Spectral possibility distribution of closed connected water and remote sensing statistical inference for lacustrine yellow substance

Weining Zhu[1,2*], Zeliang Zhang[3], Zaiqiao Yang[2], Shuna Pang[3], Jiang Chen[4], Qian Cheng[5]

1. Key Laboratory of Ocean Observation-Imaging Testbed of Zhejiang Province, Ocean College, Zhejiang University, ZJ, China
2. Department of Marine Information Science, Ocean College, Zhejiang University, ZJ, China
3. Department of Marine Science, Ocean College, Zhejiang University, ZJ, China
4. School of Remote Sensing and Information Engineering, Wuhan University, HB, China.
5. School of Tourism and Urban-Rural Planning, Zhejiang Gongshang University, ZJ, China

Abstract

The traditional ocean color remote sensing usually focuses on using optical inversion models to estimate the properties of in-water components from the above-surface spectra, so we call it the spectrum-concentration (SC) scheme. Unlike the SC scheme, this study proposed a new research scheme, distribution-distribution (DD) scheme, which uses statistical inference models to estimate the possibility distribution of these in-water components, based on the possibility distribution of the observed spectra. The DD scheme has the advantages that (1) it can rapidly give the key and overview information of the interest water, instead of using the SC scheme to compute each image pixel, (2) it can assist the SC scheme to improve their models and parameters, and (3) it can provide more valuable information for better understanding and indicating the features and dynamics of aquatic environment. In this study, based on Landsat-8 images, we analyzed the spectral possibility distributions (SPD) of 688 global water and found many of them were normal, lognormal, and exponential distributions, but with diverse patterns in distribution parameters such as the mean, standard deviation, skewness and kurtosis. Furthermore, we used Monte-Carlo and Hydrolight simulations to study the theoretical and statistical connections between the possibility distributions of in-water components and SPDs. The simulation results were basically consistent with the observations on the real water. Then by using the simulation and field measured data, we proposed a bootstrap-based DD scheme and developed some simple statistical inference models to estimate the distribution parameters of yellow substance in lakes. Since DD scheme is still on its early stage, we also suggested some potential and useful topics for the future work.

Keywords

Spectral possibility distribution (SPD), remote sensing, statistical inference, yellow substance, aquatic environment

1. Introduction

An important content of aquatic remote sensing is estimating the properties of in-water components from the above-surface spectra (Bukata et al., 1995; O'Reilly et al., 1998). See an example in Fig. 1a, the absorption coefficient of yellow substance (also called CDOM, colored dissolved organic matter) at 440 nm, i.e., $a_{CDOM}(440)$, was estimated from a spectrum at an image pixel of the interested lake (Zhu et al., 2014; Chen et al., 2017), by using a remote sensing inversion model – we call this approach the spectrum-concentration (SC) scheme of ocean color remote sensing. In this study, we will propose a new research scheme, distribution-distribution (DD) scheme, which uses statistical inference to estimate $a_{CDOM}(440)$'s distribution of the whole lake, based on the spectral possibility distributions (SPD) of all water pixels over the lake at different wavelengths, see Fig. 1b. Note that we do not call it spectral frequency distribution, because the term 'frequency' is likely to be confused with the reciprocal of the 'wavelength', a well-known term used in remote sensing.

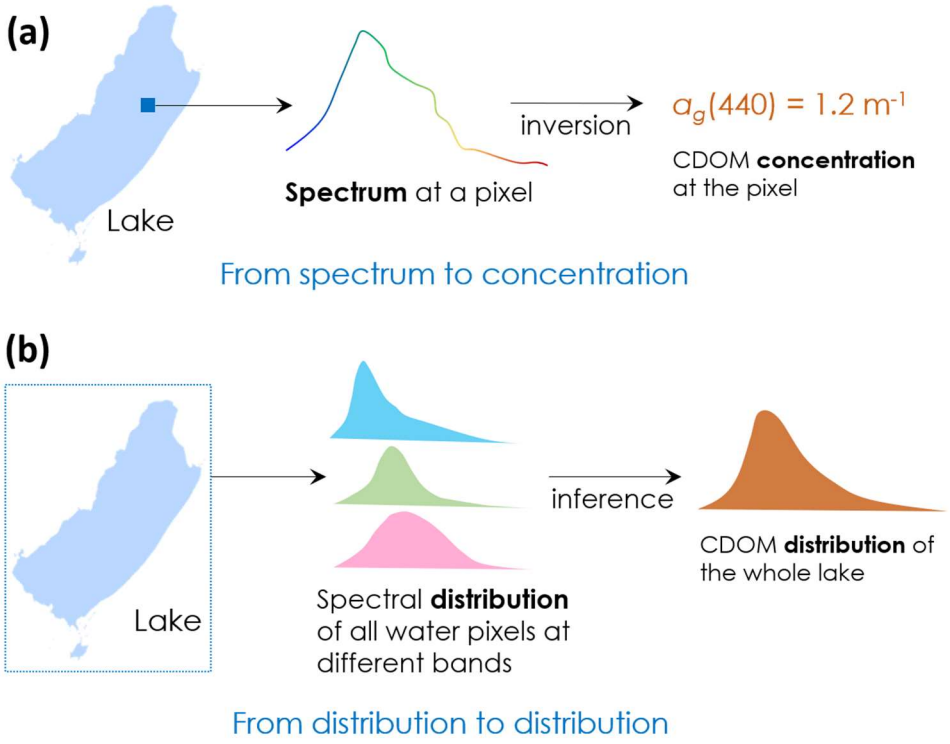

Figure 1. The new scheme for ocean color remote sensing: **(a)** SC scheme, the conventional spectrum-concentration scheme based on optical inversion, and **(b)** DD scheme, the possibility distribution-distribution scheme based on statistical inference.

The possibility distribution method (e.g., histogram) has been widely used in many disciplines. In image processing, for example, it helps image segmentation to extract the interested objects from an image (Ohlander, et al., 1978; Gonzalez & Wood, 2017). In remote sensing, it also has some applications in image classification (Battiti, et al., 2015; Demir & Beguem, 2016), image enhancement (Demirel, et al., 2010; Fu, et al., 2015), image registration (Paul & Pati, 2016), haze removal (Huang, et al., 2019), land cover change detection (Lv, et al., 2019), and etc. However, so far it has not been widely associated with the radiative transfer and applied in aquatic remote sensing – some studies only reported the distribution characteristics of the measured water samples on their physical, geographical, biological, chemical, or inherent optical properties (IOPs), and many of these studies were more on the distributions of the collected samples instead of the whole population, or their distributions were statistically counted from the results of the SC scheme (Schwarz, et al., 2005; McKee, et al., 2006; Foden, et al., 2008; Sabetta, et al., 2008; Dong, et al., 2010; Zhu, et al., 2013b; Alcântara, et al., 2016; Keller, et al., 2018; Zhao, et al., 2018; Chen, et al., 2019; Kuhn, et al., 2019). In remote sensing, possibility distributions of in-water components have not been directly connected to the possibility distributions of water's apparent optical properties (AOPs), such as the surface reflectance. Most of relevant studies still follow the SC scheme, which directly connects the IOP and AOP themselves, rather than their possibility distributions in a water area.

Like the SC scheme, the DD scheme also has a theoretical foundation of radiative transfer but with more statistical issues involved. In SC scheme, the radiative transfer process from IOPs to AOPs determines the spectrum at a given pixel (Mobley 1994; Kirk 1994), while in DD scheme, the same process also changes the possibility distributions from IOPs to AOPs within the whole underwater and above-surface light fields. If the SC scheme is based on the classic geometric optics, then the DD scheme could be more based on the statistical optics (Goodman, 2015).

One of advantages of using DD scheme is that sometimes we do not need exactly to know water properties of every pixel. Instead, quite often the water resource administrators or environment monitors only want to roughly overview the water quality – that is the DD scheme can do. In practice, it is almost impossible to know the real distribution of an ocean color component, unless we had a very high-resolution field measurement over the entire study water. One approach of knowing the real distribution is making a statistical inference from the collected samples – that is what the classic statistical inference does, but this approach may have high cost on time and labor. Another approach is making a remote sensing inversion model using the collected samples, and then applying the model to satellite images – that is what the traditional ocean color remote sensing does, but this approach may be time-consuming and inaccurate if with large-size images and/or poor inversion models. The DD scheme proposed in this study combines the above two approaches and try to directly infer the statistical distributions of ocean color components, without first making and later applying an inversion model. Extracting the SPDs from an image and executing a statistical inference model once would be much faster than repeating an inversion model millions of times for a large number of image pixels.

DD scheme can also assist SC scheme in making inversion models. To make an inversion model used in SC scheme, we usually need to collect in-situ samples, and then establish, calibrate and validate a model based on these samples, and finally applied the model to satellite images. From the perspective of statistics, the collected water samples were only the statistical samples, while the water pixels in an image is the statistical population. Without knowing the population, the sample-

based models are likely to make large errors for those pixels out of the sampled water. DD scheme can provide the information of the population, so that we can adjust inversion models correspondingly. For example, if we made a simple exponential inversion model such as $x = ay^b+c$, where x is chlorophyll concentration, y is a band-ratio, and *a, b, and c* are model coefficients and determined by fitting the field measured samples. This model may work well for the field sampled data, in which y was in the range of, say, 0.2-0.3. If this model was applied to the whole image, then in case the image-derived y might be in the range of 0.01-3, the estimated x might be enlarged to the unreasonable number, meaning that the coefficients a, b, c, and even the model itself are not accurate or at least with the limited efficiency. So, if we have the DD scheme to tell us the possible range of x and y, then we can test and adjust the inversion model to make sure that its input and output variables, x and y, will fall in appropriate ranges.

Another benefit of using DD scheme is that we can directly take SPDs as the optical properties or environmental indicators of a water area. It means that we even do not need to infer any statistical distribution for a specific ocean color component, but directly link SPDs to other natural phenomena and properties of the aquatic environment, for example, the water depth, salinity, watershed area, LULC (land uses and land cover), NDWI (normalized difference water index), and algal bloom that would have influences on water IOPs and consequently change the observed SPDs.

From the above introduction we can see that DD scheme has many advantages and application potentials for remote sensing of ocean color and aquatic environment. This study consists of three main sections: (1) extracting SPDs from hundreds of global waters, and describing and analyzing their statistical features, (2) based on radiative transfer theory as well as the Monte-Carlo/Hydrolight simulations, exploring statistical relationships and parameter transfer among the possibility distributions of in-water components, intermediate optical variables, and above-surface spectra, and (3) following DD scheme to propose a statistical method to infer the distribution of an color component. Based on our data, we particularly focused on the lacustrine yellow substance, i.e. CDOM, which is an important ocean color component, with high environmental significance, and has many indicator effects. More introduction of CDOM can be seen in reference (Zhu et al., 2013a, 2014; Chen et al., 2017).

2. Landsat-8's observation on SPDs of the global closed connected water

2.1 Study water

In theory, DD scheme can be used to any water in an image, provided it contains enough water pixels as study samples, but in practice, it is usually applicable for those waters with clear boundaries, such as lakes and reservoirs. We call such water the closed connected water. The term 'closed connected' is the similar one used in mathematics and geometrical topology, meaning an object or a polygon has a boundary and in which any two points are connected. If any open water has no clear boundary, e.g., Yellow Sea in China, then the image pixels used to determine its SPDs cannot be explicitly identified. On the other side, if any water area is not connected, then they are actually two areas rather than a single water area. Obviously, mixing two areas into one area may lead their SPDs to be distorted and meaningless. Lakes, reservoirs, and lagoons are typical closed connected water, while bays, estuaries, inlets, and a segment of river channel can be also taken as the closed connected water as long as they are approximately closed, and connected to the outside water with narrow outlets, making their water cannot be rapidly and massively exchanged.

In this study, we selected 688 waters, covering almost the entire Earth, see Fig. 2. Most of these waters are lakes and reservoirs, including some famous lakes, such as the Dead Sea, West Lake, Yellowstone Lake, and Lake Geneva. A few bays, lagoons, estuarine regions, and riverine segments were also studied, including the San Francisco Bay, Tokyo Bay, Hudson River Estuary, and Indian River Lagoon. Since the Landsat-8 images were used in this study, we only focused on those closed connected waters which can be entirely observed in the scene of a Landsat-8 image.

