Quantitative estimation of wastewater quality parameters by hyperspectral band screening using GC, VIP and SPA

Zheng Xing [1, 2], **Junying Chen** [Corresp., 1, 2], **Xiao Zhao** [1, 2], **Yu Li** [2], **Xianwen Li** [2], **Zhitao Zhang** [1, 2], **Congcong Lao** [2], **Haifeng Wang** [2]

[1] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F University, Yangling, Shaanxi, China

[2] College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, Shaanxi, China

Corresponding Author: Junying Chen
Email address: cjyrose@126.com

Water pollution has been hindering the world's sustainable development. The accurate inversion of water quality parameters in sewage with visible-near infrared (VIS-NIR) spectroscopy can improve the effectiveness and rational utilization and management of water resources. However, the accuracy of spectral models of water quality parameters is usually prone to noise information and high dimensionality of spectral data. This study aimed to enhance the model accuracy through optimizing the spectral models based on the sensitive spectral intervals of different water quality parameters. To this end, six kinds of sewage water taken from a biological sewage treatment plant went through laboratory physical and chemical tests. Eight-seven samples of sewage water were obtained by adding different amount of pure water to them. The raw reflectance (R_{raw}) of the samples were collected with Analytical Spectral Devices (ASD). The $R_{raw-SNV}$ were obtained from the R_{raw} processed with the standard normal variable (SNV). Then, the sensitive spectral intervals of each of the six water quality parameters, namely, COD, BOD, NH_3-N, the total dissolved substances (TDS), total hardness (TD) and total alkalinity (TA), were selected using three different methods: gray correlation (GC), variable importance in projection (VIP) and set pair analysis (SPA). Finally, the performance of both extreme learning machine (ELM) and partial least squares regression (PLSR) was investigated based on the sensitive spectral intervals. The results demonstrated that the model accuracy based on the sensitive spectral ranges screened through different methods appeared different. The GC method had better performance in reducing the redundancy and the VIP method was better in information preservation. The SPA method could make the optimal trade-offs between information preservation and redundancy reduction and it could retain maximal spectral band intervals with good response to the inversion parameters. The accuracy of the models based on varied sensitive spectral ranges selected by the three analysis

methods was different: the GC was the highest, the SPA came next and the VIP was the lowest. On the whole, PLSR and ELM both achieved satisfying model accuracy, but the prediction accuracy of the latter was higher than the former. Great differences existed among the optimal inversion accuracy of different water quality parameters: COD, BOD and TN were very high; TA relatively high; and TDS and TH relatively low. These findings can provide a new way to optimize the spectral model of wastewater biochemical parameters and thus improve its prediction precision.

Quantitative estimation of wastewater quality parameters by hyperspectral band screening using GC,VIP and SPA

Zheng Xing[1,2] , Junying Chen[1,2], Xiao Zhao[1,2], Yu Li[1] , Xianwen Li[1], Zhitao Zhang[1],   Congcong Lao[1] and HaiFeng Wang[1]

[1] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F University, Yangling, Shaanxi, China
[2] College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, Shaanxi, China

Corresponding Author: Junying Chen, cjyrose@126.com

34 **ABSTRACT**

35 Water pollution has been hindering the world's sustainable development. The accurate inversion

36 of water quality parameters in sewage with visible-near infrared (VIS-NIR) spectroscopy can

37 improve the effectiveness and rational utilization and management of water resources. However,

38 the accuracy of spectral models of water quality parameters is usually prone to noise information

39 and high dimensionality of spectral data. This study aimed to enhance the model accuracy

40 through optimizing the spectral models based on the sensitive spectral intervals of different water

41 quality parameters. To this end, six kinds of sewage water taken from a biological sewage

42 treatment plant went through laboratory physical and chemical tests. Eight-seven samples of

43 sewage water were obtained by adding different amount of pure water to them. The raw

44 reflectance (R_{raw}) of the samples were collected with Analytical Spectral Devices(ASD). The

45 $R_{raw-SNV}$ were obtained from the R_{raw} processed with the standard normal variable (SNV). Then,

46 the sensitive spectral intervals of each of the six water quality parameters, namely, COD, BOD,

47 NH_3-N , the total dissolved substances (TDS), total hardness (TD) and total alkalinity (TA), were

48 selected using three different methods: gray correlation (GC), variable importance in projection

49 (VIP) and set pair analysis (SPA). Finally, the performance of both extreme learning machine

50 (ELM) and partial least squares regression (PLSR) was investigated based on the sensitive

51 spectral intervals. The results demonstrated that the model accuracy based on the sensitive

52 spectral ranges screened through different methods appeared different. The GC method had

53 better performance in reducing the redundancy and the VIP method was better in information

54 preservation. The SPA method could make the optimal trade-offs between information

55 preservation and redundancy reduction and it could retain maximal spectral band intervals with

56 good response to the inversion parameters. The accuracy of the models based on varied sensitive

57 spectral ranges selected by the three analysis methods was different: the GC was the highest, the

58 SPA came next and the VIP was the lowest. On the whole, PLSR and ELM both achieved

59 satisfying model accuracy, but the prediction accuracy of the latter was higher than the former.

60 Great differences existed among the optimal inversion accuracy of different water quality

61 parameters: COD, BOD and TN were very high; TA relatively high; and TDS and TH relatively

62 low. These findings can provide a new way to optimize the spectral model of wastewater

63 biochemical parameters and thus improve its prediction precision.

64

65 **Key words**: Wastewater quality parameters; Hyperspectral band screening; Gray correlation method;

66 Variable importance projection analysis; Set pair analysis; Extreme learning machine

67 **Introduction**

68   Water pollution, one of the most important causes for the shortage of the utilizable water
69   resources, has seriously threatened the sustainable development of human society. According to
70   the statistics of The 4th World Water Forum UN World Water Development Report, about 420
71   billion cubic meters of sewage is poured into rivers and lakes every year, polluting 550 million
72   cubic meters of fresh water, which is equivalent to over 14% of the total amount of global river.
73   Water quality is affected by the synergy of multiple pollutants (Brönmark & Hansson, 2002).
74   The excessive presence of these pollutant particles would make the water quality parameters go
75   beyond the current standards, so the water becomes too contaminated to use. Therefore,
76   accurately and efficiently acquiring the water quality parameters of these pollutants can improve
77   the pertinence and effectiveness of water pollution monitoring.
78   The traditional procedure of quantitatively estimating the water pollutant parameters mainly
79   includes such steps as water sampling at fixed points, physical and chemical examinations in
80   laboratory, and comprehensive statistical analysis. Such a method is not only time-consuming
81   and laborious, but also limited in observation points and representation (Esterby, 1996; Shafique
82   et al., 2003). There are more than 117 million lakes on the earth, only a very small number of
83   which are under regular and continuous monitoring (Verpoorter et al., 2014). Due to the
84   limitations of the traditional method, it is impossible to achieve dynamic monitoring of water
85   quality in large areas.
86   Different from the conventional methods, remote sensing technology can acquire large-area
87   water spectral information quickly, continuously and inexpensively and obtain by different
88   methods the multiple components information of water from remote sensing images (Bukata &
89   Bukata, 2005; Arabi et al., 2016; Campanelli et al., 2017; Deng et al., 2017). However, due to the
90   influence of spatial and temporal variation and long transmission distance, about 90-98% of the
91   signals obtained by the remote sensing are from the surface of the water and the atmosphere, and
92   the remaining 2-10% of the signals are from the water components. This has led to the
93   complexity of the optical characteristics of water, which has made it rather difficult to identify
94   the information of water pollutants, resulting in the uncertainty in the extraction results of water
95   pollutants (Gitelson & KONDRATYEV, 1991). In the water, some of the main pollutant
96   parameters (COD, BOD) change the radiation of light through their own absorption and
97   scattering characteristics of light, and thus have different characteristic absorption spectra, which
98   strongly correlates the pollutant content and spectral reflectance (Duan. et al. 2006; Cao. et al.
99   2018; Wu. et al. 2011). In recent years, hyper-spectral techniques are still being widely used for
100 various monitoring of natural resources. Nowadays, the application of hyper-spectral techniques
101 can facilitate the elimination of the external interference, obtain fine spectral information, and
102 ensure the accuracy of spectral inversion. Therefore, more researchers have deeply studied the
103 application of the hyperspectral technology in obtaining the visible and near-infrared (VIS-NIR)
104 band in order to estimate the content of some pollutants (Palmer et al., 2015; Han & Rundquist,
105 1997)

106   At present, the inversion of water quality parameters is mainly aimed at the physiological
107   parameters of the water, such as transparency and nutrients (Zhang et al., 2017), but rarely at
108   such biochemical parameters of water as its total hardness (TH), total dissolved matter (TDS),
109   biological oxygen demand (BOD)(Yang et al., 2004), chemical oxygen demand (COD)(Rojas,
110   2009), total alkalinity (TA), and ammonia nitrogen (NH_3-N)(Lerch, 2015). These water quality
111   parameters reflect the water pollution degree, which is of great significance for the purposeful
112   and effective treatment of sewage. Among the spectral inversion methods of water quality
113   parameters, the empirical statistical method remains the mainstream. The partial least squares
114   regression (PLSR) (Wold, 1966), as a typical linear regression method, is of wide use in model
115   construction for its advantage of using all available bands without multi-collinearity problem.
116   With the gradual promotion of machine learning algorithms, such advanced semi-empirical
117   machine learning algorithms (Keller et al., 2018) as artificial neural networks (Isiyaka et al.,
118   2018; Bansal & Ganesan, 2019), support vector machine (SVM), extreme learning machine
119   (ELM) have also been gradually applied in the retrieval of water quality parameters, which have
120   greatly improved the model inversion accuracy. Compared with traditional neural networks,
121   ELM calculates faster with the learning accuracy guaranteed. However, in the process of spectral
122   data modeling, there are two major problems in data processing caused by the large amount of
123   data information of hyperspectrum (Zhang et al., 2018). One problem is that between the
124   adjacent frequency bands usually has a high correlation, which brings huge redundant
125   information. Redundant information is of no use to the processing of hyperspectral data, which
126   will be a great waste of computation and storage. Another problem is that the detailed
127   information of spectrum will increase the dimension of hyperspectral data, which may lead to the
128   "Hughes phenomenon". Hughes phenomenon means the increase of data dimension will
129   decrease the classification accuracy when the number of training samples is limited (Hughes,
130   1968). Therefore, in current researches, scholars are trying to adopt some methods to screen the
131   characteristic spectrum, to remove redundant information, and to improve the stability and
132   predictability of the model (Feng et al., 2017). The variable importance projection (VIP)
133   algorithm has successfully distinguished the hyperspectral band subsets of forest species by
134   identifying the importance of the independent variables (Peerbhay et al., 2013); The gray
135   correlation (GC) method, using the gray values of the independent variables, has achieved the
136   screening of spectral intervals of soil ion characteristics (Wang et al., 2019). So far these two
137   methods have rarely been applied in the screening of the characteristic spectrum of water
138   pollution parameters. Meanwhile, these band screening methods have different preferences and
139   significant differences in the performance of different modeling methods, which will lead to the
140   uncertainty in the model accuracy (Lerch, 2015). Iterative progressive elimination combined
141   with PLSR has a good performance in the inversion of chlorophyll A and total suspended solids
142   (Wang et al., 2018). VIP combined with PLSR and SVR has significantly improved the inversion
143   effect of characteristic spectra of water-soluble ions in soil, while GC combined with PLSR and
144   SVR showed relatively poor inversion effect (Wang et al., 2019). However, GC combined with
145   ELM has a good effect on the prediction and inversion of total nitrogen content in soil (Zhou et
146   al., 2017). Moreover, in order for the better universality and stability of band selection method,

