Supporting Information

The Estimation of Chemical Oxygen Demand of Erhai Lake Basin and Its Links with DOM Fluorescent Components Using Machine Learning

Yuquan Zhao [1,2] Jian Shen [1,2,3,*], Jimeng Feng [1,2,3], Zhitong Sun [4], Tianyang Sun [2,3], Decai Liu [2,3], Mansong Xi [5], Rui Li [6] and Xinze Wang [1,2,3,*]

1 School of Environmental Science and Engineering, Shanghai Jiao Tong University, Shanghai, China; zhaoyuquan@sjtu.edu.cn (Y.Z.); fjm520@sjtu.edu.cn (J.F.)
2 National Observation and Research Station of Erhai Lake Ecosystem in Yunnan, Dali, China; suntianyang_139872@163.com (T.S.); 595356393@qq.com (D.L.)
3 Yunnan Dali Research Institute of Shanghai Jiao Tong University, Dali, China;
4 Northwest Agriculture and Forestry University, Yangling, China; sunnyzt@nwafu.edu.cn (Z.S.)
5 Dali Erhai Lake Research Institute, Dali, China; xiweng@foxmail.com (M.X.)
6 Yunnan Institute of Water & Hydropower Engineering Investigation, Design and Research, Kunming, China
* Correspondence: Jian Shen (sjlnts@sjtu.edu.cn); Xinze Wang (xinzewang@sjtu.edu.cn)

Supporting Information contains:

DOM related index measurement, model optimization method and evaluation index calculation.

Material and Methods

DOM measurements

The measurement of the three-dimensional fluorescence spectrum was done on the instrument (RF-6000, Japan). Combined with the DOMFluor toolbox in MATLAB, according to the standard procedure, we carried out PARAFAC analysis on the EEM data [1]. After removing the Rayleigh scattering, we further removed the outliers. For the four-component model obtained, it was verified by the method of split-half analysis and random initialization. Then the four components were uploaded to the OpenFluor database (https://openfluor.lablicate.com/) for comparative analysis [2]. The fluorescence intensity of all components was expressed as the Fmax. Proportions of fluorescent components were determined by dividing each component by total fluorescence (e.g. C1% = C1/(C1 + C2 +C3 +…+ Cn) * 100, n represents the number of components selected by the researcher in the model) [3].

Fluorescence index measurement

FI was defined as the ratio of the fluorescence intensity at the emission wavelength of 470 nm and 520 nm at an excitation wavelength of 370 nm [4,5]. β/α was determined by the ratio of the fluorescence intensity at the emission wavelength of 380 nm to the maximum fluorescence intensity in the range of 420-435 nm under the excitation wavelength of 310 nm [6]. HIX was estimated as the fluorescence peak between 435-480nm divided by the sum of the fluorescence peak between 300-345nm and the fluorescence peak between 435-480nm under the excitation light of 254nm wavelength [7]. BIX was calculated as the ratio of the fluorescence intensity at the emission wavelength of 380nm and 430nm at the excitation wavelength of 310nm [8].

Model optimization

The machine learning algorithm is driven by parameters, so the GridSearchCV function from Scikit-Learn library was used to get the optimal value of each parameter [9]. The parameters and their values contained in the grid are shown in Table S2. In order to get better prediction results, cross-validation was used to reduce the error. 5-fold cross-validation (CV) was used to establish and validate the prediction results. The entire training data set was randomly divided into 5 subsets, 4 subsets were used for model training and to make predictions for the subsets not involved in training. The whole process was repeated 5 times until each subset was tested [10].

Performance indices

Three indicators are used to evaluate and compare the performance capabilities of the models. These are R-Square (R^2), root mean square error (RMSE) and residual prediction deviation (RPD), defined as follows:

$$R^2 = \frac{\sum\left(Y_{predict} - Y_{mean}\right)^2}{\sum(Y_{actual} - Y_{mean})^2} \tag{S1}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Y_{actual} - Y_{predict}\right)^2} \tag{S2}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{m}(Y_{actual} - Y_{mean})^2}{m}} \tag{S3}$$

$$RPD = \frac{SD}{RMSE} \tag{S4}$$

Where Y_{actual} is the detected CODMn data, $Y_{predict}$ is the predicted data by the models, m is the number of data and Y_{mean} is the mean of detected CODMn data.