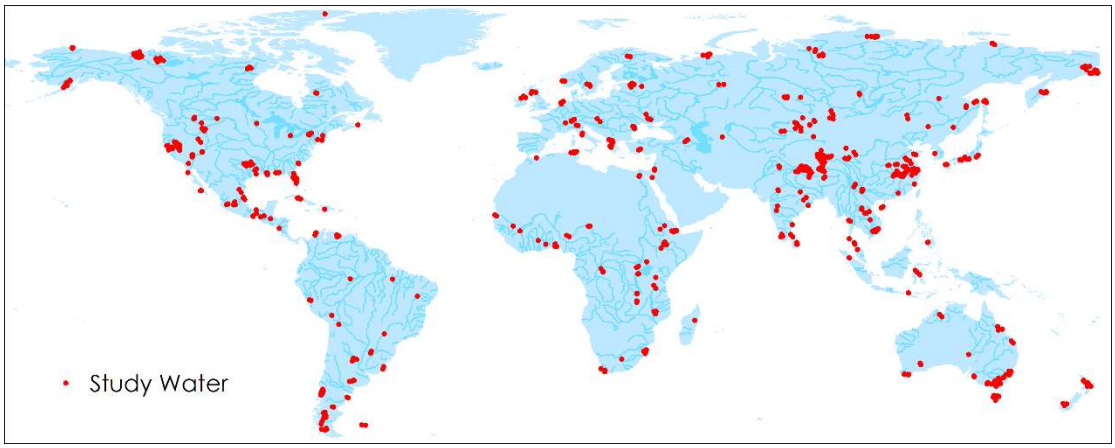

Figure 2. The study site map of the closed connected water in the world.

2.2 Data processing and statistics

Many types of spectral variables, in different levels, can be obtained from a satellite image, for example, the DN (digital number), radiance of TOA (top of atmosphere), atmospherically corrected surface reflectance (*Ref*), and remote sensing reflectance (R_{rs}), which is widely used in ocean color remote sensing. Therefore, here comes a problem that which variable should be used for SPD analysis? Certainly, R_{rs} is the best since it only carries the spectral information of the below-surface water itself, while the others may carry the useless non-water information contributed by air-water interface and the atmosphere. However, processing an image for R_{rs} may not be necessary if the DN-based SPD, for example, can still be used for DD scheme without significant errors. Like in image classification, we often do not need to make an atmospheric correction. We would like to remain this problem for the future work. In this study, R_{rs} was used for SPD analysis and it can be directly obtained from Landsat-8 images using the ACOLITE, an atmospheric correction model developed for aquatic applications of Landsat data (Vanhellemont & Ruddick, 2014, 2018).

The output R_{rs} of ACOLITE has already masked the non-water pixels and the negative values. The other pixels outside the study water should be further manually masked by setting the ROI (region of interest). Once the ROI of the study water has been determined, R_{rs} in the ROI at five bands (443, 483, 551, 665, and 865 nm of Landsat-8) were extracted and their SPDs can be represented by their respective histograms, see Fig. 3. The 2% minimum and 2% maximum R_{rs} were excluded from the SPD, because there were still possibly image processing errors and some unavoidable contaminated pixels induced by such as the shallow bottoms, whitecaps, and glints.

To consistently compare SPD characteristics of different waters, their R_{rs} histograms can be further normalized into the range 0-1, using the formula

$$R_{rs\_norm} = \frac{R_{rs} - R_{rs\_min}}{R_{rs\_max} - R_{rs\_min}} \quad (1)$$

where the R_{rs_norm} is the normalized R_{rs}, and R_{rs_min} and R_{rs_max} are the minimum (2%th) and maximum (98%th) values in the ROI, respectively.

Another issue of a SPD histogram is determining the number of bins. It is known that there is no the 'best' number of bins, since different bin number can reveal different data features. Many rules and formulas have been proposed for this issue, such as the square-root choice, Sturges' formula, and Freedman-Diaconis rule (Sturges, 1926; Scott, 1992). In this study, the number of bins was determined by the Sturges' formula

$$k = 1 + 3.32\log(N) \quad (2)$$

where k is the number of bins and N is the number of samples, namely, the number of image pixels within the study water. In the 688 study waters, the N ranges from tens of thousands to over four million, and hence k ranges from 15 to 23, so we simply set k = 20 for all water images. Moreover, to consistently compare the R_{rs}'s possibility in different water, the bin's height was also normalized to the possibility = n/N instead of the count n.

2.3 Statistic and analysis of the observed SPDs

By observing the resultant SPDs of the 688 closed connected water, we found that although their SPDs have shown the diversity in shapes and magnitudes, many of them were still share the same or similar features. Fig. 3 shows the selected typical SPDs of the study water, including freshwater/saltwater natural lakes, man-made reservoirs, estuaries, and swamps under different geographical and environmental conditions, see details in Table 1. In this preliminary study, we only focused on their statistical distribution properties that can be characterized by three aspects:

(1) What are their likely pdfs (possibility distribution functions)?

Many SPDs are highly and approximately similar to the known common distributions such as the normal, lognormal, exponential, or uniform. For example, the SPDs of Dead Sea and Garza-Little Elm Reservoir (Fig. 3a) shown the smoothed normal or lognormal distributions, SPD at 865 nm in Yishan Lake (Fig. 3c) and 655 nm in Lake Rotorua (Fig. 3q) shown the likely exponential distributions, and some bands in Lake Novosibirsk (Fig. 3j) and Lake Tahoe (Fig. 3t) shown the approximately uniform distributions. However, many SPDs were still with uncommon distributions such as the bimodal shown in Lake Kakhovske (Fig. 3d), small narrow peaks in Lake Cuitzeo (Fig. 3p), small flat peaks in Lake Duoergaicuo (Fig. 3n), small zigzags in Sandy Lake at 483 nm (Fig. 3s), and the irregular cases in Hudson River estuarine region (Fig. 3u). In addition, some common distributions were not so ideal in their statistical parameters – some were not smoothed, some were shifted, and some were skewed. In particular, the shifting and skewing from the standard distributions were often occurred in many water, making their SPDs looked symmetric with each other. For example, the exponential distributions in Yishan Lake (right shifted and left skewed, Fig. 3c) vs. Bahi Swamp (left shifted and right skewed, Fig. 3e), and the normal/lognormal distributions in Lake Eucumbene (right shifted and left skewed, Fig. 3h) vs. in Lac Maunoir (left shifted and right skewed, Fig. 3i).

Note that here we only gave a descriptive observation on their SPDs, more details of skewness and other distribution parameters were analyzed in next subsections of this study, while the even

more insight exploration on SPD's real pdfs remains for the future work, by using the more accurate statistical methods such as the parametric tests.

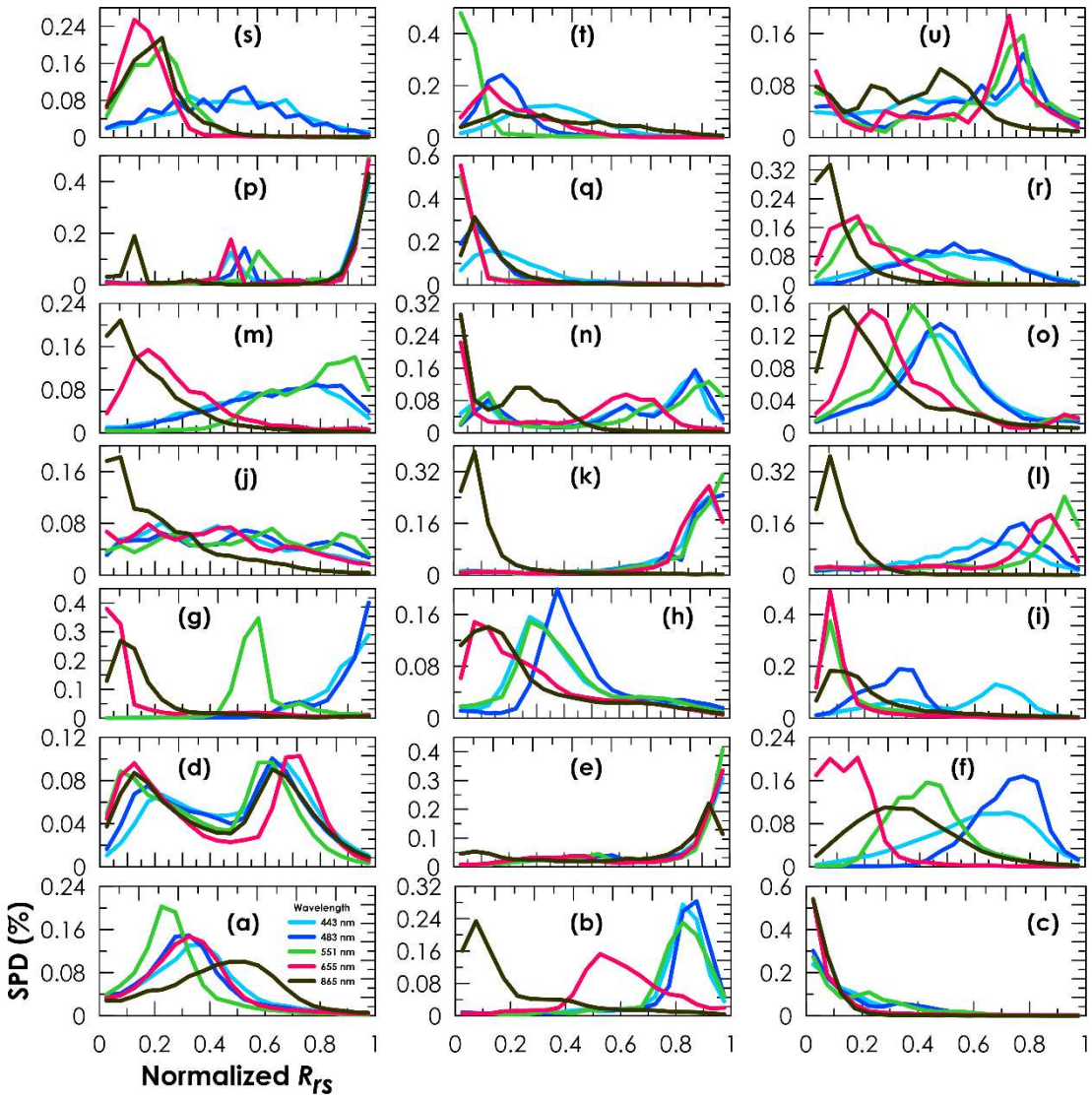

Figure 3. The Landsat-8 observed typical SPDs of 21 closed connected water: **(a)** Dead Sea, **(b)** Utah Lake, **(c)** Yishan Lake, **(d)** Kakhovka, **(e)** Bahi Swamp, **(f)** Cardiel Lake, **(g)** Lake Argentino, **(h)** Lake Eucumbene, **(i)** Lac Maunoir, **(j)** Novosibirsk, **(k)** Lake Ross, **(l)** Lake Griffin, **(m)** Pongolapoort, **(n)** Duoergaicuo, **(o)** Garza-Little Elm, **(p)** Lake Cuitzeo, **(q)** Lake Rotorua, **(r)** Lake Tarawera, **(s)** Sandy Lake, **(t)** Lake Tahoe, and **(u)** Hudson River Estuary. The detailed information of these lakes is shown in Table 1.

(2) What are the statistical features (mean, variance, skewness, etc.) of their SPDs?

Without knowing the exact pdf form of a given SPD, we may, nevertheless, calculate some well-known key statistical parameters, such as the mean, standard deviation (SD), and moments (skewness, kurtosis, etc.), to illustrate a general overview of samples or populations. The Fig. 4 shows the distributions of the mean, SD, skewness, and kurtosis of the 688 study water at 5 Landsat-8 bands. We can see that they have different trends: from 443 nm to 865 nm, the mean decreased (Fig. 4c), skewness and kurtosis increased (Fig. 4a and 4b), while SD kept almost unchanged (Fig.

4d). Mean values < 0.5 (Fig. 4c) indicates that more SPDs were shifted right with positive (left) skewness (mean > 0.2, Fig. 4a). The mean of 688 mean values at 865 nm is 0.24, indicating they were highly shifted and concentrated in the half (left) part of the SPDs, (see Fig. 3, all SPDs at 865 nm were left skewed except Fig. 3e and 3p) and consequently making they were with the larger kurtosis (mean 4.8, Fig. 4b). Unlike the other three parameters, SD were always around the 0.2 for all 5 bands.