147 the set pair analysis (SPA) theory was used to establish models which can identify the overall
148 degree as well as the important indicators and subsystems in the influencing factors (Zhang et al.,
149 2019). SPA theory, for its remarkable capability of dealing with uncertain problems, has been
150 widely applied in many fields (Wang et al., 2017; Li et al., 2016), but not yet applied in band
151 selection. This research applied the band selection methods of GC, VIP and SPA to select from
152 hyperspectral data the feature subsets, which were used to establish PLSR and ELM models to
153 estimate such water quality parameters as COD, BOD, NH$_3$-N TDS TH and TA, and to compare
154 and analyze the inversion effects of different pollutant indicators.

155   Particularly, the study aims to: (1) determine the optimal spectral interval corresponding to the
156 six water pollution indicators, build the optimal inversion models of six water pollution
157 parameters, and achieve quantitative estimation of water pollution parameters through
158 hyperspectrum; (2) verify the SPA method in weighing the different wavelength screening
159 methods and provide an approach to band selection by applying the set optimization idea; (3)
160 determine the corresponding optimal models for different water quality parameters by comparing
161 the performance of GC, VIP and SPA models.

MATERIALS AND METHODS

Sample preparation and chemical analysis

164   The water samples were taken from different spots (water inlet, anoxic tank, aerobic tank,
165 sedimentation tank and water outlet) under different treatment methods at a domestic sewage
166 treatment plant in China. The domestic sewage treatment plant gave field permit approval to us
167 (NO.51409221, 51979234). According to our pre-analysis (Table 1), the water samples collected
168 at the inlet were abundantly rich in COD, BOD and NH3-N (COD: 425mg/L, BOD: 101.15mg/L,
169 NH3-N: 34.853mg/L), whilst the water samples collected in other treatment ponds were
170 relatively low in COD of 20~140mg/L, BOD of 3.1~13mg/L and NH3-N of 0~1.723mg/L. Such
171 large variations suited well to represent the range of water quality collected from different
172 locations investigated in this study, but also leaving non-negligible deficiency in representing
173 water quality of moderate concentrations. Therefore, to make the hyperspectral model more
174 robust and responsive to minor changes of water quality at other potential locations, we further
175 diluted the water samples into subsamples to form a full range of distribution within the
176 minimum and maximum values detected in this study. To be specific, following the minimum
177 and maximum COD measured from the inlet and other treatment ponds, the water samples rich
178 in COD were diluted by adding different amount of pure water, forming different COD
179 concentrations varying at unit of 5 mg/L. In total, 87 water samples were obtained to establish
180 the hyperspectral model and strengthen its applicability. In a real application of hyperspectral
181 model, the water quality of targeted samples often covers a variety of concentrations, and thus
182 can be measured and applied directly without any dilution. The laboratory was kept in a constant

183 temperature during the experiments. The main water quality parameters included NH_3-N, TA,
184 TH, TDS, COD and BOD.

Acquisition and pretreatment of spectral data

186 The wastewater samples were put into black cylindrical cups with a depth of 5cm and a
187 diameter of 10cm for spectral data collection in the laboratory. The hyperspectral data for
188 wastewater samples were measured using an ASD(Analytical Spectral Devices, Inc. Boulder,
189 CO, USA)FieldSpec® 3 spectrometer with spectral range from 350 to 2500 nm. The
190 instrument was equipped with one sensor with 1.4 nm spectral resolution for the range of 350-
191 1000 nm and the other sensor 2 nm, 1000-2500nm. The spectral data were collected in a dark
192 room by exposing the water sample surfaces to a halogen lamp of 50 W above with a 30° of
193 incident angle and 50cm in distance and a fiber-optics probe 5°, 15cm, so as to minimize the
194 effects of external factors (Wang & Zhao, 2000). Before each measurement, we had fully
195 preheated the spectrometer and light source and checked the spectrometer with a standardized
196 white panel of 99% reflectance in order to reduce measurement errors. Each water sample was
197 measured twice vertically, and at each of which the spectral data were gathered ten times.
198 Altogether there were 20 spectrum curves for each sample (Hong et al., 2018). From the 20
199 curves the raw spectral reflectance (R_{raw}), namely, arithmetic mean, were obtained by using
200 ViewSpecPro software (6.0 version). The fluctuation of R_{raw} would influence the accuracy of
201 subsequent modeling due to the disturbance of the random error, instrument noise and external
202 environment in spectral data collection. In general, the external noise can be eliminated to some
203 degree with such effective pretreatments as resampling, smoothing and transformation, which
204 can improve the spectral characteristics (Ding et al., 2018). Therefore, two steps were adopted to
205 pretreat the R_{raw}: (1) removing the marginal wavelengths (2,401–2,500nm and 350–399nm) of
206 higher noise in each water sample, then smoothing the remaining spectrum data through filter
207 method (polynomial order = 2; window size = 5) of Savitzky-Golay (SG) (Savitzky&Golay,1964)
208 by Origin Pro software (2017SR2 version); and (2) obtaining the precise $R_{raw–SNV}$ using the
209 standard normal variable (SNV) to remove the effects of baseline shift and surface scattering on
210 the spectral data (Xiao et al., 2016). The spectral curves of R_{raw} and $R_{raw–SNV}$ are shown in Figs.
211 1A and 1B. As can be seen, the spectral curve in Fig. 1B is much smoother than that in Fig. 1A.

Wavelength selection methods

Gray correlation (GC)

214 The GC method, one of the grey system theories, seeks the primary and secondary relations and
215 analyzes the different effects of all the factors in a system (Ju-Long, 1982; Liu, Yang & Wu,
216 2015).The gray correlation degree (GCD), varying from one moment and object to the other,
217 refers to the measurement of the factor correlation between two systems. In the process of system
218 development, higher consistency of change trend of the two factors reflects higher correlation

219   degree and vice versa (Zhu et al., 2017). The main calculation steps are: first, determining
220   signature and factor sequences with the former marked as $x_0(t)$ which is a collection of m data--
221   $x_0(t)=\{x_0(1),x_0(2),..., x_0(m)\}$ and the latter marked as $x_i(t)$ which contains n sub-sequences with
222   each as a collection of m data-- $x_i(t)=\{ x_i(1),x_i(2),..., x_i(m)\}$; and second, processing the data
223   sequences dimensionlessly using averaging operators. $X_i=\{ x_i(1),x_i(2),..., x_i(m)\}$ represents the
224   behavior sequence of factor x_i and D the averaging sequence operator, and thus $X_iD=\{ x_i(1)d, x_i$
225   $(2)d,..., x_i(m)d\}$. So the mean data value is calculated as

226
$$x_i(k)d = \frac{x_i(k)}{\overline{X_i}}, \quad \overline{X_i} = \frac{1}{n}\sum_{k=1}^{n} x_i(k) \quad k=1,2,..,n \tag{1}$$

227
$$\xi_{0i}(t) = \frac{\Delta_{min} + \rho\Delta_{max}}{\Delta_{0i(t)} + \rho\Delta_{max}} \tag{2}$$

228   where Δ_{max} is the maximum of $|x_0(t)-x_i(t)|$; Δ_{min} is the minimum of $|x_0(t)-x_i(t)|$; $|x_0(t)-x_i(t)|$ is the

229   value at time t; ρ is the coefficient of discrimination. Therefore, GCD between the signature and

230   factor sequences is calculated as

$$\gamma_{0i} = \frac{1}{n}\sum_{t=1}^{n} \xi_{0i}(t)$$
231   (3)

232   From the interaction between the signature and factor sequences, the factors' primary and
233   secondary influences can be predicted (Wang et al., 2018). In this paper, the higher the GCD of a
234   certain band is, the more sensitive the band is to the water quality parameter, and vice versa.

235   *Variable importance projection*

236   The VIP is a variable screening analysis method based on PLSR model. The VIP value of a
237   spectral variable reflects the importance of this variable in the prediction of the substance to be
238   measured (Oussama et al., 2012;de Almeida et al., 2013). The VIP value reflects the explanatory
239   power of independent variable over dependent variable, and represents the importance of
240   independent variable to model fitting (Chavana-Bryant et al., 2019). If the explanatory ability of
241   each variable to Y is the same, then the VIP value of all independent variables is 1. If the VIP
242   value of an independent variable is less than 1, it means that the variable makes little
243   contribution to the model and has a very low ability to explain the dependent variable, so it can
244   be eliminated. In the PLSR model, VIP_i to an independent variable X_i is defined as:

245
$$VIP_i = \sqrt{\frac{p*\sum_{h=1}^{m} Rd(Y;t_h)W_{hj}^2}{Rd(Y;t_1,t_2,\cdots,t_m)}} \tag{4}$$

246   where p is the dimension of independent variable; t_h is the component h; m is the number of
247   components selected in the model; W_{hj} is the component of X_j corresponding to axis W_h; Rd (Y;
248   T_h) represents the variation of Y explained by the component t_h; and Rd (Y; t_1, t_2, $\cdots t_m$) represents
249   the cumulative variation accuracy of Y explained by $t_1\sim t_m$. In PLSR model, the explanatory
250   power of x_j over Y is passed through t_h, so if t_h bears a strong explanatory power over Y, and

251   meanwhile x_j plays a very important role in the process of t_h construction, t_h can be reasonably

252   regarded as very important variable in explaining Y. In this way of wavelengths selection, the

253   wavelengths with strong explanatory power are reserved while those with weak explanatory

254   power are eliminated (Chemura et al., 2017).

