean and median are the parameters that can be inferred most accurately, with the highest R^2 (∼0.35) and the lowest RMSE (∼0.021) and MAPE (∼7%), as shown in Table III. In contrast, the models for the other three parameters (std, minimum, and maximum) performed relatively worse in both modeling and validation. Among them, the std was the most difficult parameter to model, with an average R^2 of only 0.19 for modeling and 0.16 for validation. The minimum and maximum parameters were inferred with similar accuracy, with MAPEs around 10%, although the minimum parameter had a slightly better MAPE of 9.70% compared to the maximum parameter's MAPE of 11.07%.

Furthermore, R^2 and RMSE results in Fig. 7(a) and (b) and Table III indicate that all models performed better in modeling than in validation. This is reasonable given that the models were developed using the modeling datasets and validated using independent validation datasets.

E. Best Models for Inferring CDOM in Qiandao Lake

Based on the results and analysis presented above, we recommend the models listed in Table IV as the best options for inferring each statistical parameter of the CDOM in Qiandao Lake. All of these models use two band ratios, three statistical variables (as shown in Table I), and the polynomial function of (8).

We also propose alternative models, listed in Table V, which are simpler in form and use fewer bands (one band ratio), but have lower accuracy than the best models. These simplified models can be used when some of the bands used in the best models are not available in the images.

Note that the models listed in Tables IV and V are based on data that has been magnified by a factor of 10. Therefore, if they are used to infer CDOM statistics, Rrs must be multiplied by 10 and the model result should be divided by 10 to obtain the CDOM statistical parameters in their true magnitude.

IV. Discussion

This section discusses several important issues related to remote sensing statistical inference and highlights some potential topics for future work.

TABLE III
MODEL PERFORMANCE OF FIVE STATISTICAL PARAMETERS

Statistical Parameter	Modeling		Validation		
	R^2	RMSE	R^2	RMSE	MAPE
Mean	0.37, 0.73	0.023, 0.013	0.35, 0.56	0.020, 0.016	6.55%, 5.09%
Median	0.39, 0.74	0.022, 0.014	0.36, 0.60	0.022, 0.016	7.47%, 5.09%
Std	0.19, 0.66	0.025, 0.016	0.16, 0.53	0.027, 0.017	9.62%, 5.62%
Min	0.25, 0.59	0.025, 0.018	0.18, 0.48	0.028, 0.016	9.70%, 5.09%
Max	0.22, 0.53	0.025, 0.018	0.17, 0.54	0.031, 0.018	11.03%, 5.89%

* In each cell, the first number is the average value and the second number is the best value, i.e., the highest R^2, and lowest RMSE and MAPE.

TABLE IV
BEST MODELS FOR INFERRING CDOM'S STATISTICAL PARAMETERS (Y) IN QIANDAO LAKE

Y	λ_1	λ_2	c_1	c_2	c_3	c_4	c_5	c_6	c_7	MAPE
Mean	B1/B2	B1/B3	12.925	-9.768	-3.523	-2.129	6.471	1.639	-1.927	5.09%
Median	B1/B2	B1/B3	9.554	-5.596	3.413	0.612	9.416	-1.952	-4.292	5.64%
Std	B2/B4	B3/B4	-0.060	-0.007	0.080	0.003	-0.654	2.511	1.806	6.75%
Min	B2/B3	B3/B4	-2.188	-0.064	0.567	0.049	3.237	1.658	3.075	6.96%
Max	B1/B2	B3/B4	7.407	0.028	-1.753	-0.102	-1.942	2.084	-1.665	6.15%

All models use the function (8).