Table 1. The geographical information and the observed typical SPD features (shown in Fig. 3) of the selected waters

Fig.	Water	Water/Geo.	Geo. location		Region,	Obs. Date	SPD features		
No.	Name	Feature	Lat.	Lon.	Nation	Y/M/D	pdf	SK	BR type
(a)	Dead Sea	s.w. lake, endorheic	31.54N	35.48E	Jordan	19/7/12	NL	RM	U-B-RGN
(b)	Utah Lake	f.w., slightly saline	40.22N	111.82W	UT, USA	17/5/3	NL’	RML	U-B-GRN
(c)	Yishan Lake	f.w., no-man’s-land	35.24N	90.92E	QH, China	14/10/22	E	R	U-B-G-R-N
(d)	Kakhovka	f.w., reservoir	47.49N	34.25E	Ukraine	18/5/27	Irr	RL	U-B-G-R-N
(e)	Bahi Swamp	f.w., wetland	6.09S	35.18E	Tanzania	16/6/23	E’L’	L	U-B-G-R-N
(f)	Cardiel Lake	f.w., partly alkaline	48.92S	71.20W	Argentina	15/9/20	NLL’	RML	UBGRN
(g)	Lake Argentino	f.w., subalpine	50.22S	72.45W	Argentina	18/3/27	NLEE’	RML	U-BGR-N
(h)	Lake Eucumbene	f.w. reservoir, alpine	36.05S	148.71E	Australia	19/2/21	NL	ML	U-GBR-N
(i)	Lac Maunoir	f.w. lake, arctic	67.49N	124.98W	Canada	15/6/29	LL’	RML	UBG-RN
(j)	Novosibirsk	f.w. reservoir	54.74N	82.85E	Russia	17/6/14	EU	R	U-B-G-RN
(k)	Lake Ross	f.w., water supply	19.44N	146.76E	Australia	17/5/28	LL’E’	RL	U-B-G-RN
(l)	Lake Griffin	f.w., chain lakes	28.88N	81.84W	FL, USA	18/3/9	LL’	RL	UBGRN
(m)	Pongolapoort	f.w. reservoir, dam	27.41S	31.96E	South Africa	17/6/28	LL’	RL	U-BGRN
(n)	Duoergaicuo	f.w. lake, plateau	35.21N	92.17E	Tibet, China	17/10/7	Irr	RML	U-B-GRN
(o)	Garza-Little Elm	f.w. reservoir, rural	33.10N	96.97W	TX, USA	18/3/22	NL	RM	U-BGRN
(p)	Lake Cuitzeo	f.w. lake, astatic	19.94N	101.21W	Mexico	19/4/10	Irr	ML	U-B-G-RN
(q)	Lake Rotorua	f.w. lake, volcano	38.07S	176.27E	New Zealand	14/3/6	LE	R	UB-NG-R
(r)	Lake Tarawera	f.w. lake, volcano	38.20S	176.43E	New Zealand	14/3/6	NL	RM	U-BG-RN
(s)	Sandy Lake	f.w. lake, natural	67.81N	132.25W	Canada	14/9/8	NL	RM	U-BG-R-N
(t)	Lake Tahoe	f.w. lake, resort	39.10N	120.05W	CA, USA	17/9/27	NLEU	RM	UBGRN
(u)	Hudson River	f.w. estuary, urban	41.04N	73.89W	NY, USA	16/3/30	Irr	ML	UBGRN

Notations for symbols and abbreviations: s.w. (saltwater), f.w. (freshwater); in pdf column, N (normal), L (lognormal), L’ (negative lognormal), E (exponential), E’ (negative exponential), U (uniform), Irr (irregular); in SK (skewness) column: L (left), M (middle), R (right); in BR (band relationship) type column: U (443 nm), B (483 nm), G (551 nm), R (655 nm), N (865 nm).

Fig. 5 shows the correlations between two statistical parameters at 551 nm, and the correlations at other wavelengths were similar with those at 551 nm. We can see the similar relations above-mentioned and observed from Fig. 4. There is no significant correlations (R = 0.29) between mean and SD (Fig. 5a), but there are good correlations between skewness and mean/kurtosis (Fig. 5b and 5d). Their strong correlations make sense because the closer the mean is to 0.5, the closer the skewness is to 0 as well as the lower the kurtosis. As the mean turns to the maximum 1, the skewness becomes lower and lower in negative, and as the mean turns to the minimum 0, the skewness becomes higher and higher in positive (Fig. 5b). Similarly, when skewness turns to greater either in negative or positive, the SPDs are more likely to be squeezed in half left or right regions, and then the curves turn to be sharper and sharper, making kurtosis to be larger and larger (Fig. 5d). The larger kurtosis means that more values were concentrated in a narrower area near the mean value,

and hence reducing the SD – such kurtosis-SD relationship can be seen in Fig. 5c, where they do demonstrate a good negative correlation (R = 0.71).

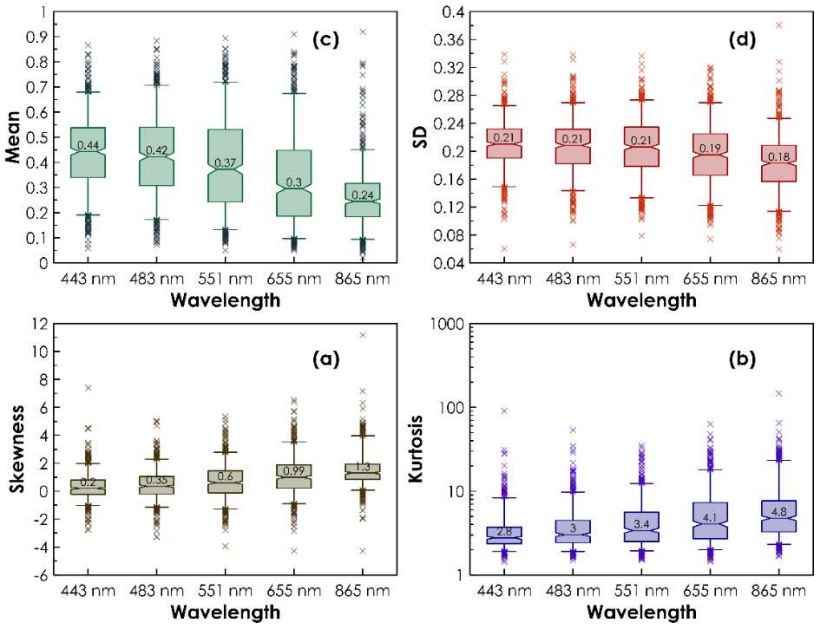

Figure 4. The distributions of four SPD statistical parameters of 688 study waters at five wavelengths: **(a)** skewness, **(b)** kurtosis, **(c)** mean, and **(d)** SD. The boxes and whiskers indicate the locations of 5%, 25%, 50%, 75%, and 95%, and the oblique cross symbols mark the locations of the outliers >95% and <5%.

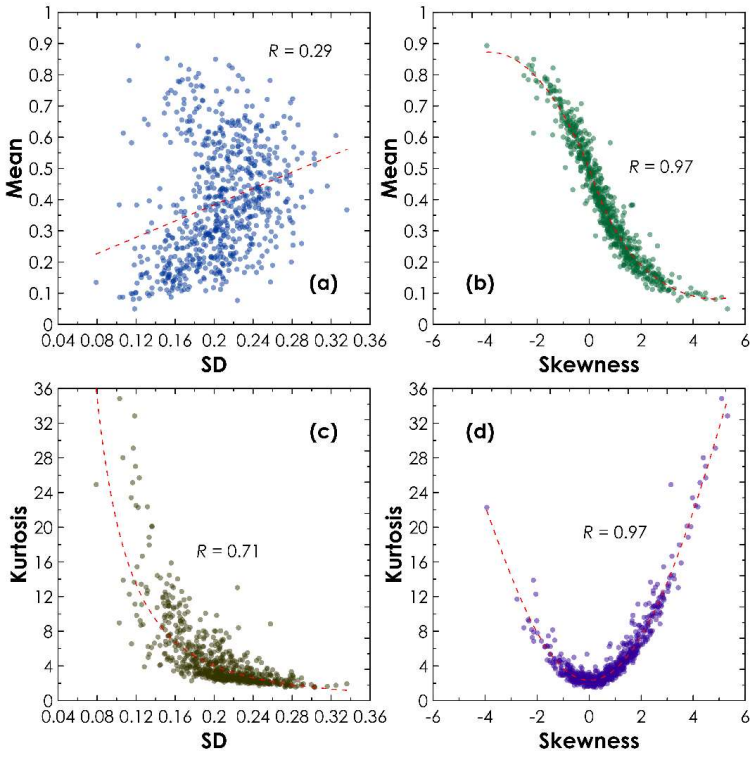

Figure 5. The correlations between two statistical parameters of 688 SPDs at 551 nm: **(a)** mean vs. SD, **(b)** mean vs. skewness, **(c)** kurtosis vs. SD, and **(d)** kurtosis vs. skewness.

(3) What are the SPD's band relationships?

From Fig. 3 we can see that, at a given water, SPD curves at different wavelengths are various. In some water, their five SPDs were quite similar (e.g., Fig. 3c, 3e, and 3d), but in other water, they looked partly different or entirely distinct from each other (e.g., Fig. 3b, 3f, and 3o). In this study, we analyzed the similarity between the SPDs of two bands by using the Jensen-Shannon divergence (JSd) (Lin, 1991; Ôsterreicher & Vajda, 2003). In statistics, the JSd is usually used to measure the similarity between two probability distributions, and it is calculated by

$$JSd(P \parallel Q) = \frac{1}{2}KLd\left(P \parallel \frac{1}{2}(P+Q)\right) + \frac{1}{2}KLd\left(Q \parallel \frac{1}{2}(P+Q)\right) \quad (3)$$

where P, Q are two probability distributions and KLd(P||Q) is the Kullback-Leibler divergence (Kullback & Leibler, 1951) between P and Q, and calculated by

$$KLd(P \parallel Q) = \sum_{x \in X} P(x) log\left(\frac{Q(x)}{P(x)}\right) \quad (4)$$

The JSd(P||Q) always falls within the range (0, 1), and $JSd = 0$ means P and Q are equal almost everywhere, while $JSd = 1$ means they are with the least similarity.

The distributions of the JSd of the study water were shown in Fig. 6, which were calculated from the 10 band pairs of 5 Landsat-8 bands. Fig. 6 shows that the higher similarity occurred when the two bands are closer, e.g., $JSd(1,2)_{mean} = 0.0089$, $JSd(2,3)_{mean} = 0.026$, and $JSd(3,4)_{mean} = 0.031$, see Fig. 6(a). The further the distance between the two bands, the less similarity they have, and the band 5 (865 nm) seems having the least similarity with all the other four bands.

To classify the SPD band relationships shown in Fig. 3, we take $JSd = 0.04$ as a threshold so that if $JSd < 0.04$, then the two SPDs were treated as the similar and set their JSd level (JSdL) = 1, otherwise, they were dissimilar with $JSdL = 0$. As the results, the 688 waters can be classified into 152 types in terms of their JSdLs between 10 band pairs, see Fig. 6b. The water's JSdLs can be represented by a mathematical graph, called SPD graph, shown in Fig. 6c, in which each vertex represents a band, and if the band-pair's $JSdL = 1$, then they were connected by an edge, and if $JSdL = 0$, they were disconnected. Furthermore, a SPD graph can be simplified by a text notation using letters U, B, G, R, and N to represent 5 bands respectively, and if two SPDs are similar, then they are connected by a hyphen '-', otherwise, we do not insert the hyphen. Therefore, if SPDs at all 5 bands were dissimilar, then the water has a SPD graph such as the Fig. 6c(1), with SPD band relationship type (BR-type) UBGRN, whereas if they were all similar, then its SPD graph seems as the Fig. 6c(2), with BR-type U-B-G-R-N.

Among the 688 waters, 23 have the BR-type U-B-G-R-N, e.g. Lake Kakhovska (Fig. 3d) and Yellowstone Lake (44.47N, 110.36W), and 11 have the BR-type UBGRN, e.g. Cardiel Lake (Fig. 3f) and Great Salt Lake (41.28N, 112.73W). The most common BR-type is U-B-G-RN, see 6c(3), and in Fig. 6b, they have the code 1110110100, and 104 waters were with the type, e.g. Bahi Swamp (Fig. 3e), Lake Ross (Fig. 3k), and Saldanha Bay (33.05S, 11.02E). The second large water group were with the type U-B-GRN, and their amount is 62 (Fig. 6c(4)), e.g. Utah Lake (Fig. 3b) and Mono Lake (33.02N, 118.98W). There were 29 waters having the type U-BG-RN (Fig. 6c(5)), e.g. Lake Tarawera (Fig. 3r) and Oyster Harbor (34.97S, 117.96E), and 33 having the type U-BGRN (Fig. 6c(6)) e.g. Lake Griffin (Fig. 3l), Lake Argentino (Fig. 3g), and Mtera Reservoir (7.05S, 35.83E) – these BR-types were also frequently observed among the 688 waters.

We have noticed that in some waters, the SPDs at one or more bands were significantly isolated from the SPDs at the other bands, for example, the green band of Lake Argentino (Fig. 3g), the red band of Utah Lake (Fig. 3b), and the blue bands (443 and 483 nm) of Lac Maunoir (Fig. 3i). The occurrence of these ungrouped SPDs may indicate some special water cases such that, for example, the algal blooms led the SPDs of the green band deviated from the other bands.