255   ***Set pair analysis (SPA)***

256   SPA integrates uncertainty analysis and determinacy analysis (Zhao, 1996). The basis of SPA is

257   set pair whose key is correlation degree, which has been applied as a criterion for the analysis of

258   certainty and uncertainty (Zhang et al., 2019). During the hyperspectral band selection, the GC

259   value $Y_i = (y_1,\ y_2,\cdots,y_n)$ can be defined as set A, and VIP value $Z_i = (z_1,z_2,\cdots,z_n)$ as set B, and

260   then the two sets compose a set pair H = (A, B). To study the correlation degree of the set pair,

261   the formulas are as follows:

262
$$\mu_i = S_i + F_i * I_i + P_i * J = \frac{S_i}{n} + \frac{f_i}{n} I_i + \frac{p_i}{n} J \tag{5}$$

263
$$I_i = \frac{S_i - P_i}{S_i + P_i} \tag{6}$$

264
$$v_i = 1/n + 1/n * u_i \tag{7}$$

265
$$W_i = \frac{v_i}{\sum_i^N v_i} (i = 1, 2, \dots, N) \tag{8}$$

266   *where* S_i, F_i, P_i are the identity degree, discrepancy degree and opposite degree of the two sets in

267   the context of the same problem, respectively. They describe the association of the two sets from

268   different aspects. S_i, F_i, P_i meet the relationship: $S_i + F_i + P_i = 1$, n denotes the total number of

269   characteristics of the set pair; s_i is the number of the common characteristics of the two sets; f_i is

270   the number of the different characteristics of the two sets (Li et al., 2016); p_i is the number of the

271   opposite characteristics of the two sets. When $J = -1$,the contact number u_i, the relative

272   membership degree v_i, and the weighted value, W_i , of the VIP value and GC value, were

273   calculated.

274   **Model construction and validation**

275   **Classification of modeling and validation sets**

276   Kennard-Stone algorithm (K-S) (Kennard & Stone, 1969) was used to classify the sample sets to

277   select representative calibration sample sets. K-S was used to calculate the minimum Euclidean

278   distance of the unselected samples according to the selected samples. Then the samples with the

279   maximum Euclidean distance were selected into the calibration set in a repeated way until the

280   samples of a specified number were selected. K-S algorithm has turned out to be effective in

281   selecting representative samples (Morais et al., 2019). Two-third of the water quality samples

282   were selected for the modeling set (n=58) and one-third for the validation set (n=29).

283 *Modeling methods and evaluation*

284 The PLSR and ELM methods were applied to the quantitative inversion of different water quality

285 parameters in this paper. The PLSR model has been widely applied and can overcome the

286 multicollinearity of independent variables because of its dimensionality reduction, information

287 integration and band optimization in the modeling process, which has greatly improved the

288 ability of the system to extract principal components (Ryan & Ali, 2016). The ELM is a machine

289 learning algorithm based on feed forward neural network. It has the advantages of high learning

290 efficiency, high accuracy and simple parameter adjustment over the traditional feed forward

291 neural network (Song et al., 2013). The optimal model was selected to inverse the water quality

292 parameters through comparing the models' root mean square error of calibration (RMSEC),

293 determination of coefficients (R^2_c), root mean square error of prediction (RMSEP), prediction

294 determination of coefficients (R^2_p) and relative prediction deviation (RPD). R^2_c and R^2_p are used

295 to evaluate the stability degree of a model. The closer its value is to 1, the higher stability the

296 model has. RMSEC and RMSEP represent the accuracy of the model. The smaller their value is,

297 the higher the model accuracy is. A model is considered excellent when RPD>3.0, very good

298 when 2.5<RPD≤3, good when 2<RPD≤2.5, satisfactory when 1.5 < RPD ≤ 2 and poor when

299 RPD≤1.5(Williams, 2001) .

RESULT AND ANALYSIS

301 **Correlation between water quality parameters content and spectral reflectance**

302 The Pearson correlation coefficients between each water quality parameter and the $R_{raw–SNV}$

303 (400–2400nm) were tested with the significance level (P < 0.001, |r|=0. 324 or above). The

304 curves of correlation coefficients of water quality parameters were plotted in Fig. 2.

305 The change of the correlation coefficients of water quality parameters was complex. There

306 were significant differences in the correlation coefficients between different water quality

307 parameters and the wavelengths but their curve patterns were similar (Peng et al., 2018). There

308 were many concentrated similar sensitive ranges but in general, the curves took "sharp fall,

309 fluctuation, and oscillation" (Fig. 2). The COD, BOD, NH_3-N and TA were all highly correlated

310 and represented a positive correlation (the correlation coefficients > 0.6) at the wavelengths of

311 900~1300nm and 1500~1800nm. In contrast, the correlation between the total dissolved

312 substances and spectral data and that between the TH and spectral data were not so satisfied (the

313 correlation coefficients < 0.6). The six parameters could pass the test of significance (P=0.001) at

314 the wavelengths of 400~450nm and 850~1800nm. The band range was wide and the peak value

315 of the correlation coefficients was relatively flat, so it was difficult to screen the characteristic

316   bands based on the correlation coefficients.

317   **Selection of characteristic wavelength**

318   *GC-based selection of Characteristic wavelength*

319   The curves of GCD for water quality parameters and $R_{raw-SNV}$ are shown in Fig. 3. An obvious
320   peak wavelength appeared in the curves with the similar patterns of "sharp rise, fall and
321   stabilization" on the whole (Fig. 3). The gray value of each water quality parameter gradually
322   went up from 400nm to a peak at about 820nm, and then gradually down to 1000nm where it
323   remained flat. The overall gray values of TDS and TH were relatively high, generally above 0.5,
324   while those of COD, BOD, NH_3-N and TA were generally between 0.3 and 0.4 with only a few
325   peak bands greater than 0.4 and above 0.8 at the most. It reflected that the gray correlation
326   method excels in eliminating a great amount of redundant information and selecting a few most
327   sensitive bands.

328   The sensitive bands of the six parameters were counted based on their GCD analysis (Table 2).
329   Of all the GCD values, the maximum ones were concentrated within the spectral wavelengths of
330   815~830nm. The numbers of the sensitive bands were sequenced from large to small as: TH (601)
331   > TDS (381) > TA (93) > BOD (50) > COD (49) > NH_3-N (46), and those of the maximum GCD
332   values as: TH (0.8504) > TDS (0.802) > BOD (0.7974) > NH_3-N (0.7973) > COD (0.7949) > TA
333   (0.7878). The comparison between the correlation coefficients and the GCD indicated a great
334   difference: the higher correlation coefficient the parameters had, the lower the GCD values was,
335   and the smaller number of sensitive bands were selected. This result revealed that for the
336   parameters with strong spectral response the GCD method had better band screening ability.

337   *VIP-based selection of Characteristic wavelength*

338   The curves of VIP scores for water quality parameters and $R_{raw-SNV}$ are shown in Fig. 4. The
339   curves patterns of the six parameters were similar and the overall scores of the VIP were
340   relatively high (Fig. 4). These curves exhibited a sharp drop in the wavelength intervals of
341   400~450nm, a sharp rise in 460~500nm, a violent fluctuation in 700~1000nm, but in
342   1100~2000nm TH and TDS displayed a less violent fluctuation and COD, BOD, NH_3-N and TA
343   showed flat fluctuation. Many peak points existed in the VIP curves, and their peak intervals
344   were relatively scattered, completely different from the single peak value in the GC curves. The
345   principle of VIP > 1 was used to select and count the spectral sensitive bands of the six
346   parameters (Table 3). The comparison of the maximum spectral response bands showed five
347   parameters were found in the wavelength intervals of 460~480nm except for NH_3-N in
348   990~999nm.

349   The numbers of sensitive bands selected through VIP method were sequenced from large to
350   small as BOD (770) > NH_3-N (768) > COD (753) > TA (709) > TH (543) > TDS (497), and
351   those of the maximum VIP scores as: TH (4.893) > TDS (4.275) > TA (2.466) > COD (1.634) >

352 BOD (1.439) > NH$_3$-N (1.397). The comparison between the two sequences indicated a great
353 difference: the larger the VIP score was, the smaller number of selected sensitive bands was. On
354 the whole, the effective bands could be retained as many as possible with the VIP method during
355 bands screening.

SPA-based selection of Characteristic wavelength

357 The curves of SPA scores for water quality parameters and $R_{raw-SNV}$ are shown in Fig.5. The
358 curves patterns of TDS and TH were similar and exhibited a gentle flatness except for obvious
359 peaks at the wavelengths of about 460nm and 990nm. Those of BOD, NH$_3$-N and TA were
360 similar and showed notable valleys at the wavelengths of about 500nm and 700nm and notable
361 peaks at the wavelengths of about 760nm and 1000nm. Those of COD displayed notable peaks at
362 the wavelengths of about 460nm and 1000nm and fluctuation at other wavelengths. On the whole,
363 great difference existed among the six curves, and the sensitive interval of each parameter was
364 quite notable and different from each other.