TABLE V
ALTERNATIVE SIMPLE MODELS FOR INFERRING CDOM'S STATISTICAL PARAMETERS (Y) IN QIANDAO LAKE

Y	λ	*Function*	c_1	c_2	c_3	c_4	MAPE
Mean	B1/B2	1S, exponential (3)	0.325	0.229			6.23%
Median	B1/B2	1S, linear (1)	16.527	-11.993			7.37%
Std	B2/B4	1S, linear (1)	-1.361	3.471			7.42%
Min	B1/B4	3S, linear (7)	-0.003	-0.032	0.008	0.0029	7.89%
Max	B3/B4	1S, exponential (3)	4.573	0.140			8.03%

A. *Field Sampling and Resampling of the Field-Measured Data*

As our results show, to construct an accurate inference model, it is crucial to ensure that the modeling data have a good statistical distribution, rather than using all the raw data. Therefore, we recommend that field sampling sites are distributed as evenly as possible across the study area to ensure that the samples collected represent the true statistical distribution of the population. However, it can be challenging to achieve evenly distributed sampling for aquatic environmental studies, particularly for a lake with a large area and frequently changing water quality. In this study, the sampling sites were essentially evenly distributed across Qiandao Lake but were sampled at different times (either spring or autumn), which means that the statistical distribution of CDOM in the raw samples may not exactly match the ground truth.

We strongly recommend that a data screening process is performed before inference modeling to remove redundant samples and use the remaining samples to form a smooth probability distribution curve. For example, if the raw histogram shows a normal distribution, we can remove some samples to obtain a clear normal distribution, ensuring that the remaining samples are still evenly distributed in the space of the study area, as shown in Fig. 2.

Alternatively, we can add some samples to the dataset to obtain a statistical distribution with more samples. For example, we can add seven or eight samples for aCDOM (440) between 0.4 and 0.6 m^{-1} (the third bin in Fig. 3) and remove seven or eight samples for a_{CDOM} (440) between 0.6 and 0.8 m^{-1} (the fourth bin in Fig. 3). This allows us to form a newly generated exponential distribution with 100 samples, which is more than the 67 samples in the clean dataset used in this study. The values of the new samples added to the dataset can be estimated using random interpolations of the neighboring samples.

B. *Statistical Variables Used in Inference Models*

In our previous studies, we assumed that there were relationships between the same statistical parameters of the optical and nonoptical properties of a lake. For instance, we used the mean spectra to infer the mean CDOM, and the std of spectra to infer the std of CDOM. However, these relationships have not been theoretically proven and may not necessarily be true. We need to pay particular attention to the parameters min and max. The minimum spectrum may have a relationship with the maximum CDOM, rather than the minimum because CDOM has strong absorption but no backscattering. Hence, if a_{CDOM} (440) is high, more light will be absorbed, and the water's reflectance will be low. The statistical variables used in Table I were only tentative

tests, and further studies on this topic are required for future work.

C. Inference of Statistical Probability Distribution

In this study, we focused solely on deriving the five statistical parameters of the CDOM: mean, median, std, minimum, and maximum. These parameters provide crucial insight into the main features of the statistical distribution of CDOM and are sometimes sufficient to describe its complete statistical probability distribution, assuming a prior knowledge of this distribution. For example, if we know that CDOM in a lake follows a normal distribution, then inferring the mean and std would allow us to fully understand the statistical probability distribution of CDOM in the lake. However, in cases where we lack prior knowledge, these key parameters may not be sufficient to determine the full statistical probability distribution. In such cases, we need additional parameters to approximate the full distribution. In this study, we inferred the median, minimum, and maximum values, which were set as the 50%, 5%, and 95% percentiles of the CDOM distribution histogram, respectively. By inferring more percentiles, such as 5%, 10%, 15%, and so on, up to 95%, we can bring the histogram closer and closer to the full statistical probability distribution of CDOM.

D. Types of Inference Models

The bootstrap-based method presented in this study is essentially empirical, regardless of the number of bands/band ratios and statistical variables employed. While using more bands, band ratios, and variables is likely to yield more accurate results, the resulting models become increasingly complex and overfitting. Consequently, their accuracies may not improve significantly. We recommend the use of the ternary and cubic polynomial function (8) for CDOM inference in complex waters. However, the function coefficients listed in Tables IV and V may not be applicable for CDOM inference in other water bodies. In addition to the bootstrap-based method, other empirical, analytical, or semianalytical inference methods remain to be explored in future research.