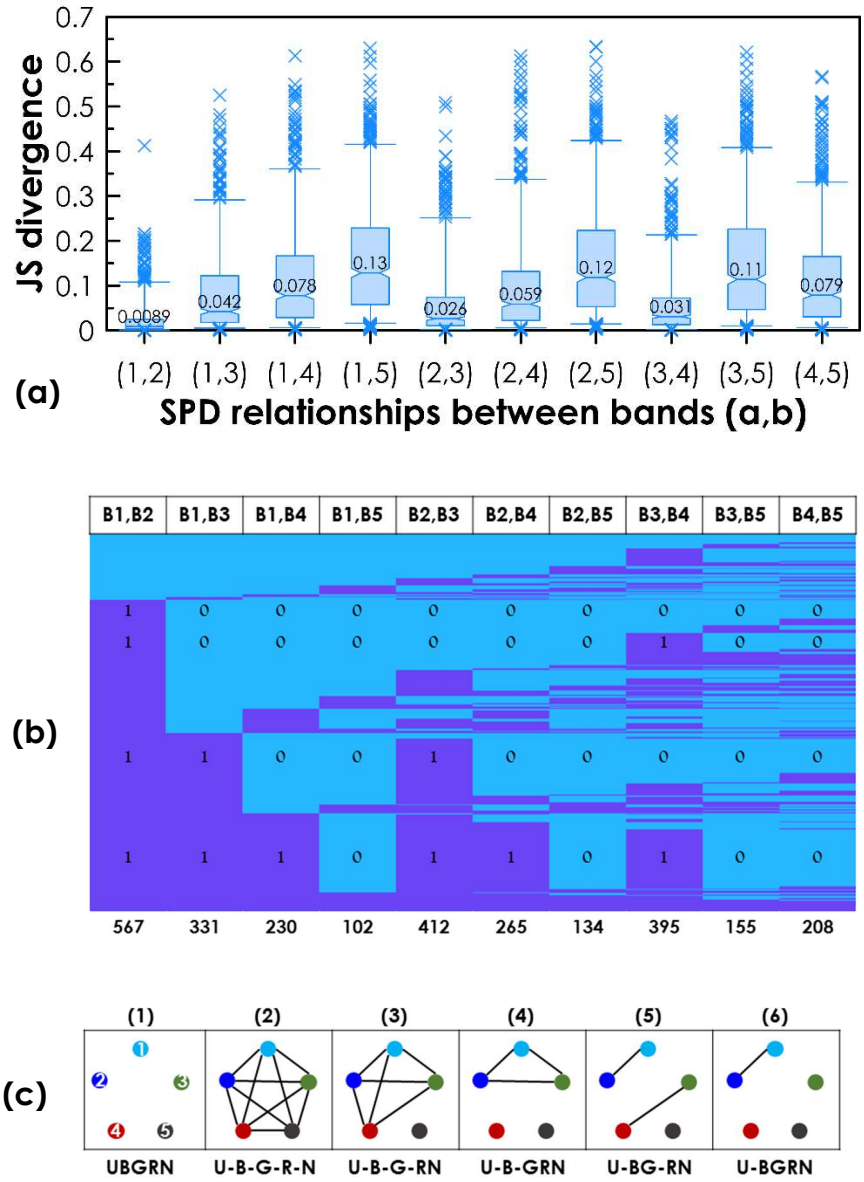

Figure 6. SPD band relationships observed in 688 study waters: **(a)** the distributions of the JSd between 10 band pairs, **(b)** the JSd-based SPD band connection table to classify a water into its SPD band type, with threshold $JSd = 0.04$, and **(c)** the common SPD band relationship graphs and their notations.

Since sometimes the typical histograms or pdf curves shown in Fig. 3 are not easy to discern, we therefore represent the SPD of a given water by using a so-called SPD diagram, see Fig. 7. The x-axis of a SPD diagram represents the normalized R_{rs}, the number in y-axis corresponds to each wavelength, and the color in the diagram represents the possibility at the given wavelength and for the given normalized R_{rs}. Note that, in this study, we only have the five bands used in Landsat-8,

the SPDs of the other wavelengths were interpolated from the known values at the five Landsat-8 bands. The above-mentioned as well as some new SPD features can be clearly observed in SPD diagrams, for example, the skewness in Fig. 7b and 7c. Fig. 7g shows there was a highland in the range of the red wavelength, indicating that the SPDs of the red band was distinct from the other SPDs. We can also see that the SPD diagrams of Lake Tahoe and Lake Griffin are approximately symmetric to each other (Fig. 7h and 7i).

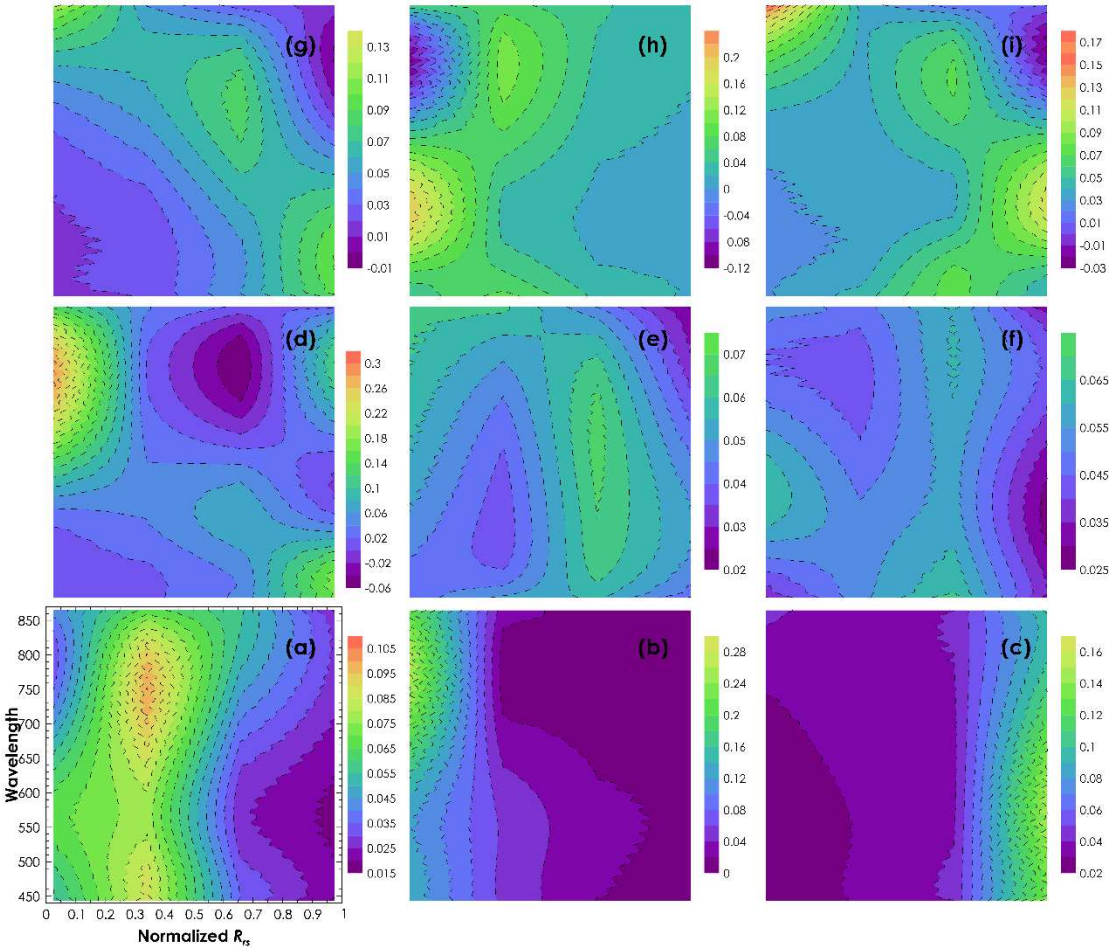

Figure 7. SPD diagrams of 9 closed connected waters: **(a)** Dead Sea, **(b)** Yishan Lake, **(c)** Bahi Swamp, **(d)** Lake Argentino, **(e)** Hudson River Estuary, **(f)** Kakhovka, **(g)** Utah Lake, **(h)** Lake Tahoe, and **(i)** Lake Griffin. The number in color bars shows the possibilities of SPD, and the negative values can be ignored or treated as nearly zero because they were caused by the spline interpolation.

In this section we only gave a tentative and qualitative analysis on the characteristics of the observed SPDs. Due to the lack of the detailed geophysical, biochemical, optical, and environmental properties of many study waters, how these properties were related to their SPDs has not been fully explored. We expect to give more exhaustive and quantitative studies on water's SPDs in future.

3. Forward DD process: theory and simulation

We all know that in classic aquatic optics, water's AOPs are determined by its IOPs, so do the SPDs. Because many variables, such as in-water components, underwater light field, and surface

conditions, are involved in the process of the radiative transfer from IOPs to AOPs, the possibility distribution of each variable may have impact on the remote sensing observed spectra as well as their SPDs. In this section, we made a simplified theoretical calculation and some simulations on this forward DD process.

3.1 A semi-analytical forward DD process for lognormal distributed ocean color components

In this study, we focused only on a 4-component water model (the classic case-2 model used in Hydrolight), which contains pure freshwater as the baseline, plus three ocean color components, chlorophyll (CHL), CDOM, and mineral (or non-algal particles, NAP). Note that we only considered the cases that the three components are independent with each other, but in real water, sometimes two or three of them might be correlated, particularly in those water without input from terrigenous sources, such as the open sea water. However, most of the closed connected water are inland or coastal water, and their components can be approximately taken as the independent random variables.

Assuming the concentrations of three components are lognormal distributed random variables, with their respective parameters μ and variances σ^2, that is,

$$C_{chl} \sim lognormal(\mu_{chl}, \sigma_{chl}^2) \tag{4}$$

$$a_{cdom}(\lambda_0) \sim lognormal(\mu_{cdom}, \sigma_{cdom}^2) \tag{5}$$

$$C_{nap} \sim lognormal(\mu_{nap}, \sigma_{nap}^2) \tag{6}$$

where C_{chl} and C_{nap} denote the concentrations of the CHL and NAP, and they are usually in units of weight of mass in unit volumes, i.e., mg/m^3 or μg/L, while the concentration of CDOM is often represented by its absorption coefficients (in unit m^{-1}) at a typical reference wavelength λ_0 = 440 m.

Because there are good and inherent correlations between physical concentrations and optical concentrations of CHL/NAP, such as the CHL-based mode for case-1 water and the power law Loisel-Morel or Gordon-Morel near surface models used in Hydrolight, then we can obtain the IOPs, i.e., the absorption and backscattering coefficients of CHL, NAP, and CDOM at different wavelengths, using the below models,

$$a_{chl}(\lambda) = t a_{chl}^*(\lambda)[C_{chl}]^k \tag{7}$$

$$b_{chl}(\lambda) = g[C_{chl}]^n(\lambda_0/\lambda)^m \tag{8}$$

$$bb_{chl}(\lambda) = b_{chl}(\lambda) \times r \tag{9}$$

$$a_{cdom}(\lambda) = a_{cdom}(\lambda_0) \times \exp\left(S(\lambda - \lambda_0)\right) \tag{10}$$

where a, b, and bb denote the absorption, scattering, and backscattering coefficients of the respective components, variables t, k, g, n, m, and r are model parameters, and S is the slope indicating a_{cdom}'s decay from the short wavelengths to long wavelengths. Note that the NAP has the same models as the CHL's Eq. (7)-(9), but with the different parameters t_{nap}, k_{nap}, g_{nap}, n_{nap}, m_{nap}, and r_{nap}.