365   The principles of GCD>0.5 and VIP > 1 were used to select and count the spectral sensitive
366 bands of the six parameters (Table 4). The comparison of the maximum spectral response bands
367 revealed NH$_3$-N was in the wavelength intervals of 990~999nm, BOD in 760~770nm and the
368 other four in 460~475nm. The numbers of sensitive bands selected through SPA method were
369 arranged in a descending order as: BOD (767) > NH$_3$-N (765) > COD (750) > TA (696) > TDS
370 (280) >TH (223), the scores of the maximum SPA as: TA (2.409) >TH (1.944) > TDS (1.928) >
371 COD (1.598) > BOD (1.409) > NH$_3$-N (1.364). The comparison between the two orders
372 indicated a great difference: the larger the SPA score was, the smaller number of the selected
373 sensitive bands had.

Construction and analysis of PLSR model

375 The sensitive bands selected by different methods as GC, VIP and SPA were applied to PLSR
376 modeling. The results of PLSR model are shown in Table 5.

377   In general, the PLSR model had a good spectral prediction of water quality parameters (Wang
378 et al,. 2019). The GC-PLSR model was proved to be optimal for COD, BOD, NH3-N and TH
379 with the best prediction. SPA-PLSR model was proved to be optimal for TDS and TA with the
380 best prediction (Table 5). The VIP-PLSR model of BOD had the best modeling effect (R^2_c =
381 0.99). The GC-PLSR model of NH$_3$-N had the best prediction effect (R^2_p = 0.962, RPD = 5.894).
382 The inversion effects of TDS and TH were relatively less satisfying, with the R^2_c and R^2_p
383 generally around 0.8. Overall, the PLSR models based on the characteristic bands of the six
384 parameters exhibited good modeling and prediction effect as well as a greater improvement than
385 those full band-based models.

Construction and analysis of ELM model

387 The sensitive bands selected using GC, VIP and SPA methods were applied to build ELM model.

388   The results of ELM model are shown in Table 6.

389   The ELM models based on the characteristic bands of the six parameters exhibited good

390   modeling and prediction effect as well as a greater improvement than those based on the full

391   band. The GC-ELM model was proved optimal for COD, BOD and NH$_3$-N with the best

392   prediction. SPA-ELM was proved optimal for TDS, TA and TH with the best prediction. The

393   VIP-ELM model of BOD had the best modeling effect (R^2_c = 0.986). The GC-ELM model of

394   NH$_3$-N exhibited the best prediction effect (R^2_p=0.976, RPD = 6.596). The ELM models of COD,

395   TH and TA had the satisfying modeling and validation effect. The validation effect of TDS was

396   relatively less satisfying (R^2_c = 0.820, R^2_p = 0.790, RPD = 2.198).

DISCUSSION

Comparison among the estimating results of different water quality parameters

399   The optimal wavelength selection methods varied when the optimal modeling methods were

400   different (Tables 5 and 6). The hyper-spectrally estimated value and the chemically measured

401   value of the six water quality parameters were compared under the optimal model (Fig. 6).

402   Different water components can change the spectral radiation by their own absorption and

403   scattering characteristics of light. The water components (such as SPM, CDOM, phytoplankton,

404   blue-green algae) can be obtained by separating the factors of various parameters from the total

405   water radiation (Dörnhöfer & Oppelt, 2016). However, the problem of "lower ion content and

406   weaker spectral response" often exists in the spectral analysis of water components, and was also

407   reflected in the verification results of this study. All the six parameters produced satisfying

408   inversion results, which indicated the feasibility of quantitative inversion of sewage water quality

409   parameters by hyperspectrum (Pu et al., 2017). The sequence of the predicting power of the

410   water quality parameters was NH$_3$-N > BOD > COD > TA > TH > TDS. The validation results

411   showed that most data points of the five water parameters, NH$_3$-N, BOD, COD, TA and TH,

412   were concentrated near line 1:1. This indicated that the hyperspectral analysis values of the five

413   parameters were very close to the chemically estimated values and the optimal models of these

414   five parameters had strong prediction ability (RPD>3.0) (Tables 5 and 6). Comparatively, due to

415   its lower content, the data points of TDS were relatively discrete, indicating that the model

416   prediction ability was average (RPD = 2.198).

Correlation analysis and inversion performance

418   The raw spectral reflectance curve of each water sample exhibited different shapes (Fig. 2A).

419   This difference results from the inconsistency between the contents and types of the water

420   quality parameters, and the different characteristics of spectrum obtained through absorption and

421   scattering of light. The results accord with those in previous studies (Abdelmalik, 2018; Abd-

422   Elrahman et al., 2011; Rostom et al., 2017; Wang et al., 2016), which demonstrates that the VIS-

423   NIR spectrum can be used to quantitatively determine water quality parameters.

424   Traditionally, correlation analysis is helpful to reveal the relationship between water quality

425   parameters and the VIS-NIR spectrum (Peng et al., 2018). In this paper, the numbers of the

426   significant wavebands of the six water quality parameters can be arranged in an order, in which

427   the numbers of NH_3-N, BOD and COD were close to one another but larger than those of the

428   other three parameters. The number of TA followed, and those of TH and TDS came last. This

429   order agreed with that of their correlation coefficient ranges. Therefore, significant correlations

430   exist between NH_3-N, BOD, COD and their reflectance spectrum, which reflects the optimal

431   models of the three parameters have a very good prediction performance (Fig. 6). By contrast,

432   the correlations between the other three water quality parameters and their reflectance spectrum

433   were relatively low, which led to poor model prediction performance, especially for the TDS

434   model. However, by comparing the three screening methods, the numbers of the sensitive

435   wavebands of TA, TH and TDS were not as small as had been expected (Tables 3, 4, 5). This is

436   because different waveband screening methods use different calculation mechanisms, which

437   leads to different methods to screen optimal response spectrum for different parameters (Zhang

438   et al., 2018). From the perspective of modeling methods, ELM model performed better than

439   PLSR model, which also indicated the superiority and strong learning ability of machine learning

440   algorithm (Keller S et al. 2018). From the perspective of band selection methods, GC and SPA

441   are superior to VIP. Therefore, the selection of band screening methods will be of practical

442   significance for the quantitative inversion of water quality parameters in the future research.

443   **Effects of wavelength selection on hyperspectral estimation models**

444   In the process of hyperspectral analysis, due to the mass of hyperspectral information, there

445   exists a large amount of redundant information irrelevant to parameters content. This makes it

446   difficult to mine and extract effective information. This problem complicates the hyperspectral

447   analysis and decrease the analysis accuracy, which hinders the development and utilization of

448   related testing equipment. Therefore, characteristic selection of hyperspectral data information is

449   required. The characteristic selection methods are devised to find the smallest subset from the

450   original characteristics without reducing the accuracy (Zhang et al., 2018). For the hyperspectral

451   band selection, there are two key aspects, namely, effective information preservation and

452   redundancy elimination, which has a great impact on subsequent modeling and validation.

453   In this study, GC and VIP methods were used to explore their spectral characteristic response

454   intervals for six main water quality parameters in wastewater. These two methods had their fixed

455   preferences in terms of information preservation and redundancy elimination. Therefore, the two

456   methods had quite different performance in the band selection for the parameters and modeling

457   methods, and their performance was very unstable (Wang et al., 2019). Of the two methods, the

458   GC method had better performance in redundancy elimination. Better spectral response of the

459 water quality parameters means smaller number of bands filtered by the GC method, so the GC
460 method is the best screening method for COD, BOD and NH$_3$-N with a satisfying screening
461 performance in this study.

462   However, the VIP method performed better in information preservation, and it could retain as
463 many spectral band intervals with good response to the inversion parameters as possible
464 (Oussama et al., 2012; de Almeida et al., 2013). So, the VIP method retained the largest number
465 of bands during band selection. Meanwhile, compared with the GC method, VIP was more
466 applicable to characteristic band selection of TA and TDS.

467   As to this condition, some researchers considered it necessary to make the optimal trade-offs
468 between information preservation and redundancy elimination according to the different
469 characteristics of the datasets, and have conducted some explorations (Zhang et al., 2018). This
470 study put forward a band screening method based on SPA to optimize the trade-offs between GC
471 and VIP. From its application to the datasets in this study, the SPA method could well balance
472 effective information retaining and redundant information eliminating. Compared with GC and
473 VIP, the SPA method had a more stable ability of band screening. It had a strong applicability to
474 predicting the six water quality parameters as well as building the linear modeling method PLSR
475 and the nonlinear modeling method ELM. Therefore, the SPA method had the superiority in
476 certainty and uncertainty analysis of characteristic bands, and it is applicable to the integrative
477 optimization of GC and VIP.

478 **Research limitations and future research**

479 This study was only the statistical application of characteristic band screening methods to the
480 inversion of different water quality parameters and failed to explore the spectral response
481 mechanism of the various water quality parameters (Hu & Wang, 2017), which would lead to
482 some limitations in the applicability of the model. Meanwhile, the optimized band screening
483 method based on the SPA needs further study, which can focus on further:

484 (1) exploring the different spectral sensitive intervals and analyzing the spectral response
485 mechanism of each water quality parameter.

486 (2) probing into the band screening mechanism and optimizing various screening methods so as
487 to propose a band screening method suitable for various models; and applying various screening
488 methods at the same time in the longitudinal study to achieve multiple screening.

CONCLUSIONS

490 This paper studied the feasibility of quantitatively estimating water quality parameters via VIS-
491 NIR spectral model. The GC, VIP and SPA methods were used to screen the sensitive intervals
492 based on the spectral responses of different water quality parameters in sewage. The inversion
493 accuracy of the six water quality parameters by the linear PLSR and nonlinear ELM modeling

494 methods was compared, and finally the research arrived at the following conclusions:

495 1. Band screening plays an important role in spectral data processing and spectral model

496 accuracy improvement. In this study, the GC had the best performance in redundancy

497 elimination (up to 97%), the VIP performed best in information preservation, and the SPA could

498 make the optimal trade-offs between information preservation and redundancy reduction, so the

499 SPA had the best applicability.

500 2. Both the PLSR and ELM models have good performance in modeling and validation. The

501 PLSR model was more applicable and the ELM model had higher prediction accuracy. Among

502 the six corresponding optimal models for the six water quality parameters, five were ELM

503 models while only one was PLSR model.