The empirical inference models of Qiandao Lake cannot be directly applied to other lakes. We do not expect there is a universal inference model which can be used in all closed-connected water bodies—it is like there is no universal inversion model which works well for all waters. However, the methods, functions, bands or band ratios, and statistical variables used in Qiandao Lake can be adopted by other lakes, and use their own *in situ* data to fit the lake-specific model coefficients.

V. Conclusion

We estimated the five statistical features (mean, median, std, min, and max) of CDOM concentration a_{CDOM} (440) in Qiandao Lake using their respective remote sensing statistical inference models. These models were built based on field-measured data and bootstrap methods, and they generally performed well with MAPEs of 6%–11%. We recommend reselecting the clean data from the raw field-measured data to obtain CDOM samples with known statistical distributions, such as normal, log-normal, or exponential distributions, before building inference models. Using three spectral statistical variables at two band ratios, along with polynomial functions, is the best approach for constructing accurate inference models. Among the five statistical parameters, mean and median can be inferred more accurately than std, min, and max.

Acknowledgment

The author would like to thank Professors Qian Cheng and Shuangyan He and graduate students Nan Sun, Litong Huang, Shuna Pang, and Zeliang Zhang for their contributions to the measurements, data, and funding acquisition.

References

[1] W. N. Zhu, Z. L. Zhang, Z. Q. Yang, S. N. Pang, J. Chen, and Q. Cheng, "Spectral probability distribution of closed connected water and remote sensing statistical inference of yellow substance," *Photogramm. Eng. Remote Sens.*, vol. 87, pp. 807–819, 2021.

[2] W. N. Zhu, "Remote sensing statistical inference: Basic theory and forward simulation of water–air statistical radiative transfer," *Earth Sci. Inform.*, vol. 14, pp. 2145–2159, 2021.

[3] W. Zhu, Z. Yang, S. He, and Q. Cheng, "Skewness-based classification and environmental indication of spectral probability distribution of global closed connected waters," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 1504005.

[4] W. N. Zhu and W. Xia, "Effects of atmospheric correction on remote sensing statistical inference in an aquatic environment," *Remote Sens.*, vol. 15, 2023, Art. no. 1907.

[5] J. W. Goodman, *Statistical Optics*, 2nd ed. Hoboken, NJ, USA: Wiley, 2015.

[6] V. P. Kandidov, "Monte Carlo method in nonlinear statistical optics," *Uspekhi Fizicheskikh Nauk*, vol. 166, pp. 1309–1338, 1996.

[7] A. C. Fassoni-Andrade and R. C. D. de Paiva, "Mapping spatial-temporal sediment dynamics of river-floodplains in the Amazon," *Remote Sens. Environ.*, vol. 221, pp. 94–107, 2019.

[8] A. P. Yunus, Y. Masago, and Y. Hijioka, "COVID-19 and surface water quality: Improved lake water quality during the lockdown," *Sci. Total Environ.*, vol. 731, 2020, Art. no. 139012.

[9] J. Gazeaux, E. Flaounas, P. Naveau, and A. Hannart, "Inferring change points and nonlinear trends in multivariate time series: Application to West African monsoon onset timings estimation," *J. Geophys. Res., Atmos.*, vol. 116, 2011, Art. no. D05101.

[10] S. D. Prager and J. J. Barber, "Modeling unobserved true position using multiple sources and information semantics," *Int. J. Geograph. Inf. Sci.*, vol. 26, pp. 15–37, 2012.

[11] P. Meyfroidt, T. K. Rudel, and E. F. Lambin, "Forest transitions, trade, and the global displacement of land use," *Proc. Nat. Acad. Sci. USA*, vol. 107, pp. 20917–20922, 2010.

[12] D. H. Card, "Using known map category marginal frequencies to improve estimates of thematic accuracy," *Photogramm. Eng. Remote Sens.*, vol. 48, pp. 431–439, 1982.