COD and CODMn response of DOM's Fmax reduction

Fmax of each component of the DOM would be sequentially reduced (Fmax was reduced in steps of 10% until it became 0). For example, the COD reduction concentration at the 20% reduction level of C1 is calculated as the value calculated by substituting the Fmax of the original unchanged DOM into the model minus the Fmax of 20% C1 and other unchanged Fmax of C2, C3 and C4 into the model, the degree of COD reduction at the 20% reduction level of C1 is calculated as the reduction concentration divided by the value calculated by substituting the original DOM into the model. The calculation of CODMn reduction is consistent with COD.

Figure S1. The sketch map of 5-fold cross validation.

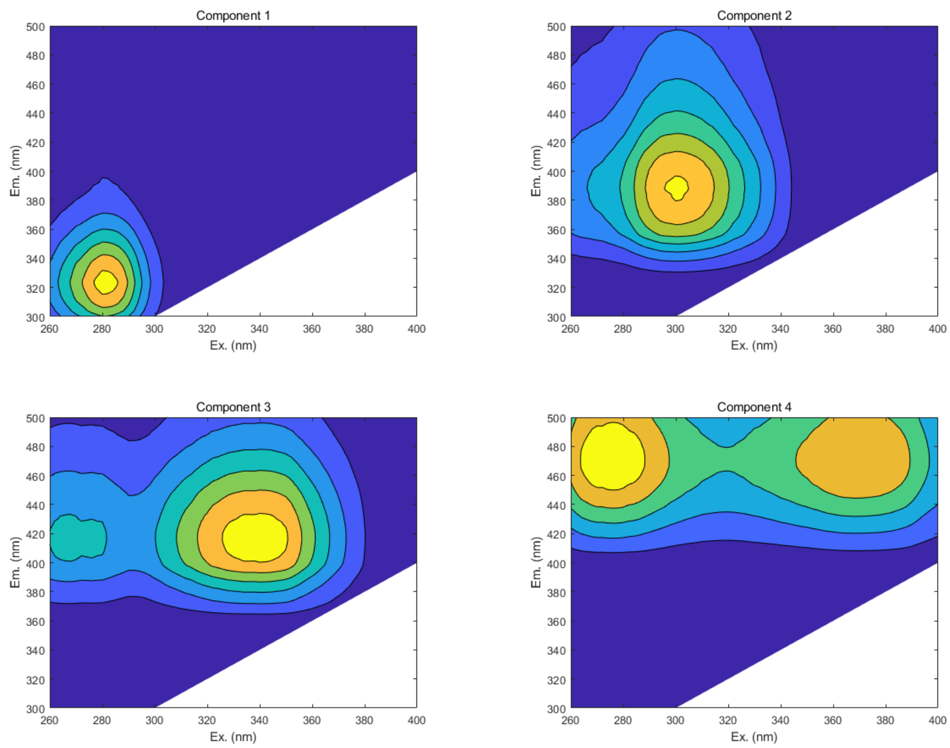

Figure S2. PARAFAC model output showing fluorescence signatures of the four DOM components.

.