504 3. The optimal spectral inversion models of the six water quality parameters had quite different

505 validation results: the prediction of the COD, BOD, NH_3-N and TA models was quite satisfying

506 (their RPD values are 4.69, 6.192, 6.596 and 3.651, respectively), while that of the TH and TDS

507 was relatively poor.

ACKNOWLEDGEMENTS

509 The authors want to thank A.P. Ying Meng and Yinwen Chen for their help in language

510 standardization of this manuscript and providing helpful suggestions. We are especially grateful

511 to the reviewers and editors for appraising our manuscript and for offering instructive comments.

REFERENCE

513 Abdelmalik, K.W. 2018. Role of statistical remote sensing for Inland water quality parameters prediction. *The*

514 *Egyptian Journal of Remote Sensing and Space Science* 21:193-200. 10.1016/j.ejrs.2016.12.002

515 Abd-Elrahman, A., Croxton, M., Pande-Chettri, R., Toor, G.S., Smith, S., and Hill, J. 2011. In situ estimation of

516 water quality parameters in freshwater aquaculture ponds using hyperspectral imaging system. *ISPRS JOURNAL OF*

517 *PHOTOGRAMMETRY AND REMOTE SENSING* 66:463-472. 10.1016/j.isprsjprs.2011.02.005

518 Arabi, B., Salama, M., Wernand, M., and Verhoef, W. 2016a. MOD2SEA: A Coupled Atmosphere-Hydro-Optical

519 Model for the Retrieval of Chlorophyll-a from Remote Sensing Observations in Complex Turbid Waters. *Remote*

520 *Sensing* 8:722. 10.3390/rs8090722

521 Arabi, B., Salama, M., Wernand, M., and Verhoef, W. 2016b. MOD2SEA: A Coupled Atmosphere-Hydro-Optical

522 Model for the Retrieval of Chlorophyll-a from Remote Sensing Observations in Complex Turbid Waters. *Remote*

523 *Sensing* 8:722. 10.3390/rs8090722

524 Bansal, S., and Ganesan, G. 2019. Advanced Evaluation Methodology for Water Quality Assessment Using

525 Artificial Neural Network Approach. *WATER RESOURCES MANAGEMENT* 33:3127-3141. 10.1007/s11269-019-

526 02289-6

527 Brönmark, C., and Hansson, L. 2002. Environmental issues in lakes and ponds: current state and perspectives.

528 *ENVIRONMENTAL CONSERVATION* 29:290-307. 10.1017/S0376892902000218

529 Bukata, R.P., and Bukata, R.P. 2005. Satellite monitoring of inland and coastal water quality: retrospection,

530 introspection, future directions. *Crc Taylor & Francis*.

531   Campanelli, A., Pascucci, S., Betti, M., Grilli, F., Marini, M., Pignatti, S., and Guicciardi, S. 2017. An Empirical
532   Ocean Colour Algorithm for Estimating the Contribution of Coloured Dissolved Organic Matter in North-Central
533   Western Adriatic Sea. *Remote Sensing* 9:180. 10.3390/rs9020180

534   Cao Y, Ye Y., Zhang X., Shi Y., Jiang Y., and Zhao h. 2015. Turbidity hyperspectral inversion model of nansihu
535   water body. Water diversion from south to north and water conservancy technology   13:883-887 (in Chinses).

536   CESP. 2002. Analysis methods for water and wastewater monitoring . 4th ed., Beijing: China Environmental
537   Science Press(in Chinese).

538   Chavana-Bryant, C., Malhi, Y., Anastasiou, A., Enquist, B.J., Cosio, E.G., Keenan, T.F., and Gerard, F.F. 2019.
539   Leaf age effects on the spectral predictability of leaf traits in Amazonian canopy trees. *SCIENCE OF THE TOTAL*
540   *ENVIRONMENT* 666:1301-1315. 10.1016/j.scitotenv.2019.01.379

541   Chemura, A., Mutanga, O., and Dube, T. 2017. Separability of coffee leaf rust infection levels with machine
542   learning methods at Sentinel-2 MSI spectral resolutions. *PRECISION AGRICULTURE* 18:859-881.
543   10.1007/s11119-016-9495-0

544   Duan H., Wen Y, Zhang B, Song K., and Wang Z.. 2006. Quantitative inversion of water quality parameters in
545   chagan lake using hyperspectral data. *Journal of Arid Land Resources and Environment*:104-108(in Chinses).

546   Deng, Y., Zhang, Y., Li, D., Shi, K., and Zhang, Y. 2017. Temporal and Spatial Dynamics of Phytoplankton
547   Primary Production in Lake Taihu Derived from MODIS Data. *Remote Sensing* 9:195. 10.3390/rs9030195

548   Ding, J., Yang, A., Wang, J., Sagan, V., and Yu, D. 2018. Machine-learning-based quantitative estimation of soil
549   organic carbon content by VIS/NIR spectroscopy. *PeerJ* 6. 10.7717/peerj.5714

550   Dörnhöfer, K., and Oppelt, N. 2016. Remote sensing for lake research and monitoring – Recent advances.
551   *ECOLOGICAL INDICATORS* 64:105-122. 10.1016/j.ecolind.2015.12.009

552   Esterby, S.R. 1996. Review of methods for the detection and estimation of trends with emphasis on water quality
553   applications. *HYDROLOGICAL PROCESSES* 10:127-149.

554   Feng, X., Li, G., Yu, H., Wang, S., Yi, X., and Lin, L. 2017. Wavelength selection for portable noninvasive blood
555   component measurement system based on spectral difference coefficient and dynamic spectrum. *Spectrochimica*
556   *Acta Part A Molecular & Biomolecular Spectroscopy* 193:40.

557   Gitelson, A.A., and KONDRATYEV, K.Y. 1991. Optical models of mesotrophic and eutrophic water bodies.
558   *INTERNATIONAL JOURNAL OF REMOTE SENSING* 12:373-385.

559   Han, L.H., and Rundquist, D.C. 1997. Comparison of NIR/RED ratio and first derivative of reflectance in estimating
560   algal-chlorophyll concentration: A case study in a turbid reservoir. *REMOTE SENSING OF ENVIRONMENT*
561   62:253-261. 10.1016/S0034-4257(97)00106-5

562   Hong, Y., Chen, Y., Lei, Y., Liu, Y., Liu, Y., Yong, Z., Yi, L., and Hang, C. 2018. Combining Fractional Order
563   Derivative and Spectral Variable Selection for Organic Matter Estimation of Homogeneous Soil Samples by VIS –
564   NIR Spectroscopy. *Remote Sensing* 10:479.

565   Hu, Y., and Wang, X. 2017. Application of surrogate parameters in characteristic UV – vis absorption bands for
566   rapid analysis of water contaminants. *Sensors and Actuators B: Chemical* 239:718-726. 10.1016/j.snb.2016.08.072

567   Hughes, G. 1968. On the mean accuracy of statistical pattern recognizers. *Information Theory IEEE Transactions on*
568   14:55-63.

569   Isiyaka, H.A., Mustapha, A., Juahir, H., and Phil-Eze, P. 2018. Water quality modelling using artificial neural
570   network and multivariate statistical techniques.

571   Ju-Long, D. 1982. Control problem of grey systems. *SYSTEMS & CONTROL LETTERS* 1:288-294.

572 Keller, S., Maier, P.M., Riese, F.M., Norra, S., Holbach, A., Boersig, N., Wilhelms, A., Moldaenke, C., Zaake, A.,
573 and Hinz, S. 2018. Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll a, Diatoms,
574 Green Algae and Turbidity. *International Journal of Environmental Research and Public Health* 15.
575 10.3390/ijerph15091881

576 Kennard, R.W., and Stone, L.A. 1969. Computer Aided Design of Experiments. *TECHNOMETRICS* 11:137-148.

577 Lerch, R. 2015. Long-term agroecosystem research in the Central Mississippi River Basin: dissolved nitrogen and
578 phosphorus transport in a high-runoff-potential watershed. *JOURNAL OF ENVIRONMENTAL QUALITY* 44:44.

579 Kotti, M., Zacharioudaki, D., Kokinou, E., and Stavroulakis, G. 2018. Characterization of water quality of Almiros
580 river (Northeastern Crete, Greece): physicochemical parameters, polycyclic aromatic hydrocarbons and anionic
581 detergents. *Modeling Earth Systems and Environment* 4:1285-1296. 10.1007/s40808-018-0504-3Li, C., Sun, L., Jia,
582 J., Cai, Y., and Wang, X. 2016. Risk assessment of water pollution sources based on an integrated k -means
583 clustering and set pair analysis method in the region of Shiyan, China. *SCIENCE OF THE TOTAL ENVIRONMENT*
584 557-558:307-316. 10.1016/j.scitotenv.2016.03.069

585 Morais, C.L.M., Santos, M.C.D., Lima, K.M.G., and Martin, F.L. 2019. Improving data splitting for classification
586 applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach.
587 *BIOINFORMATICS*. 10.1093/bioinformatics/btz421

588 Oussama, A., Elabadi, F., Platikanov, S., Kzaiber, F., and Tauler, R. 2012. Detection of Olive Oil Adulteration
589 Using FT-IR Spectroscopy and PLS with Variable Importance of Projection (VIP) Scores. *JOURNAL OF THE*
590 *AMERICAN OIL CHEMISTS SOCIETY* 89:1807-1812.

591 Palmer, S.C.J., Kutser, T., and Hunter, P.D. 2015. Remote sensing of inland waters: Challenges, progress and future
592 directions. *REMOTE SENSING OF ENVIRONMENT* 157:1-8.

593 Peerbhay, K.Y., Mutanga, O., and Ismail, R. 2013. Commercial tree species discrimination using airborne AISA
594 Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu‐Natal, South
595 Africa. *Isprs Journal of Photogrammetry & Remote Sensing* 79:19-28.

596 Peng J, Xu F, Deng K, and Wu J. 2018. Spectral difference of water quality different index concentrations: in
597 langya mountain scenic area . *Spectroscopy and Spectral Analysis* 38:1499-1507(in Chinses).