Based on the known statistical properties of lognormal distribution, the possibility distributions of the above optical properties (absorption and backscattering coefficients) can be obtained in forms

$$a_{chl}(\lambda)\sim\text{lognormal}(k_{chl}\mu_{chl}+\ln(t_{chl}a^*_{chl}(\lambda)),k^2_{chl}\sigma^2_{chl}) \quad (11)$$

$$a_{cdom}(\lambda)\sim\text{lognormal}(\mu_{cdom}+S\lambda-S\lambda_{cdom\_0},\sigma^2_{cdom}) \quad (12)$$

$$a_{nap}(\lambda)\sim\text{lognormal}(k_{nap}\mu_{nap}+\ln(t_{nap}a^*_{nap}(\lambda)),k^2_{nap}\sigma^2_{nap}) \quad (13)$$

$$bb_{chl}(\lambda)\sim\text{lognormal}\left(n_{chl}\mu_{chl}+m_{chl}\ln\left(g_{chl}\lambda_{chl\_0}/\lambda\right)+\ln(r_{chl}),n^2_{chl}\sigma^2_{chl}\right) \quad (14)$$

$$bb_{nap}(\lambda)\sim\text{lognormal}\left(n_{nap}\mu_{nap}+m_{nap}\ln\left(g_{nap}\lambda_{nap\_0}/\lambda\right)+\ln(r_{nap}),n^2_{nap}\sigma^2_{nap}\right) \quad (15)$$

CDOM usually has no scattering effect because of its dissolved status, thus the sums of the absorption and backscattering coefficients of three or two components are

$$a_{3com}(\lambda)=a_{chl}(\lambda)+a_{cdom}(\lambda)+a_{nap}(\lambda) \quad (16)$$

$$bb_{2com}(\lambda)=bb_{chl}(\lambda)+bb_{nap}(\lambda) \quad (17)$$

It is known that the additive distributions of a series of lognormal distributions have no closed-form expressions, but still can be reasonably approximated by other lognormal distributions, so the distributions of a_{3com} and bb_{2com} can be expressed by

$$a_{3com}(\lambda)\sim\text{lognormal}(\mu_{a_{3com}},\sigma^2_{a_{3com}}) \quad (18)$$

$$bb_{2com}(\lambda)\sim\text{lognormal}(\mu_{bb_{2com}},\sigma^2_{bb_{2com}}) \quad (19)$$

where

$$\sigma^2_{a_{3com}}=\ln\left(\frac{\sum_{i=chl,cdom,nap}\exp\left(2\mu_{a_i}+\sigma^2_{a_i}\right)\left(\exp\left(\sigma^2_{a_i}\right)-1\right)}{\left(\sum_{i=chl,cdom,nap}\exp\left(\mu_{a_i}+\sigma^2_{a_i}/2\right)\right)^2}+1\right) \quad (20)$$

$$\sigma^2_{bb_{2co}}=\ln\left(\frac{\sum_{i=chl,nap}\exp\left(2\mu_{bb_i}+\sigma^2_{bb_i}\right)\left(\exp\left(\sigma^2_{bb_i}\right)-1\right)}{\left(\sum_{i=chl,nap}\exp\left(\mu_{bb_i}+\sigma^2_{bb_i}/2\right)\right)^2}+1\right) \quad (21)$$

$$\mu_{a_{3com}}=\ln\left(\sum_{i=chl,cdom,nap}\exp\left(\mu_{a_i}+\sigma^2_{a_i}/2\right)\right)-\frac{\sigma^2_{a_{3com}}}{2} \quad (22)$$

$$\mu_{bb_{2com}}=\ln\left(\sum_{i=chl,nap}\exp\left(\mu_{bb_i}+\sigma^2_{bb_i}/2\right)\right)-\frac{\sigma^2_{bb_{2com}}}{2} \quad (23)$$

To calculate the water's total absorption and backscattering coefficients, the absorption and backscattering coefficients of the pure water should be added, that is,

$$a_t(\lambda)=a_{3com}(\lambda)+a_w(\lambda) \quad (24)$$

$$bb_t(\lambda)=bb_{2com}(\lambda)+bb_w(\lambda) \quad (25)$$

Since a_w and bb_w are constants at a given wavelength λ, so the a_t and bb_t should turn to the 3-parameter lognormal distributions such that

$$a_t(\lambda)\sim LN3(a_w,\mu_{a_{3com}},\sigma^2_{a_{3com}}) \quad (26)$$

$$bb_t(\lambda)\sim LN3(bb_w,\mu_{bb_{2com}},\sigma^2_{bb_{2com}}) \quad (27)$$

Next, we can use the well-known QAA (Quasi-Analytical Algorithm) model (citation) to further calculate R_{rs}, namely,

$$u=\frac{bb_t}{a_t+bb_t} \quad (28)$$

$$r_{rs} = g_0 u + g_1 u^2 \tag{29}$$

$$R_{rs} = \frac{Tr_{rs}}{1-\gamma Q r_{rs}} \tag{30}$$

However, these variables, u, r_{rs}, and R_{rs}, are based on the addition, multiplication, reciprocal, power, and other operations of two or more 3-parameter lognormal distributions, and so far, one do not know whether the resultant distributions of these operations are still the 3-parameters lognormal. Let us ignore the intermediate variables u and r_{rs}, and guess the R_{rs}'s distribution is still a 3-parameter lognormal such that

$$R_{rs}(\lambda) \sim LN3(\gamma_{Rrs}, \mu_{Rrs}, \sigma^2_{Rrs}) \tag{31}$$

So we can use some suggested numerical methods to estimate the parameters γ_{Rrs}, μ_{Rrs}, and σ^2_{Rrs}, but the procedures might be too complicated. In this study, we made a simplification that only focused on the parameter bb_t/a_t, which could be regarded as the approximation of surface reflectance R. Then we have

$$R \propto \frac{bb_t}{a_t} = \frac{bb_{2com}+bb_w}{a_{3com}+a_w} = \frac{1}{\frac{a_{3com}}{bb_{2com}+bb_w}+\frac{+a_w}{bb_{2com}+bb_w}} \tag{32}$$

If we further neglected the bb_w that is relatively tiny to the other parameters, then the Eq. (32) is simplified to

$$\frac{bb_t}{a_t} \approx \frac{1}{\frac{a_{3com}}{bb_{2com}}+\frac{a_w}{bb_{2com}}} = R_{sp} \tag{33}$$

Let $V_1 = \frac{a_{3com}}{bb_{2com}}$, $V_2 = \frac{a_w}{bb_{2com}}$, and $V_3 = V_1 + V_2$, then V_1, V_2, and V_3 should be lognormal distributions with parameters

$$\mu_{V_1} = \mu_{a_{3com}} - \mu_{bb_{2com}} \tag{34}$$

$$\sigma_{V_1} = \sigma^2_{a_{3c}} + \sigma^2_{bb_{2co}} \tag{35}$$

$$\mu_{V_2} = \ln(a_w) - \mu_{bb_{2com}} \tag{36}$$

$$\sigma_{V_2} = \sigma^2_{bb_{2com}} \tag{37}$$

$$\sigma^2_{V_3} = \ln\left(\frac{\sum_{i=1,2} e^{2\mu_{V_i}+\sigma^2_{V_i}}(e^{\sigma^2_{V_i}}-1)}{\left(\sum_{i=1,2} e^{\mu_{V_i}+\sigma^2_{V_i}/2}\right)^2}+1\right) \tag{38}$$

$$\mu_{V_3} = \ln\left(\sum_{i=1,2} e^{\mu_j+\sigma_j^2/2}\right) - \frac{\sigma^2_{V_3}}{2} \tag{39}$$

Then the variable R_{sp} should be a lognormal distribution such that

$$R_{sp} \sim \text{lognormal}(-\mu_{V_3}, \sigma^2_{V_3}). \tag{40}$$

The above equations tell how to theoretically calculate distribution parameters μ and σ^2 of each optical variable, while we can also use Monte-Carlo method to simulate these variables, and with assuming that they were lognormal distributions, their μ and σ^2 can be calculated from the simulated samples by using

$$\mu = \ln \frac{E(X)^2}{\sqrt{\mathrm{Var}(X)+E(X)^2)}} \tag{41}$$

$$\sigma^2 = \ln\left(1 + \frac{\mathrm{Var}(X)}{E(X)^2}\right) \tag{42}$$

where $E(X)$ and $Var(X)$ are the expectation (mean) and variance of the random variable X.

Therefore, we could compare the theoretically calculated and the Monte-Carlo simulated parameters μ and σ^2 that were used to characterize the lognormal distributions of a_x, bb_x, and R_{sp}. The results of these comparisons, see Table 2, show that the most of relative errors between the theoretical and simulated parameters were very small (< 2%), and many of them were smaller than 0.5% or even 0.1% for parameter μ. The relative errors of the parameter σ^2 were slightly larger than those of μ. However, there were also relatively larger errors (30%-55%) between the σ^2s of the theoretical R_{sp} and simulated R_{sp}, particularly within the visible bands (Fig. 8b), while the relative errors of R_{sp}'s μ still remained very small (Fig. 8a). The large errors of R_{sp}'s σ^2 were possibly caused by the errors of the variable $V_1 = a_{3com}/bb_{2com}$. Because the distributions of a_{3com} and bb_{2com} are both formed by the additions of the other three or two lognormal distributions, while the results of these additions are only the approximation of the lognormal distribution, and hence the division of two approximations may bring more errors into the next result. We also noticed that for some variables, such as the bb_{2com} and the a_x and/or bb_x of CHL, CDOM and NAP, their σ^2s are independent of the wavelength, making their errors are the same at all bands.

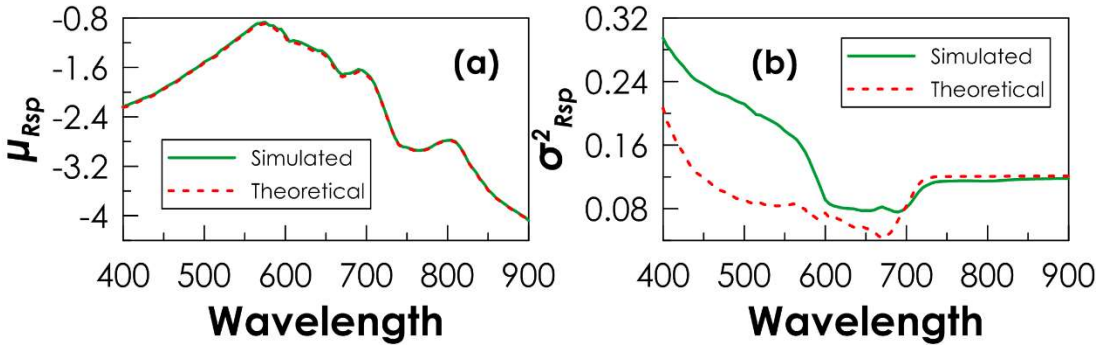

Figure 8. The comparisons between the Monte-Carlo simulated and the theoretically calculated statistical parameters μ and σ^2 for R_{sp}, which is assumed to be with an approximate lognormal distribution.

Table 2. The relative errors (%) between the Monte-Carlo simulated and the theoretically calculated statistical parameters at 5 wavelengths for the lognormal distributed variables used in a simplified 3-component radiative transfer.

	440 nm		480 nm		550 nm		655 nm		865 nm	
variable	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
a_{chl}	1.32	1.57	0.92	1.57	0.45	1.57	0.47	1.57	0.09	1.57
bb_{chl}	0.54	1.50	0.51	1.50	0.47	1.50	0.43	1.50	0.38	1.50
a_{nap}	0.11	1.96	0.10	1.96	0.09	1.96	0.09	1.96	0.10	1.96
bb_{nap}	0.10	1.90	0.09	1.90	0.09	1.90	0.09	1.90	0.08	1.90
a_{cdom}	5.17	1.51	1.12	1.51	0.36	1.51	0.18	1.51	0.09	1.51
a_{3com}	0.01	0.65	0.55	0.53	0.18	0.44	0.39	1.21	0.09	1.32
bb_{2com}	0.49	1.80	0.46	1.80	0.42	1.80	0.38	1.80	0.33	1.80
R_{sp}	0.74	48.03	0.83	54.95	1.91	52.90	2.35	30.25	0.21	2.96

Through the above study we can see that there are possibly the theoretical connections between the observed parameters of SPD and the distribution parameters of the underlying components as well as the intermediate variables. Following the radiative transfer of the light, these distribution parameters also experience a series of transfer. Despite many approximation and simplification being made, their theoretical connections may still be very complicated. Nevertheless, if these connections can be approximately established for at least some particular cases, for example, all three components are with lognormal distributions, then their statistical parameters might be analytically calculated via some inverse mathematics, such as the MIM (matrix inversion method) used in the traditional ocean color remote sensing inversion.

3.2 Forward DD process observed from Hydrolight simulation

The above Monte-Carlo simulation was based on the simplified radiative transfer model QAA or only calculating R_{sp}, the more accurate approach is using the Hydrolight, a well-known radiative transfer model widely used in ocean color remote sensing community. By using Hydrolight, here we studied 9 cases of different distribution combinations of three components, including the uniform, lognormal, exponential, see Table 3. We particularly paid attention to the CDOM, so in the first three cases, we let CHL and NAP's concentrations to be the constants, and then we were able to observe how CDOM's distribution change AOP's distribution. The datasets of the three components were produced by the MATLAB's random functions, in which the used parameters and the mean and SD of each resultant component dataset were also shown in Table 3. Each case contains three component datasets, and each dataset contains 10,000 samples, that is, each case contains 10,000 water samples in which CHL, CDOM, and NAP had different possibility distributions. Then these water samples were put into the Hydrolight to obtain its outcomes, namely, the simulated R_{rs}. The other input models, parameters, constants, and coefficients used by the Hydrolight were set by their defaults.