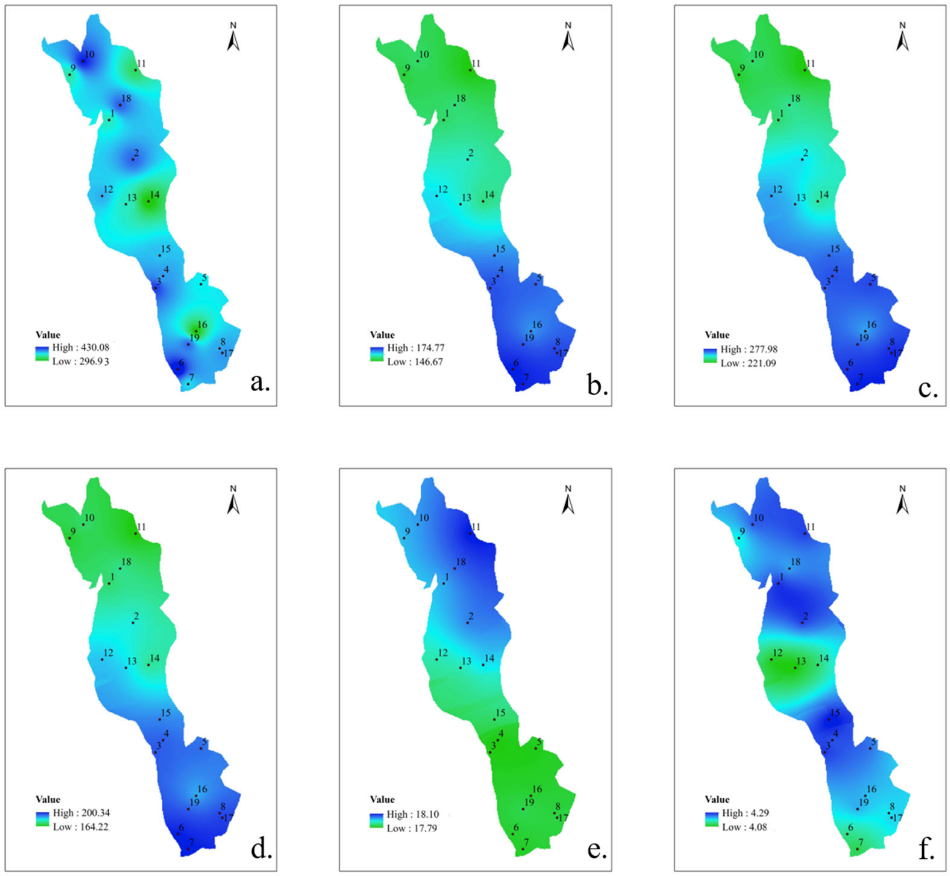

Figure S3. Spatial distribution of COD, CODMn and DOM in Summer. (**a**): C1; (**b**): C2; (**c**): C3; (**d**): C4; (**e**): COD; (f): CODMn.

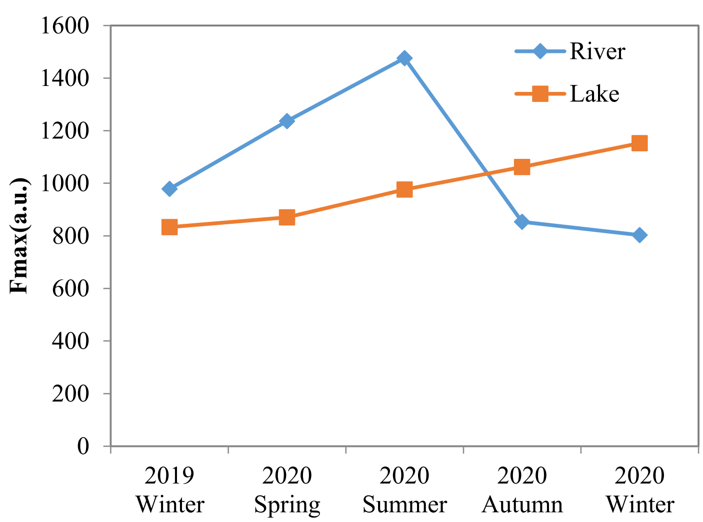

Figure S4. Sum of Fmax for C1–C4.

Table S1. Spectral parameter description.

Index	Method	Description	References
FI	λEx＝370nm, F470/520nm	The source of humus, >1.8 means DOM mainly comes from microbial metabolism and other processes, < 1.4 indicates DOM is mainly land-based input	[4, 5]
β/α	λEx＝280nm, F380/Fmax420~435nm	The proportion of newborn DOM in the overall DOM	[6]
HIX	λEx＝254nm, Fmax435~480nm/（Fmax300~345+Fmax435~480nm）	The degree of humification of organic matter, >4 means high humification degree of DOM, < 4 represents low humification degree	[7]
BIX	λEx＝310nm, F380/430nm	The ratio of microbial-derived organic matter and exogenous organic matter, > 0.8 symbolizes obvious autogenous characteristics, < 0.8 indicates that the characteristics of authigenic source are not obvious	[8]

Table S2. Values of the parameters contained in the grid search.

Parameters	Values
Bootstrap	True, False
Number of the trees	2,3,4,5,6,8,10,15,20,30,50,100
Depth of the trees	3,5,7,9,10,15,20,50
Alpha	0.1,0.5,1,5

Table S3. Characteristics of four PARAFAC components.