598 Pu, H., Liu, D., Qu, J., and Sun, D. 2017. Applications of Imaging Spectrometry in Inland Water Quality
599 Monitoring—a Review of Recent Developments. *Water, Air, & Soil Pollution* 228. 10.1007/s11270-017-3294-
600 8Rojas, F.S. 2009. Process Analytical Chemistry: Applications of Ultraviolet/Visible Spectrometry in
601 Environmental Analysis: An Overview. *APPLIED SPECTROSCOPY REVIEWS* 44:245-265.

602 Rostom, N.G., Shalaby, A.A., Issa, Y.M., and Afifi, A.A. 2017. Evaluation of Mariut Lake water quality using
603 Hyperspectral Remote Sensing and laboratory works. *The Egyptian Journal of Remote Sensing and Space Science*
604 20:S39-S48. 10.1016/j.ejrs.2016.11.002

605 Ryan, K., and Ali, K. 2016. Application of a partial least-squares regression model to retrieve chlorophyll-a
606 concentrations in coastal waters using hyper-spectral data. *Ocean Science Journal* 51:209-221.

607 Shafique, N.A., Fulk, F., Autrey, B.C., and Flotemersch, J. 2003. Hyperspectral remote sensing of water quality
608 parameters for large rivers in the Ohio River basin. *Renard*.

609 Song, K., Lin, L., Shuai, L., Tedesco, L., Duan, H., Li, Z., Shi, K., Jia, D., Ying, Z., and Shao, T. 2013. Using
610 Partial Least Squares-Artificial Neural Network for Inversion of Inland Water Chlorophyll-a. *IEEE Transactions on*
611 *Geoscience & Remote Sensing* 52:1502-1517.

612   Verpoorter, C., Kutser, T., Seekell, D.A., and Tranvik, L.J. 2014. A global inventory of lakes based on high-
613   resolution satellite imagery. *GEOPHYSICAL RESEARCH LETTERS* 41:6396-6402. 10.1002/2014GL060641

614   Wang, H., Chen, Y., Zhang, Z., Chen, H., Li, X., Wang, M., and Chai, H. 2019. Quantitatively estimating main soil
615   water-soluble salt ions content based on Visible-near infrared wavelength selected using GC, SR and VIP. *PeerJ*
616   7:e6310. 10.7717/peerj.6310

617   Wang H, Zhang Z, Arnon K, Chen J, Han W. 2018. Hyperspectral estimation of desert soil organic matter content
618   based on gray correlation-ridge regression model. *Transactions of the Chinese Society of Agricultural Engineering*
619   34:124–131 (in Chinese)

620   Wang, J., Tiyip, T., Ding, J., Dong, Z., Wei, L., and Fei, W. 2017. Quantitative Estimation of Organic Matter
621   Content in Arid Soil Using Vis-NIR Spectroscopy Preprocessed by Fractional Derivative. *Journal of Spectroscopy*
622   2017:1-9.

623   Wang, L., Pu, H., and Sun, D. 2016. Estimation of chlorophyll-a concentration of different seasons in outdoor ponds
624   using hyperspectral imaging. *TALANTA* 147:422-429. 10.1016/j.talanta.2015.09.018

625   Wang Y., and Zhao Z. 2000. Comparison of two methods and applications for determination of indoor visible - near
626   infrared reflectance spectra of water. *remote sensing information*:2-5.(in Chinses)

627   Wang, Y., Jing, H., Yu, L., Su, H., and Luo, N. 2017. Set pair analysis for risk assessment of water inrush in karst
628   tunnels. *Bulletin of Engineering Geology and the Environment* 76:1199-1207. 10.1007/s10064-016-0918-y

629   Wang, Z., Sakuno, Y., Koike, K., and Ohara, S. 2018. Evaluation of Chlorophyll-a Estimation Approaches Using
630   Iterative Stepwise Elimination Partial Least Squares (ISE-PLS) Regression and Several Traditional Algorithms from
631   Field Hyperspectral Measurements in the Seto Inland Sea, Japan. *SENSORS* 18.

632   Williams, P. 2001. Implementation of near-infrared technology. *Near-Infrared Technology in the Agricultural and*
633   *Food Industries*.

634   Wold, H. 1966. Estimation of Principal Components and Related Models by Iterative Least Squares. *Multivariate*
635   *Analysis* 1:391-420.

636   Wu Y., Du S., and Yan Z. 2011. Ultraviolet spectral analysis method for the concentration detection of organic
637   pollutants in water bodies. *Spectroscopy and Spectral Analysis*   31:233-237(in Chinses).

638   Xiao, Z., Yi, L., and Hao, F. 2016. Modeling Soil Cation Concentration and Sodium Adsorption Ratio Using
639   Observed Diffuse Reflectance Spectra. *CANADIAN JOURNAL OF SOIL SCIENCE* 96:2012-2016.

640   Yang, Y.L., Yan, G.Y., and Lin, Q. 2004. Determination of heavy metal ions in Chinese herbal medicine by
641   microwave digestion and RP-HPLC with UV-Vis detection. *MICROCHIMICA ACTA* 144:297-302.
642   10.1007/s00604-003-0020-6

643   Zhang, M., Maoguo, G., and Yongqiang, C. 2018. Hyperspectral band selection based on multi-objective
644   optimization with high information and low redundancy. *APPLIED SOFT COMPUTING*:S1668392098.

645   Zhang, R., Wang, Y., Wang, K., Zhao, H., Xu, S., Mu, L., and Zhou, G. 2019. An evaluating model for smart
646   growth plan based on BP neural network and set pair analysis. *JOURNAL OF CLEANER PRODUCTION* 226:928-
647   939. 10.1016/j.jclepro.2019.03.053

648   Zhang, Y., Giardino, C., and Li, L. 2017. Water Optics and Water Colour Remote Sensing. *Remote Sensing* 9:818.
649   10.3390/rs9080818

650   Zhu, T.T., Wu, L., Wang, X.L., Zhu, H., Zhu, X.C., Zhou, Q.G., Liu, X., and Cai, B.C. 2017. Investigation on
651   relationships between chemical spectrum and bioeffect of prepared rhubarb decoction in rats by UPLC-ESI-Q-TOF-
652   MS method coupled with gray correlation analysis. *Journal of Functional Foods* 31:104-112.

653   Zhao, K., Xuan, A., 1996. Set pair theory-a new theory method of non-define and its applications. Syst. Eng. 14, 18–

654   23 (in Chinese).

655   Zhou P, Yang W, Li M., Zheng L., and Chen Y.. 2017. Prediction of soil total nitrogen based on gray correlation

656   and extreme learning machine. *Transactions of the Chinese Society for Agricultural Machinery* 48:271-276(in

657   Chinses).

658

Figure 1

spectral data curve

(A) Reflectance spectral curves (B) Standard normal variable reflectance curves.

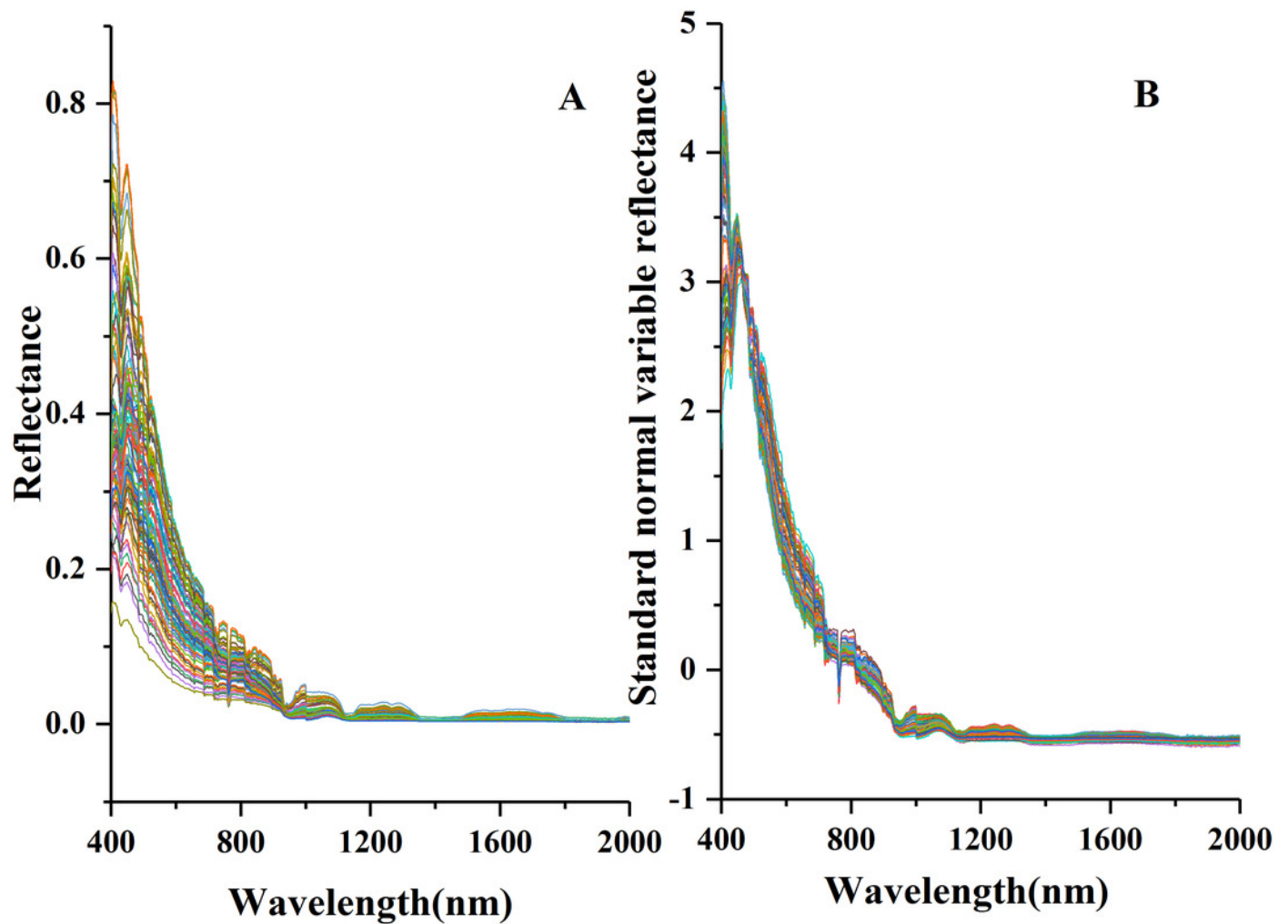



Figure 2

Correlation coefficients of water quality parameters with standard normal variable reflectance