Table 3. Study cases with different distributions of three ocean color components for Hydrolight simulation

Cases	CHL (mg/m^3)			CDOM (m^{-1})			NAP (mg/L)		
	pdf & para.	mean	SD	pdf & para.	mean	SD	pdf & para.	mean	SD
1	Const(20)	20	0	Unif(0.2,2)	1.10	0.52	Const(10)	10	0
2	Const(20)	20	0	LN(0.1,0.5)	1.26	0.67	Const(10)	10	0
3	Const(20)	20	0	Exp(1)	1.00	0.99	Const(10)	10	0
4	Unif(5,50)	27.55	12.95	Unif(0.2,2)	1.10	0.52	Unif(2,20)	11.03	5.17
5	Unif(5,50)	27.40	13.05	LN(0.1,0.5)	1.24	0.65	Unif(2,20)	10.00	5.18
6	Unif(5,50)	27.47	13.02	Exp(1)	0.98	0.99	Unif(2,20)	11.00	5.18
7	Exp(30)	30.05	30.19	LN(0.1,0.5)	1.25	0.66	Exp(20)	20.23	20.10
8	LN(3.5,0.5)	37.30	19.98	LN(0.1,0.5)	1.25	0.66	LN(2.5,0.5)	13.92	7.49
9	LN(3.5,0.5)	37.34	19.94	Exp(1)	1.01	1.00	LN(2.5,0.5)	13.92	7.38

Note: Const(c) denotes the optical property is a constant c, Unif(a,b) denotes a uniform distribution between a and b, Exp(λ) denotes an exponential distribution with a parameter λ, and LN(μ,σ) denotes a lognormal distribution with its parameters μ and σ used for MATLAB's random number function lognrnd(μ,σ^2). The mean and SD were calculated from their respective simulation results.

The Hydrolight-simulated results are shown in Fig. 9. Due to the high computation cost of the simulation, we only simulated the output R_{rs} at the Landsat-8 bands B2-B5. Given the CHL and NAP concentrations were constants, we can see that the R_{rs}'s distributions were consistent with the CDOM's distributions, that is, if CDOM was with uniform, lognormal, and exponential distributions, then the Rrs, at all four bands, was also with uniform, lognormal, and exponential distributions, respectively, see cases 1-3 in Fig. 9. If the three components were all with uniform distributions, then the R_{rs}'s distributions seemed to be normal or lognormal, see case-4 in Fig. 9. If CDOM was with either lognormal or exponential distributions, while the CHL and NAP were with one of the uniform, lognormal and exponential distributions, then the R_{rs}'s distributions seemed to be always the lognormal, but with different parameters such as the SD and kurtosis, see cases 5-9 in Fig. 9. It is also easy to see that in all cases, except the case-1, SPDs had the negative skewness, and moreover, their SPDs looked all similar at the four bands. Although such scenarios, i.e., negative skewness and all-band similarity, were indeed frequently found in real water (see examples introduced in Section 2.3), there were also many other scenarios happened in nature (such as the diverse SPDs shown in Fig. 3) but not observed in the simulated results. Why there were so many various SPDs in real water and how to form them using simulation are still unknown.

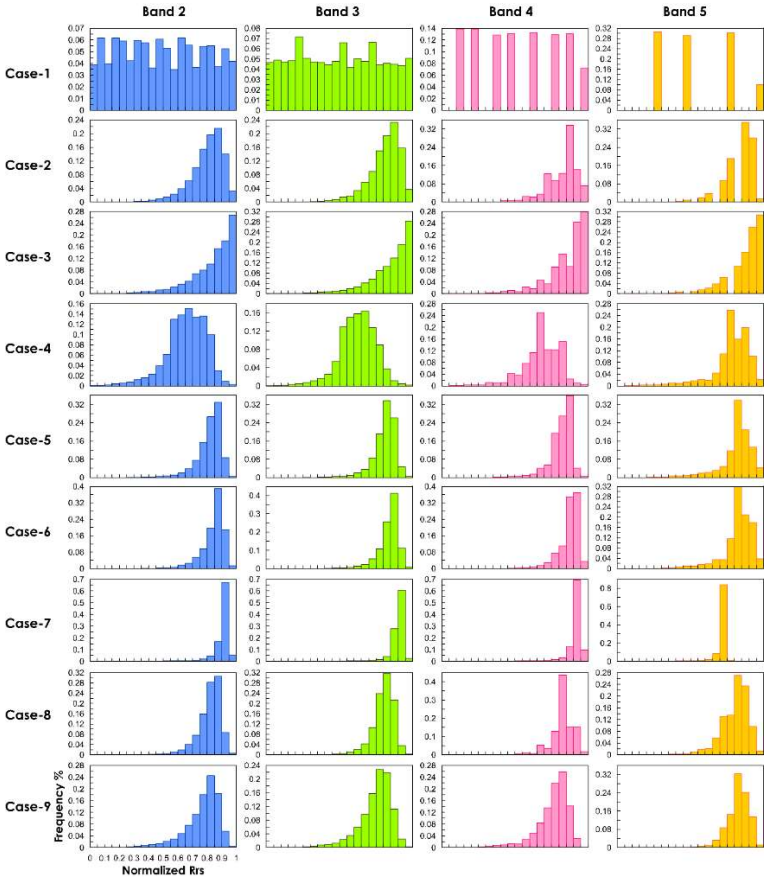

Figure 9. The Hydrolight simulated SPDs at 4 Landsat-8 bands for 9 cases listed in Table 3.

4. Backward DD scheme: statistical inference

The backward DD (bDD) scheme is of the most importance for the DD scheme, and is also the useful tool for ocean color remote sensing applications. Basically, there are two types of bDD scenarios when it is used for the image-based statistical inference: (1) with field measured data, and

(2) without field measured data. Here we only study the scenario that we have the field measured data, which contain the samples obtained from their population. Based on the measured samples, what we need to do is establishing a statistical inference model to estimate CDOM's possibility distribution from the SPDs observed in an image scene.

4.1 Bootstrap method and data process

Many statistical inference techniques may be used in bDD scheme, but in this study, we only used a bootstrap-based redistribution/resample method. Given a field measured dataset **D**, see Fig. 10, we first resample a series of subset ($\mathbf{d_1}$, $\mathbf{d_2}$, …, $\mathbf{d_i}$), and in each subset, CDOM and SPDs may have different distributions. Then from each subset of $\mathbf{d_1}$, $\mathbf{d_2}$, …, $\mathbf{d_i}$, we calculate CDOM's statistical parameters Sp_{cdom_1}, Sp_{cdom_2}, …, Sp_{cdom_k} as well as the SPD's statistical parameters Sp_{spec_11}, Sp_{spec_12}, …, Sp_{spec_21}, Sp_{spec_22},…, Sp_{spec_mn} at a single band or some band-combinations, and hence to have a series of CDOM parameter subsets $\mathbf{C_1}$, $\mathbf{C_2}$, …, $\mathbf{C_i}$, and their corresponding SPD parameter subsets $\mathbf{S_1}$, $\mathbf{S_2}$, …, $\mathbf{S_i}$. These statistical parameters could be the mean, SD, skewness, and other parameters used for describing a possibility distribution, and the band-combinations could be two band ratios or other forms with more bands combined. Next, from $\mathbf{C_1}$, $\mathbf{C_2}$, …, $\mathbf{C_i}$, we extract the parameter $Sp_{\text{cdom\_1}}$ such as $CDOM_{mean}$ or $CDOM_{SD}$, and from $\mathbf{S_1}$, $\mathbf{S_2}$, …, $\mathbf{S_i}$, we extract the SPD parameters such as $Rrs_{\text{\_mean}}$ and $R_{\text{rs\_SD}}$, and then we can establish the statistical inference or simply the fitting model between, for example, $CDOM_{mean}$ and $Rrs_{\text{\_mean}}$. Other CDOM's distribution parameters can be obtained using the similar way, and the more parameters we get, the more accurate we know its distribution.

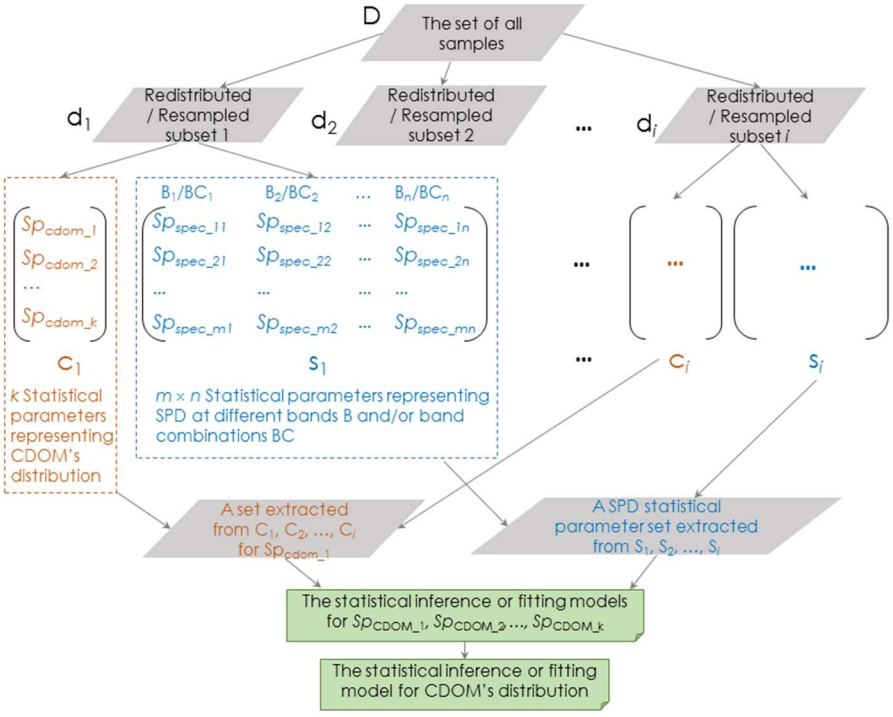

Figure 10. The bootstrap process of making a statistical inference model for one statistical parameter (e.g., the mean or SD) of CDOM distribution. Other parameters can follow the similar process to make their models.

We tested the bootstrap-based method by using 11 datasets, in which ten sets are Hydrolight-simulated datasets and one set is a field-measured dataset (QDH) from Qiandaohu Lake. Among the ten simulated datasets, nine sets are the cases listed in Table 3, and the one is the IOCCG (International Ocean-Color Coordinating Group) synthetic dataset (IOCCG, 2006), which contains 500 samples and is used for ocean color algorithm tests and validation, and its data can be downloaded from and more information can be referred to IOCCG's website.

The Qiandaohu Lake (29.57N, 118.94E), with water area 580 km^2, is a water-supply reservoir in Zhejiang Province, east of China, see the study map and sampling locations in Fig. 11a. More details of the lake and our field/lab measurements can be found in reference (Zhu et al., 2020). Note that the QDH samples were evenly distributed over the lake but they were collected from five field trips at different dates in April and September of 2018 and 2019.

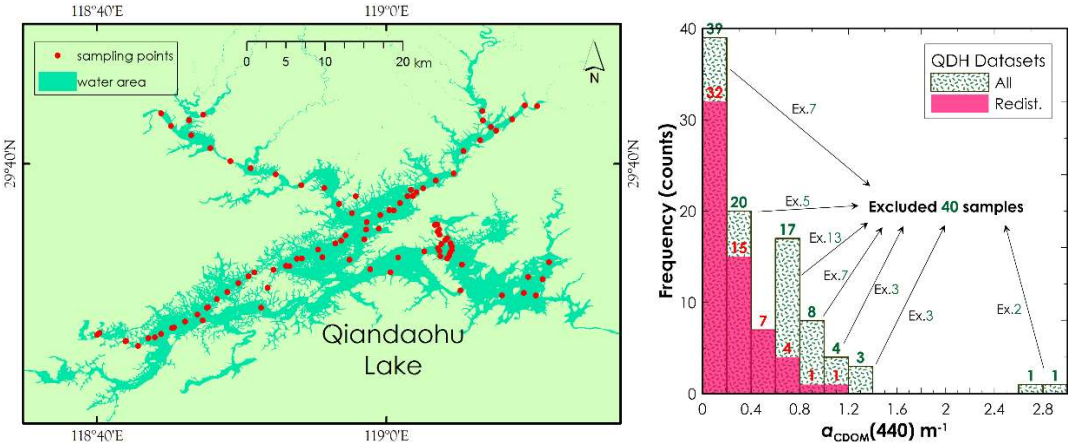

Figure 11. **(a)** The study site map and sampling locations in Qiandaohu Lake, and **(b)** the histogram of the measured CDOM concentrations in five field trips, which collected 100 samples in total (QDH_{all}), and by excluding 40 samples, it was redistributed to a new dataset QDH_{rd}, containing 60 sample, and these samples form a better exponential possibility distribution than the original 100 samples in QDH_{all} do.

In this study, we only focused on two widely-used statistical parameters, mean and SD, which are the key variables to indicate a possibility distribution. The bootstrap processing used for simulated datasets and field-measured datasets were slighted different. For the simulated datasets, we treated them as the population samples, that is, which can be regarded as the samples covering the entire enclosed water. Therefore, we first randomly selected a subset from the population set, simulating the process that we go to the field and sample water from an interested lake. For the field-measured dataset QDH, it is already a subset of the population samples, so it is no need to make a subset resampling as we did for the simulated datasets.