Component	Excaitation and emission maxima	Description
C1	Ex=278; Em=318	Protein-like, Biological production, Freshly production[11–13]
C2	Ex=296; Em=376	Humic-like, Dominates the estuarine DOM signal[14]
C3	Ex=332; Em=406	Humic-like, Prouced in the water or the coastal zooe, Related to water salinity[15–17]
C4	Ex=270(362); Em=462	Humic-like, Resembled plant-derived material[16, 18]

Table S4. One-way ANOVA of variables.

Parameter	C1	C2	C3	C4	COD	CODMn	FI	HIX
F	0.0157	0.0089	0.0234	0.0140	0.0309	0.6443	0.1365	0.3691
P-value	0.9015	0.9258	0.8797	0.9067	0.8606	0.4226	0.7119	0.5438

Using a cutoff value of $p < 0.05$.

Table S5. Seasonal Changes of COD, CODMn and DOM.

Location	Time	C1 (a.u.)	C2 (a.u.)	C3 (a.u.)	C4 (a.u.)	COD (mg/L)	CODMn (mg/L)	FI	HIX
Rivers	2019 Winter	224.65	238.40	266.07	249.48	8.38	2.77	1.74	4.79
	2020 Spring	241.79	314.92	351.86	328.03	14.17	4.12	1.73	4.94
	2020 Summer	200.73	355.75	448.28	471.09	9.21	4.06	1.67	6.62

	2020 Autumn	169.41	206.34	237.88	239.61	6.84	1.96	1.68	4.31
	2020 Winter	148.45	192.99	225.93	235.49	7.79	2.06	1.72	4.92
Erhai Lake	2019 Winter	338.61	216.36	141.94	136.42	15.55	4.13	1.63	1.97
	2020 Spring	352.91	215.92	161.20	140.56	16.56	4.10	1.72	1.88
	2020 Summer	425.26	217.70	180.75	152.19	17.52	4.24	1.73	1.79
	2020 Autumn	398.83	261.92	216.39	184.93	16.23	3.90	1.70	1.94
	2020 Winter	409.20	301.67	230.51	210.79	15.37	3.64	1.69	2.20

Table S6. Correlation coefficient of COD, CODMn and DOM in rivers.

	C1	C2	C3	C4	COD	CODMn	FI	HIX
C1	1	0.745**	0.661**	0.604**	0.746**	0.689**	0.166*	-0.375**
C2		1	0.978**	0.942**	0.730**	0.839**	0.015	0.041
C3			1	0.978**	0.691**	0.834**	0.005	0.145*
C4				1	0.651**	0.849**	-0.102	0.235**
COD					1	0.842**	0.058	-0.112
CODMn						1	-0.113	0.094
FI							1	-0.164*
HIX								1

*: *P* <.05; **: *P* <.01; and ***: *P* <.001 (*n* = 216).

Table S7. Correlation coefficient of COD, CODMn and DOM in lake.

	C1	C2	C3	C4	COD	CODMn	FI	HIX
C1	1	0.161**	0.352**	0.249**	0.241**	0.128**	0.175**	-0.238**
C2		1	0.809**	0.916**	-0.293**	-0.572**	0.031	-0.191**
C3			1	0.941**	-0.142**	-0.387**	0.299**	-0.512**
C4				1	-0.224**	-0.515**	0.129**	-0.302**
COD					1	0.220**	0.139**	-0.100*
CODMn						1	0.044	-0.051
FI							1	-0.570**
HIX								1

*: *P* <.05; **: *P* <.01; and ***: *P* <.001 (*n* = 456).

Table S8. COD response of DOM's Fmax reduction.

DOM reduction ratio	Concentration of COD reduction				Degree of COD reduction			
	C1	C2	C3	C4	C1	C2	C3	C4
10%	0.28	0.40	0.15	0.11	3%	4%	2%	1%
20%	0.57	0.73	0.22	0.23	6%	8%	3%	3%
30%	0.93	1.00	0.28	0.39	10%	11%	3%	4%
40%	1.30	1.26	0.31	0.60	15%	14%	3%	7%
50%	1.73	1.56	0.37	0.69	19%	17%	4%	8%
60%	2.19	1.80	0.40	0.80	24%	20%	4%	9%
70%	2.55	2.09	0.43	0.89	28%	23%	5%	10%
80%	2.65	2.28	0.50	1.01	30%	25%	6%	11%
90%	2.61	2.34	0.53	1.16	29%	26%	6%	13%
100%	2.60	2.34	0.53	1.29	29%	26%	6%	14%

Table S9. CODMn response of DOM's Fmax reduction.