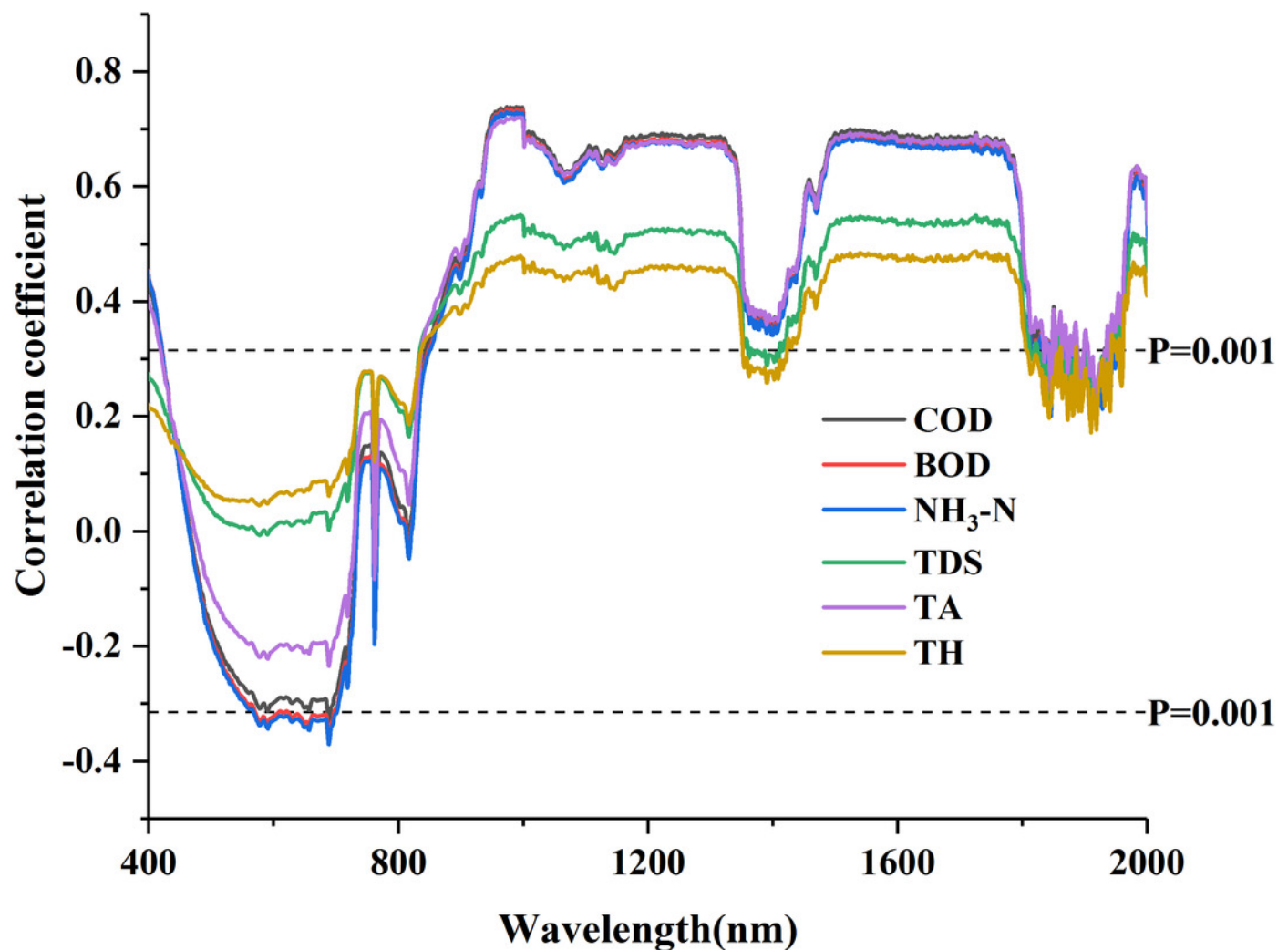



Figure 3

Gray correlation degree (GCD) for water quality parameters with standard normal variable reflectance

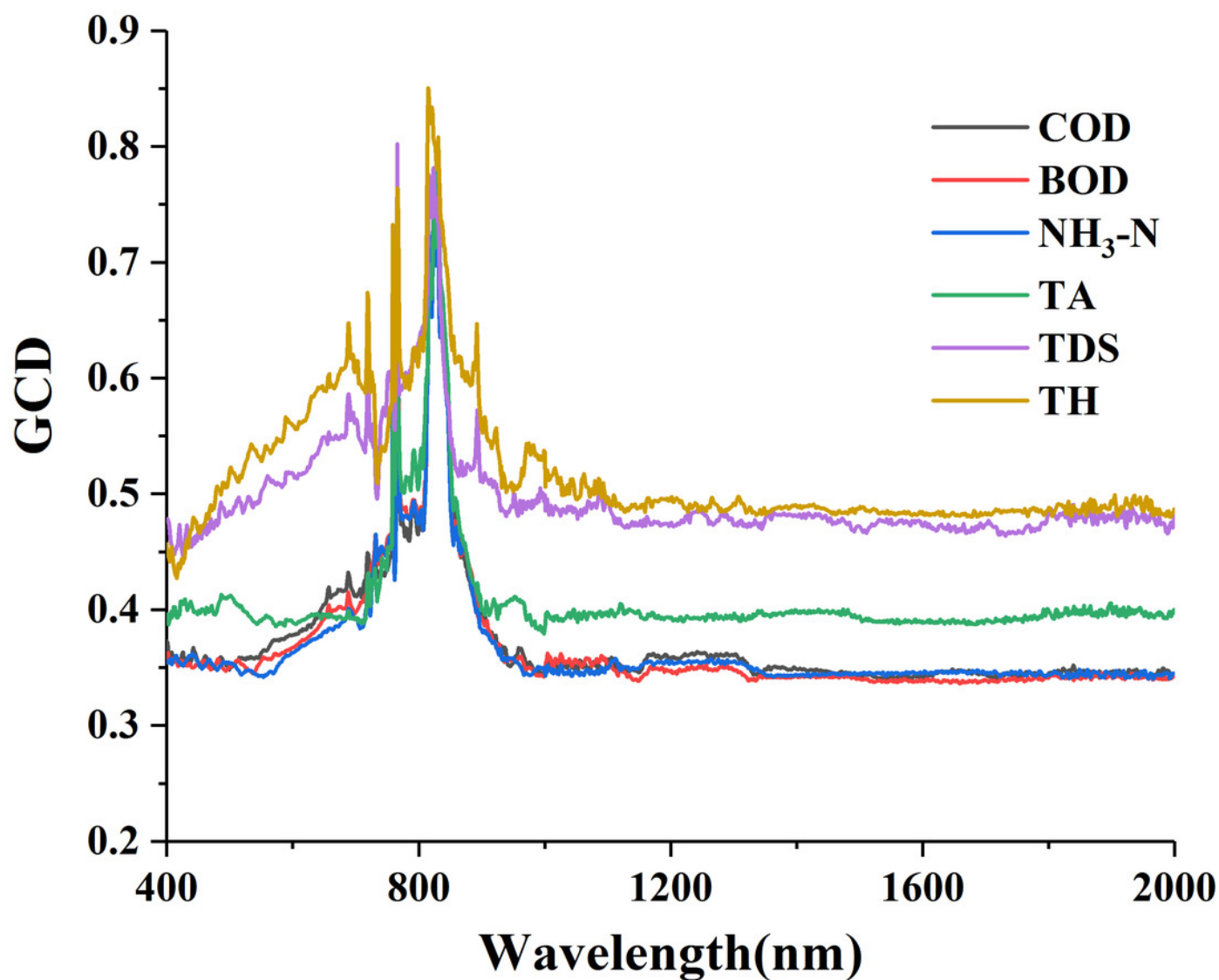



Figure 4

The VIP scores of water quality parameters with standard normal variable reflectance