[13] R. L. Czaplewski and G. P. Catts, "Calibration of remotely sensed proportion or area estimates for misclassification error," *Remote Sens. Environ.*, vol. 39, pp. 29–43, 1992.

[14] R. E. McRoberts and B. F. Walters, "Statistical inference for remote sensing-based estimates of net deforestation," *Remote Sens. Environ.*, vol. 124, pp. 394–401, 2012.

[15] W. Xia, "Comparison of remote sensing statistical inference and remote sensing inversion in calculating CDOM statistical characteristics," M.S. thesis, Zhejiang Univ., 2023.

[16] W. N. Zhu, Q. Yu, Y. Q. Tian, R. F. Chen, and G. B. Gardner, "Estimation of chromophoric dissolved organic matter in the Mississippi and Atchafalaya River plume regions using above-surface hyperspectral remote sensing," *J. Geophys. Res., Oceans*, vol. 116, 2011, Art. no. C02011.

[17] T. Kutser, "The possibility of using the Landsat image archive for monitoring long time trends in coloured dissolved organic matter concentration in lake waters," *Remote Sens. Environ.*, vol. 123, pp. 334–338, 2012.

[18] W. N. Zhu, Y. Q. Tian, Q. Yu, and B. L. Becker, "Using Hyperion imagery to monitor the spatial and temporal distribution of colored dissolved organic matter in estuarine and coastal regions," *Remote Sens. Environ.*, vol. 134, pp. 342–354, 2013.

[19] N. B. Nelson and D. A. Siegel, "Chromophoric DOM in the open ocean," in *Biogeochemistry of Marine Dissolved Organic Matter*, D. A. Hansell and C. A. Carlson, Eds. New York, NY, USA: Academic, 2002, pp. 547–578.

[20] L. Prieur and S. Sathyendranath, "An optical classification of coastal and oceanic waters based on the specific spectral absorption curves of phytoplankton pigments, dissolved organic matter, and other particulate materials," *Limnol. Oceanogr.*, vol. 26, pp. 671–689, 1981.

[21] Y. L. Zhang et al., "Chromophoric dissolved organic matter in inland waters: Present knowledge and future challenges," *Sci. Total Environ.*, vol. 759, 2021, Art. no. 143550.

[22] W. N. Zhu et al., "An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments," *Remote Sens. Environ.*, vol. 140, pp. 766–778, 2014.

[23] J. Chen, W. N. Zhu, Y. Q. Tian, and Q. Yu, "Monitoring dissolved organic carbon by combining Landsat-8 and Sentinel-2 satellites: Case study in Saginaw River estuary, Lake Huron," *Sci. Total Environ.*, vol. 718, 2020, Art. no. 137374.

[24] J. H. Mueller, G. S. Fargion, and C. R. McClain, "Biogeochemical and bio-optical measurements and data analysis methods," National Aeronautical and Space Administration, Goddard Space Flight Space Center, Greenbelt, MD, USA, NASA Tech. Memo., NASA TM-2003-211621, rev. 4, vol. 2, 2003.

[25] Z. L. Zhang, W. N. Zhu, J. Chen, and Q. Cheng, "Remotely observed variations of reservoir low concentration chromophoric dissolved organic matter and its response to upstream hydrological and meteorological conditions using Sentinel-2 imagery and Gradient Boosting Regression Tree," *Water Supply*, vol. 21, pp. 668–682, 2021.

[26] 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.

Weining Zhu received the B.S. and M.S. degrees in computer science and geography from Nanjing University, Nanjing, China, in 1995 and 2004, respectively, and the Ph.D. degree in geosciences from the University of Massachusetts, Amherst, MA, USA, in 2011.

From 2012 to 2013, he was an Assistant Research Professor with the Institute of Great Lakes Research, Department of Geography, Central Michigan University, Mount Pleasant, MI, USA. Since 2013, he has been a Faculty Member with the Department of Ocean Informatics, Ocean College, Zhejiang University, and Donghai Laboratory, Zhoushan, China. His current research interests include the remote sensing of aquatic environments and geographical information science.