Each of case 1-9 datasets contains n_{pop} = 10,000 samples, so we made a start subset containing n_{start} = 1,000 random selected samples for each case. The IOCCG dataset contains n_{pop} = 500 samples, and we made a subset containing n_{start} = 100 samples. The QDH dataset contains n_{start} = 100 samples, but we also test a QDH's subset containing n_{start} = 60 samples. Note that the 60 samples were not randomly selected from the all 100 QDH samples. Instead, they were the artificially selected in order to make a neat distribution which would be likely to represent the true distribution of the

whole lacustrine CDOM. Thereafter we denote the dataset containing all 100 samples QDH_{all}, and its redistributed subset QDH_{rd}.

The processing of the bootstrap-based statistical inference followed the Fig. 10 as the below steps:

(1) Resampling a test subset from the start simulated/measured set/subset, and the test subset contains n_{test} = 100, 20, 50, 30, for Case1-9, IOCCG, QDH_{all}, and QDH_{rd} datasets, respectively.

(2) For each test subset, calculating its mean and SD values for CDOM concentration (a_{cdom} at 440 nm) and R_{rs} of a single band or band-combinations. In this study we used four Landsat-8 bands B1-B4 and their six band-ratios B1/B2, B1/B3, B1/B4, B2/B3, B2/B4, and B3/B4 for the datasets IOCCG and QDH, but used the bands B2-B5 and their six band-ratios B2/B3, B2/B4, B2/B5, B3/B4, B3/B5, and B4/B5 for the datasets Cases 1-9. Note that in these calculations, R_{rs} should use their original values in unit sr-1, rather than those normalized R_{rs} used for representing and comparing SPDs.

(3) Repeating the above steps (1) and (2) k = 1,000 times, and then we obtained the datasets containing 1,000 samples which can be used for modeling the relationships between the CDOM's mean\SD and SPD's mean\SD.

(4) Making a simple linear regression using the data acquired by the Step (3), and the regressed equations can be used for validating and further inference on new datasets.

4.2 Inference results and discussion

The results of inference modeling for CDOM's mean and SD are shown in Table 4 and Fig. 12. We can see that there are good correlations between CDOM's mean/SD and SPD's mean/SD. The best correlations occurred in Case 1-3, showing that if the concentrations of the other two components CHL and NAP were exactly the constants, that is, the R_{rs} was only determined by the CDOM, then CDOM's mean and SD were highly correlated with SPDs ($R^2 > 0.95$), and the high correlations occurred for almost all single bands or band-ratios. However, the constant concentration of an ocean color component is almost impossible to happen in real water, if CHL and NAP were both variables, such as in Case 4-9, then their correlations were about R^2_{mean} = 0.35-0.85 and R^2_{SD} = 0.02-0.75%. Except the Case 7, all other Cases had the $R^2_{mean} > 0.85$ and except Case 4 and Case 7, all other Cases had the $R^2_{SD} > 0.58$.

For the IOCCG dataset, the mean and SD inference models were both accurate with R^2_{mean} = 0.85 and R^2_{SD} = 0.81, while for the QDH datasets, the SD models become even worse with R^2_{SD} = 0.05 for QDH_{all} and R^2_{SD} = 0.12 for QDH_{rd}, meaning that these models were actually failed for the expected statistical inference. How to further improve SD inference model for field measured data remains for the future work.

Unlike the SD models, the performance of the mean models based on QDH_{all} and QDH_{rd} were significantly different. If directly using all data, the best R^2_{mean} was only 0.10, as poor as the SD model. The reason is that all field measured data may not be a real random sampling to represent the true SPD of the given water. For example, the samples' spatial and temporal intervals were not equal, and sometimes samples were collected at different dates. The changes of the flow velocity, vessel speed, plus that the weather may bring uncertainties to the field measurement. Therefore, it

is important to screen the raw samples to redistribute them as a possible or hypothesized possibility distribution. In this study, we can see that CDOM's distribution of all samples did not match any known distribution - it looks like an exponential distribution, but within some bins, samples were possibly either missing or redundant (Fig. 11b). Because we cannot make up those missing samples, we had to remove some redundant samples to let the distribution neat and clear. Thus, we randomly excluded 1-13 samples from each bin, and in total 40 samples were removed from the QDH_{all}, and the rest samples QDH_{rd} were redistributed into a good exponential distribution, see Fig. 11b. As the result, see Table 4 and Fig. 12i and 12j, the mean inference model based on QDH_{rd} was much better (R^2_{mean} = 0.36) than the model based on QDH_{all} (R^2_{mean} = 0.10), which was actually a failed model.

Below we listed some inference functions.

For the Case-8 dataset (see Fig. 12a and 12b),

$$a_{CDOM}(440)_{mean} = -33.2697 \times \left(\frac{R_{rs}(B3)}{R_{rs}(B5)}\right)_{mean} + 39.1417 \quad (43)$$

$$a_{CDOM}(440)_{SD} = 26.1672 \times \left(\frac{R_{rs}(B3)}{R_{rs}(B5)}\right)_{SD} + 0.1611 \quad (44)$$

For IOCCG dataset (see Fig. 12e and 12f),

$$a_{CDOM}(440)_{mean} = 0.0214 \times \left(\frac{R_{rs}(B2)}{R_{rs}(B4)}\right)_{mean} + 0.022 \quad (45)$$

$$a_{CDOM}(440)_{SD} = 0.0133 \times \left(\frac{R_{rs}(B2)}{R_{rs}(B4)}\right)_{SD} + 0.0013 \quad (46)$$

For Qiandaohu Lake (see Fig. 12j),

$$a_{CDOM}(440)_{mean} = 1.536 \times \left(\frac{R_{rs}(B1)}{R_{rs}(B2)}\right)_{mean} - 1.0312 \quad (47)$$

The above inference models Eq. (43)-(47) were validated using the datasets resampled from the population datasets, and hence part of their data were never used for inference modeling. For Case-8, IOCCG, and QDH_{rd} datasets, each one was resampled 500 times that makes 500 validation subsets, and in each validation subset, sample number was randomly set but within the ranges 100-5,000, 20-500, and 10-50 for each dataset, respectively. The validation results were shown in Fig. 12c, d, g, h, and k. For the Case-8 dataset, the $MAPE_{mean}$ = 0.77% (MAPE, Mean absolute percentage error), $RMSE_{mean}$ = 0.010 m^{-1} (RMSE, root mean square error), $MAPE_{SD}$ = 2.00%, $RMSE_{SD}$ = 0.017 m^{-1}; for IOCCG dataset, the $MAPE_{mean}$ = 2.45%, $RMSE_{mean}$ = 0.00014 m^{-1}, $MAPE_{mean}$ = 2.63%, $RMSE_{SD}$ = 0.00013 m^{-1}, and for QDH_{rd} dataset, the $MAPE_{mean}$ = 7.25%, $RMSE_{SD}$ = 0.023 m^{-1}. We can see that even for the field measured QDH_{rd} dataset with only 60 samples, the inference model still achieved the excellent accuracy that the relative error on CDOM's inferred mean was less than 8%.

The results of inference modeling and validation also show that the accuracy of inferring the parameter mean was always better than that of inferring the parameter SD, indicating that the mean is a more sensitive and easy-inferred parameter than the SD. See Table 4, except the Cases 1-3, the best R^2_{SD} was only 0.75 in the Case-9, while, except the Case-7, the worst R^2_{mean} was still 0.84 in the Case-6.

The Case-7 seems to be a difficult case for which the mean and SD inference models both performed not well: $R^2_{mean} = 0.35$ and $R^2_{SD} = 0.02$. So far, we do not know the real reasons that led the inference's difficulty. We believe it could be caused by the specific distributions of CHL (exponential), CDOM (lognormal), and NAP (exponential).

Table 4. The R-squared values for modeling the distribution parameters between CDOM and bands/band-ratios for the datasets Case 1-9, IOCCG, QDH_{all}, and QDH_{rd}. The underlined number are the best three results for each dataset.

(a) For CDOM's mean

Dataset	B2	B3	B4	B5	B2/B3	B2/B4	B2/B5	B3/B4	B3/B5	B4/B5
Cases 1	1.00	1.00	0.99	0.94	0.99	0.97	1.00	0.74	**1.00**	0.98
Cases 2	1.00	1.00	0.99	0.96	0.99	0.97	1.00	0.82	**1.00**	0.99
Cases 3	1.00	1.00	0.99	0.98	0.99	0.98	1.00	0.89	**1.00**	0.99
Cases 4	0.67	0.67	0.61	0.02	0.51	0.69	0.83	0.55	**0.85**	0.85
Cases 5	0.73	0.74	0.69	0.08	0.59	0.74	0.85	0.67	**0.88**	0.87
Cases 6	0.82	0.81	0.78	0.10	0.76	0.83	**0.86**	0.72	0.84	0.84
Cases 7	0.14	0.12	0.11	0.00	0.11	0.11	0.33	0.04	**0.35**	0.48
Cases 8	0.80	0.79	0.75	0.10	0.66	0.79	0.88	0.67	**0.91**	0.89
Cases 9	0.89	0.87	0.86	0.16	0.84	0.89	0.92	0.75	0.91	**0.92**
	B1	B2	B3	B4	B1/B2	B1/B3	B1/B4	B2/B3	B2/B4	B3/B4
IOCCG	0.79	0.07	0.21	0.30	0.72	0.37	0.73	0.41	**0.85**	0.03
QDH_{rd}	0.20	0.09	0.08	0.13	**0.36**	0.05	0.00	0.01	0.09	0.07
QDH_{all}	0.09	0.05	0.05	0.06	**0.10**	0.01	0.01	0.00	0.00	0.01

(b) For CDOM's SD

Dataset	B2	B3	B4	B5	B2/B3	B2/B4	B2/B5	B3/B4	B3/B5	B4/B5
Cases 1	1.00	**1.00**	0.94	0.79	0.96	0.84	0.99	0.32	0.99	0.92
Cases 2	0.99	1.00	0.99	0.98	0.99	0.98	0.99	0.85	**1.00**	0.99
Cases 3	0.99	1.00	1.00	0.99	0.99	0.98	0.99	0.89	**1.00**	0.99
Cases 4	0.14	0.15	0.10	0.00	0.07	0.18	0.34	0.11	0.34	**0.37**
Cases 5	0.40	0.42	0.36	0.00	0.33	0.45	0.56	0.54	**0.62**	0.60
Cases 6	0.48	0.49	0.44	0.02	0.39	0.51	0.56	0.55	**0.58**	0.56
Cases 7	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.00	**0.02**	0.09
Cases 8	0.54	0.55	0.51	0.05	0.47	0.57	0.66	0.62	**0.71**	0.70
Cases 9	0.67	0.68	0.66	0.02	0.60	0.67	0.72	0.70	**0.75**	0.74
	B1	B2	B3	B4	B1/B2	B1/B3	B1/B4	B2/B3	B2/B4	B3/B4
IOCCG	0.50	0.02	0.00	0.00	0.42	0.14	0.70	0.12	**0.81**	0.02
QDH_{rd}	0.00	0.02	0.01	0.00	0.00	**0.12**	0.00	0.07	0.02	0.06
QDH_{all}	0.04	**0.05**	0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.00

We have also tested for using different sample number in the start and test subsets, i.e., changing n_{start} and n_{test}, as well as the bootstrap number k and the sample number of the validation subset. We found that, by changing these variables, the accuracies of inference modeling and validation did not have significant changes, meaning that these variables would not have great impact on SPD statistical inference, provided they were reasonably and appropriately configured.

Bands or band combinations also play an important role in statistical inference. For all datasets used in this study, the models based on simple band-ratios performed always better than or nearly equal to those models based on a single band. This is reasonable because in traditional ocean color remote sensing inversion, the algorithms based on multiple bands are usually better than those single band ones. According to the results shown in Table 4, the best three band ratios for CDOM's mean/SD inference are B3/B5, B4/B5, and B2/B5 for the Case 1-9; for the IOCCG dataset, the best three are B2/B4, B1/B4, and B1; but for QDH data, the best one is B1/B2 for mean's inference. The

band changes in different datasets may be caused by their different CDOM levels. The similar scenarios also occurred in SC scheme – our previous study found that for high-CDOM water, the best Landsat-8 inversion model used the band ratio B3/B4, while for low-CDOM water, the best ratio was shifted to B2/B4 (Chen et al., 2017; Chen et al., 2019). In this study, CDOM concentration level in IOCCG dataset were as very low as those in open seas, so its best inference model using B2/B4 is consistent with the previous results. In addition, the relationships between CDOM mean/SD and spectral mean/SD were regressed using the linear functions. We would like to remain for the future work to explore whether the more complicated band combinations and regression functions would be more suitable for CDOM's statistical inference.