DOM reduction ratio	Concentration of CODMn reduction				Degree of CODMn reduction			
	C1	C2	C3	C4	C1	C2	C3	C4
10%	0.06	0.07	0.03	0.08	2%	2%	1%	3%
20%	0.09	0.10	0.05	0.13	3%	3%	2%	5%

30%	0.12	0.11	0.07	0.18	4%	4%	2%	6%
40%	0.14	0.12	0.07	0.25	5%	4%	2%	8%
50%	0.14	0.15	0.12	0.35	5%	5%	4%	12%
60%	0.14	0.16	0.15	0.41	5%	5%	5%	14%
70%	0.14	0.18	0.16	0.48	5%	6%	6%	16%
80%	0.13	0.18	0.18	0.49	4%	6%	6%	17%
90%	0.12	0.18	0.18	0.47	4%	6%	6%	16%
100%	0.12	0.18	0.18	0.47	4%	6%	6%	16%

1. Stedmon, C.A. and R. Bro, *Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial.* Limnology and Oceanography: Methods, 2008. **6**(11): p. 572-579.
2. Murphy, K.R., et al., *OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment.* Analytical Methods, 2014. **6**: p. 658-661.
3. Dainard, P.G., et al., *Photobleaching of fluorescent dissolved organic matter in Beaufort Sea and North Atlantic Subtropical Gyre.* Marine Chemistry, 2015. **177**: p. 630-637.
4. Cory, R.M. and D.M. Mcknight, *Fluorescence spectroscopy reveals ubiquitous presence of oxidized and reduced quinones in dissolved organic matter.* Environmental Science and Technology, 2005. **39**(21): p. 8142-8149.
5. McKnight, D.M., et al., *Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity.* Limnology and Oceanography, 2001. **46**(1): p. 38-48.
6. Parlanti, E., et al., *Dissolved organic matter fluorescence spectroscopy as a tool to estimate biological activity in a coastal zone submitted to anthropogenic inputs.* Organic Geochemistry, 2000. **31**(12): p. 1765-1781.
7. Ohno and Tsutomu, *Fluorescence inner-filtering correction for determining the humification index of dissolved organic matter.* Environmental Science and Technology, 2002. **36**(4): p. 742-746.
8. Huguet, A., et al., *Properties of Fluorescent Dissolved Organic Matter in the Gironde Estuary.* Organic Geochemistry, 2008. **40**(6): p. 706-719.
9. Castrillo, M. and A. Lopez Garcia, *Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods.* Water Research, 2020. **172**.
10. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
11. Coble, P.G., *Characterization of marine and terrestrial DOM in seawater using excitation-emission matrix spectroscopy.* Marine Chemistry, 1996. **51**(4): p. 325-346.
12. Kim, J., et al., *Tracing water mass fractions in the deep western Indian Ocean using fluorescent dissolved organic matter.* Marine Chemistry, 2020. **218**: p. 103720.
13. Hambly, A.C., et al., *Characterising organic matter in recirculating aquaculture systems with fluorescence EEM spectroscopy.* Water Research, 2015. **83**: p. 112-120.
14. Sondergaard, M., C.A. Stedmon, and N.H. Borch, *Fate of terrigenous dissolved organic matter (DOM) in estuaries: Aggregation and bioavailability.* Ophelia, 2003. **57**(3): p. 161-176.
15. Ren, W., et al., *Characteristics of dissolved organic matter in lakes with different eutrophic levels in southeastern Hubei Province, China.* Journal of Oceanology and Limnology, 2021. **39**(4): p. 1256-1276.
16. Peleato, N.M., et al., *Investigation of ozone and peroxone impacts on natural organic matter*

character and biofiltration performance using fluorescence spectroscopy. Chemosphere, 2017. **172**: p. 225-233.

17. Liu, D., et al., *Human activities determine quantity and composition of dissolved organic matter in lakes along the Yangtze River.* Water Research, 2020. **168**.
18. Yan, C., et al., *Relationship between the characterization of natural colloids and metal elements in surface waters.* Environmental Science and Pollution Research, 2020. **27**(25): p. 31872-31883.