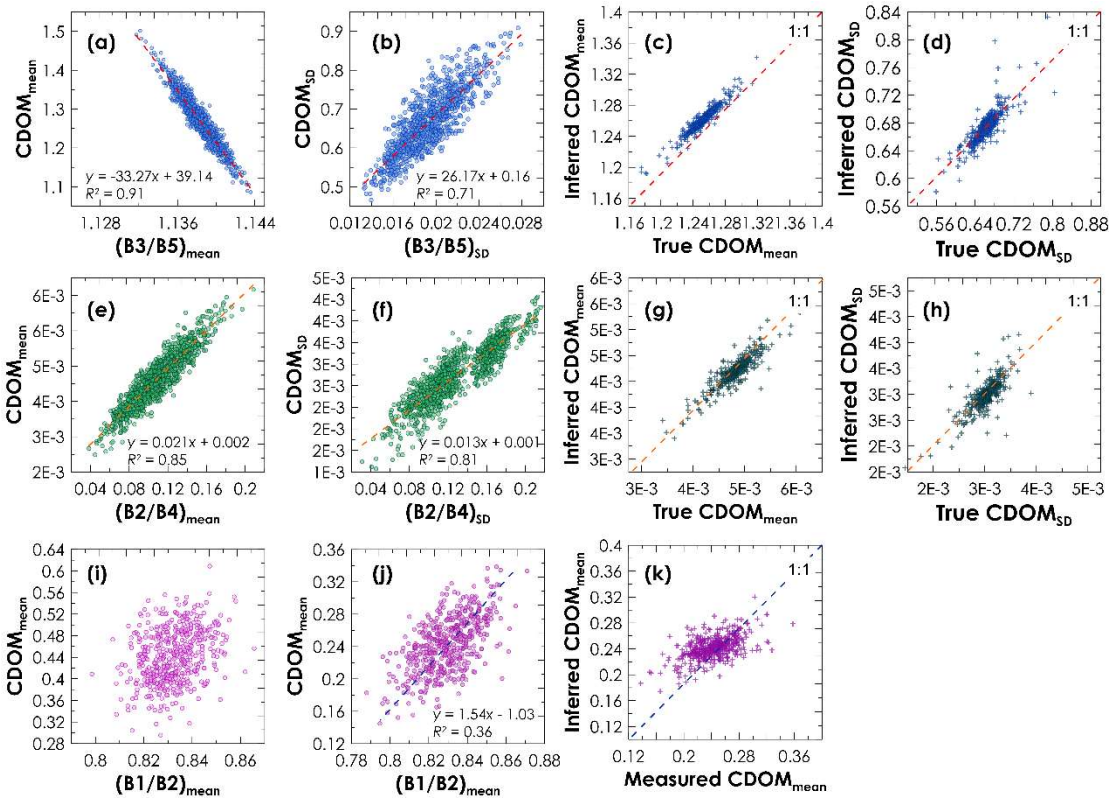

Figure 12. Statistical inference modeling and validating for CDOM's distribution parameters mean and SD using different datasets: **(a)** Case 8, modeling mean, **(b)** Case 8, modeling SD, **(c)** Case 8, validating mean, **(d)** Case 8, validating SD, **(e)** IOCCG, modeling mean, **(f)** IOCCG, modeling SD, **(g)** IOCCG, validating mean, **(h)** IOCCG, validating SD, **(i)** QDH_{all}, modeling mean for all data, **(j)** QDH_{rd}, modeling mean for the redistributed data, **(k)** QDH_{rd}, validating mean.

5. Conclusion and future work

This study suggested a new topic for ocean color and aquatic remote sensing which explores the relationship between the possibility distributions of IOPs and AOPs. This topic is still on its early stage and many concepts, methods, observations, and conclusions might be tentative. Here we

particularly listed a few notes and conclusions deserving more attentions and future work.

The Landsat-8 observed SPDs from 688 global waters show that many of them were normal, lognormal, and exponential distributions, but also with different mean, SD, skewness, and kurtosis. The similarity and diversity observed among these SPDs indicate that SPDs are very likely the response to the optical properties of the in-water components and hence the features of aquatic environment – more studies should be conducted on this work.

In this study, we have not analyzed the SPD's temporal variations on a given water area, that would be an interest topic for the future work. We have made a tentative analysis on the SPD diagrams, which we think can be taken as the spectral 'face' of a lake at a moment. Given the hyperspectral images are more and more available, the SPD diagrams will be expected to have higher spectral resolutions, and then more advanced analytical techniques, such as the artificial intelligence and machine learning, can be used to recognize the patterns in SPDs and reveal their relationships to the other natural properties of water and aquatic environment. In addition, although the band-combinations were used in statistical inference, the SPDs of band-combinations of the real water have not been extracted and analyzed yet. We expect the band-combination-based SPDs may be more closely associated to the possibility distributions of the in-water components.

The results of the forward DD analysis have demonstrated that the IOP's distributions do have the impacts on the observed SPDs. If the distributions of the three components are lognormal, then the corresponding SPDs are also with the lognormal-like distributions. The results of the image observation and Hydrolight simulation both show that the lognormal distribution is widely occurred, and we may pay more attention on it.

Our study shows that based on the bootstrap method, the mean and SD of CDOM's distribution can be accurately inferred, simply using some linear functions and band ratios. Compared to the mean's inferring model, the SD's performed relatively poor, probably because, as a distribution parameter, SD is not sensitive to the observed SPDs. Whether there are more appropriate parameters as well as other statistical inference techniques can be used in SPD's analysis can be further explored.

At last, we would like to emphasize that due to the extreme complexity of the radiative transfer, light-water interaction, and ambient uncertainties, the process of how the AOP's SPD responds to component or IOP's possibility distributions might be also extremely complex, and hence making the inversed statistical inference from AOP to components or IOP is even more complex. We are looking forward to more theoretical and practical future studies on all relevant topics of SPD-water relations and remote sensing statistical inference principles and methods.

Acknowledgement

This study is supported by National Natural Science Foundation of China (No. 41971373, 41876031, 41471346) and National Key R&D Program of China (2017YFB0503902).

Reference

Alcântara, E., Bernardo, N., Watanabe, F., Rodrigues, T., Rotta, L., Carmo, A., Shimabukuro, M., Gonçalves, S. & Imai, N. (2016). Estimating the CDOM absorption coefficient in tropical inland waters using OLI/Landsat-8 images. *Remote Sensing Letters*, 7(7), 661-70.

Battiti, R., Demir, B., & Bruzzone, L. (2015). Compressed histogram attribute profiles for the classification of VHR remote sensing images. *Proceedings of SPIE*, 9643, Conference on Image and Signal Processing for Remote Sensing XXI, Toulouse, France.

Bukata, R. P., Jerome, J. H., Kondratyev, K. Y., & Pozdnyakov, D. V. (1995). *Optical properties and remote sensing of inland and coastal waters*, CRC press, Boca Raton, FL, USA

Chen, J., Zhu W. N., Pang, S. N., & Cheng, Q. (2019). Applicability Evaluation of Landsat-8 for Estimating Low Concentration Colored Dissolved Organic Matter in Inland Water, *Geocarto International*, Published online, 1-15.

Chen, J., Zhu, W. N., Tian, Y. Q., & Yu, Q. (2017). Estimation of colored dissolved organic matter from Landsat-8 imagery for complex inland water: Case study of Lake Huron. *IEEE Transactions on Geosciences and Remote Sensing*, 55(4), 2201-2212.

Demir, B., & Beguem, L. (2016). Histogram-based attribute profiles for classification of very high resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4), 2096-2107.

Demirel, H., Ozcinar, C., & Anbarjafari, G. (2010). Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition. *IEEE Geoscience and remote sensing letters*, 7(2), 333-337.

Dong, Z. C., Bao, Z. Y., Wu, G. A., Fu, Y. R., & Yang, Y. (2010). Lead Concentration Distribution and Source Tracing of Urban/Suburban Aquatic Sediments in Two Typical Famous Tourist Cities: Haikou and Sanya, China. Bulletin of Environmental Contamination and Toxicology, 85(5), 509-514.

Foden, J., Sivyer, D., Mills, D. & Devlin, M. (2008). Spatial and temporal distribution of chromophoric dissolved organic matter (CDOM) fluorescence and its contribution to light attenuation in UK waterbodies. *Estuarine, Coastal and Shelf Science*, 79(4), 707-17.

Fu, X. Y., Wang, J. Y., Zeng, D. L., Huang, Y., & Ding, X. H., (2015). Remote sensing image enhancement using regularized-histogram equalization and DCT. *IEEE Geoscience and Remote Sensing Letters*, 12(11), 2301-2305.

Gonzalez, R. C., & Woods, R. E, (2017). *Digital Image Processing* (4th Ed.). Pearson, New York, NY, USA

Goodman, J. W. (2015). Statistical Optics (2nd Ed.). John Wiley & Sons, New Jersey, USA

Huang, S. Q., Li, D., Zhao, W. W., & Liu, Y., (2019). Haze removal algorithm for optical remote sensing image based on multi-scale model and histogram characteristic. *IEEE Access*, 7, 104179-104196.

IOCCG (2006). *Remote sensing of inherent optical properties: Fundamentals, tests of algorithms, and applications*. In Z. P. Lee (Ed.), Report of International Ocean-Colour Coordinating Group.

Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake, A., & Hinz, S. (2018). Hyperspectral data and machine learning

for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. *International Journal of Environmental Research and Public Health*, 15(9), 1881.

Kirk, J. T. O. (1994). *Light and photosynthesis in aquatic ecosystems* (2nd Ed.), Cambridge University Press, New York, NY, USA.

Kuhn, C., de Matos Valerio, A., Ward, N., Loken, L., Sawakuchi, H. O., Kampel, M., Richey, J., Stadler, P., Crawford, J. & Striegl, R. (2019). Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. *Remote Sensing of Environment*, 224, 104-18.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics,* 22(1), 79-86.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*. 37(1), 145–151.

Lv, Z. Y., Liu, T. F., Zhang, P. L., Benediktsson, J. A., Lei, T., & Zhang, X. K. (2019). Novel adaptive histogram trend similarity approach for land cover change detection by using bitemporal very-high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57, 9554-9574.

McKee, D. & Cunningham, A. (2006). Identification and characterization of two optical water types in the Irish Sea from in situ inherent optical properties and seawater constituents. *Estuarine, Coastal and Shelf Science*, 68(1-2), 305-16.

Mobley, C.D. (1994). *Light and Water: Radiative Transfer in Natural Waters*. Academic, San Diego.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., & McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWIFS. *Journal of Geophysical Research-Oceans*, 103(C11), 24,937-24,953.

Ohlander, R., Price, K., & Reddy, D. R. (1978). Picture Segmentation Using a Recursive Region Splitting Method. *Computer Graphics and Image Processing,* 8(3), 313-333.

Ôsterreicher, F., & Vajda I. (2003). A new class of metric divergences on probability spaces and its statistical applications. *Annals of the Institute of Statistical Mathematics*, 55(3), 639-653.

Paul, S., & Pati, U. C. (2016). Remote sensing optical image registration using modified uniform robust SIFT. *IEEE Geoscience and Remote Sensing Letters*, 13(9), 1300-1304.

Sabetta, L., Basset, A., & Spezie, G. (2008). Marine phytoplankton size-frequency distributions: Spatial patterns and decoding mechanisms. *Estuarine Coastal and Shelf Science*, 80(1), 181-192.

Schwarz, J. N. (2005). Derivation of dissolved organic carbon concentrations from SeaWiFS data. *International Journal of Remote Sensing*, 26(2), 283-93.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association,* 21 (153), 65-66.

Vanhellemont, Q., & Ruddick, K. (2014). Turbid wakes associated with offshore wind turbines observed with Landsat 8. *Remote Sensing of Environment*, 145, 105-115.

Vanhellemont, Q., & Ruddick, K. (2018). Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. *Remote Sensing of Environment*, 216, 586-597.

Zhao, J., Cao, W., Xu, Z., Ye, H., Yang, Y., Wang, G., Zhou, W. & Sun, Z. (2018). Estimation of suspended particulate matter in turbid coastal waters: application to hyperspectral satellite imagery. *Optics Express*, 26(8), 10476-93.

Zhu, W. N., Huang, L. T., Sun, N., Chen, J., & Pang, S. N. (2020). Landsat 8-observed water quality and its coupled environmental factors for urban scenery lakes: A case study of West Lake. *Water Environment Research*, 92(2), 255-265.

Zhu, W. N., Tian, Y. Q., Yu, Q., & Becker, B. L. (2013a). Using Hyperion imagery to monitor the spatial and temporal distribution of colored dissolved organic matter in estuarine and coastal regions. *Remote Sensing of Environment*, 134, 342-354.

Zhu, W. N., Yu, Q., & Tian, Y. Q. (2013b). Uncertainty analysis of remote sensing of colored dissolved organic matter: evaluations and comparisons for three rivers in North America. *ISPRS Journal of Photogrammetry and Remote Sensing*, 84, 12-22.

Zhu, W. N., Yu, Q., Tian, Y. Q., Becker, B. L., Zheng, T., Carrick, H. J., & Uzarski, D. J. (2014). An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments. *Remote Sensing of Environment*, 140, 766-778.