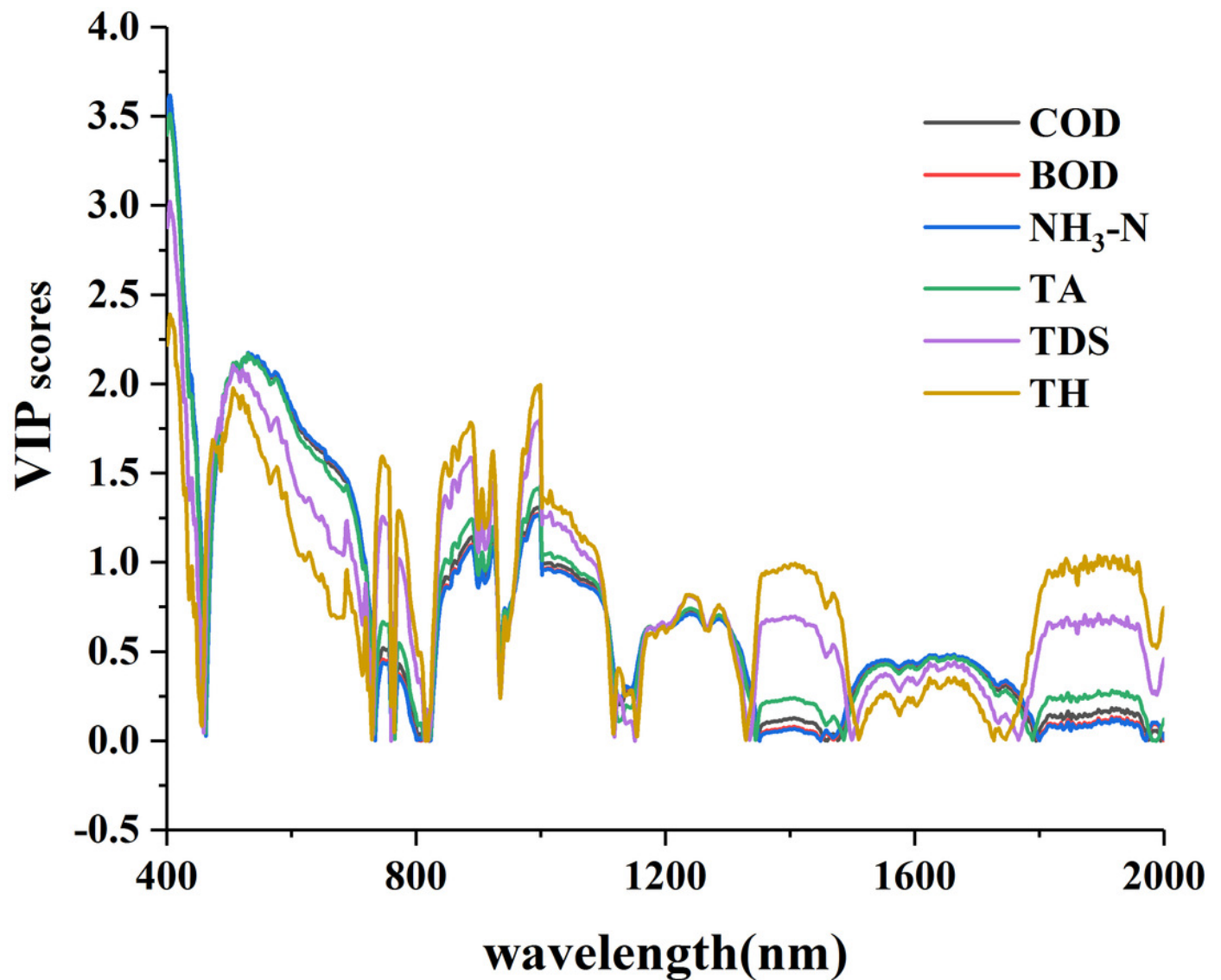



Figure 5

The SPA scores of water quality parameters with standard normal variable reflectance

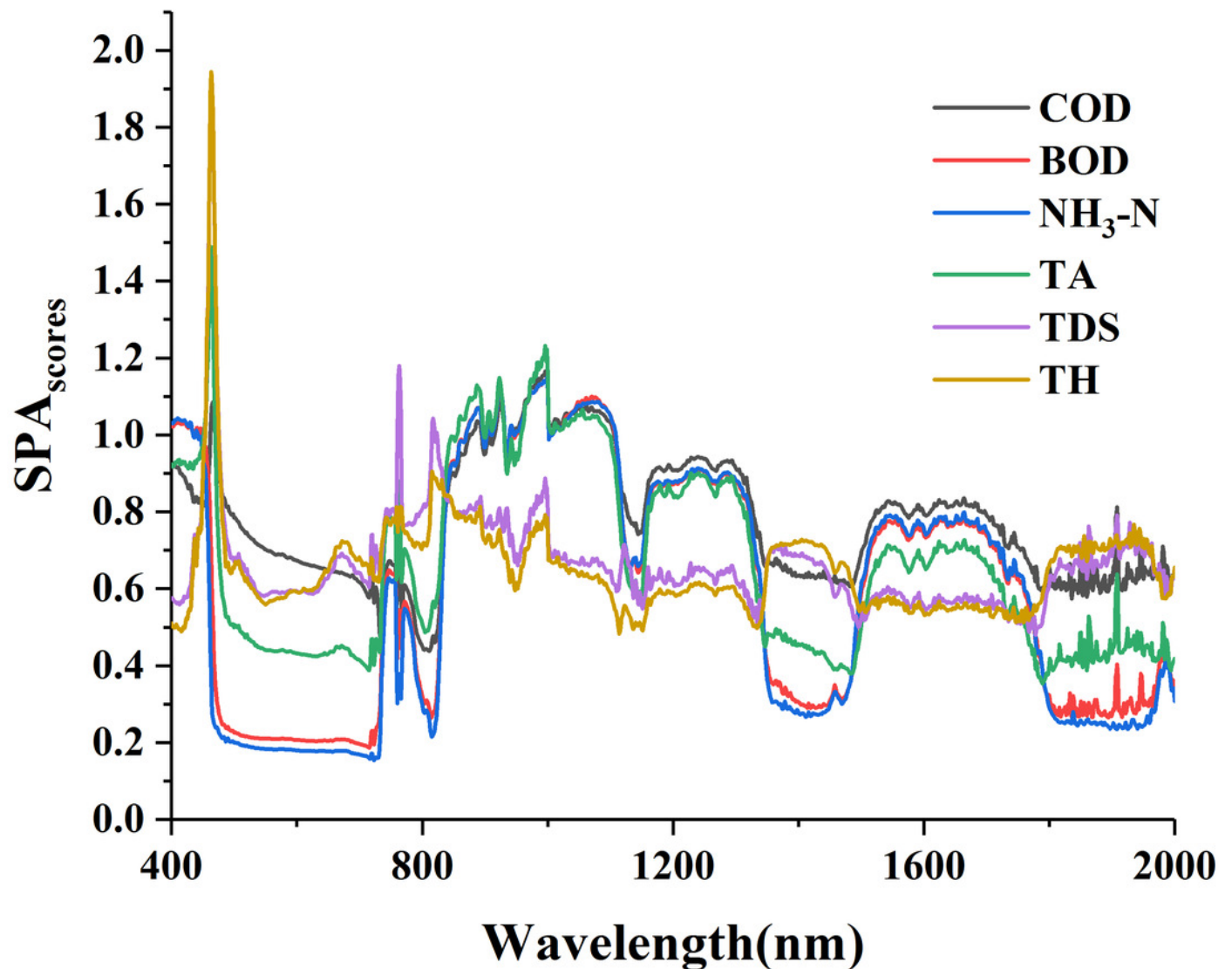
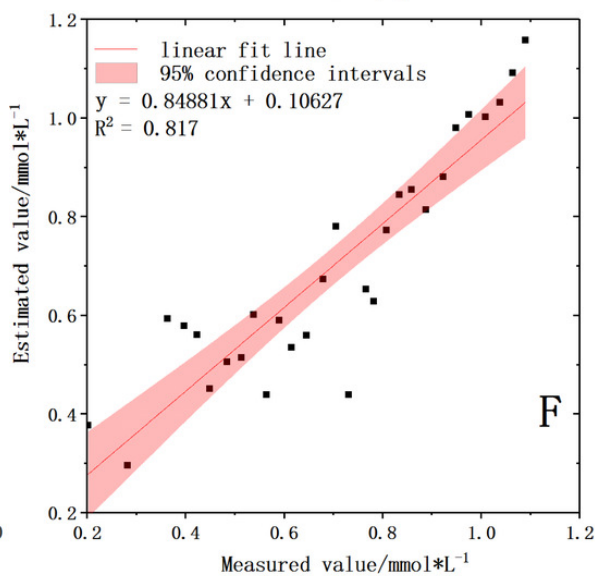
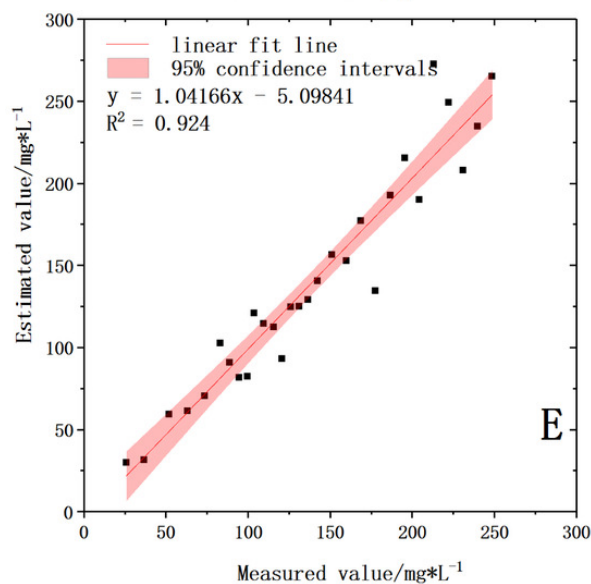
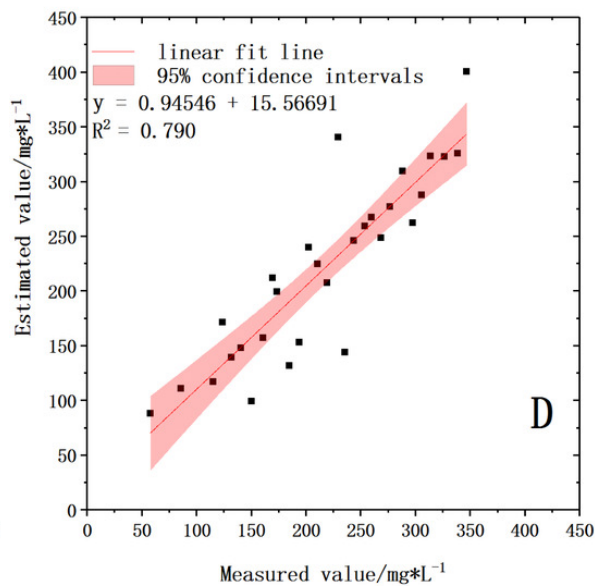
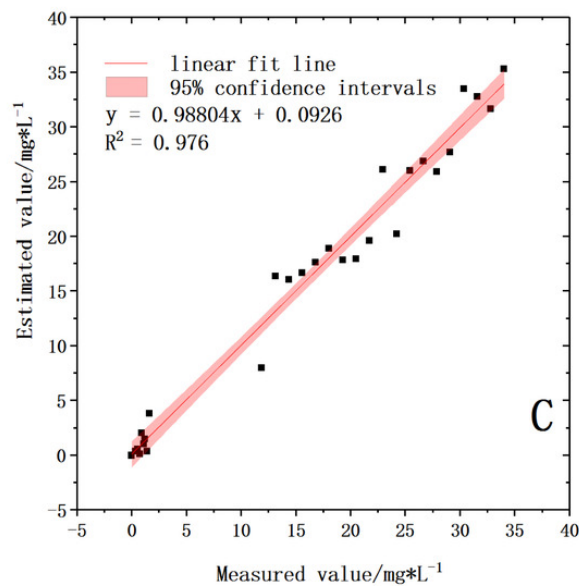
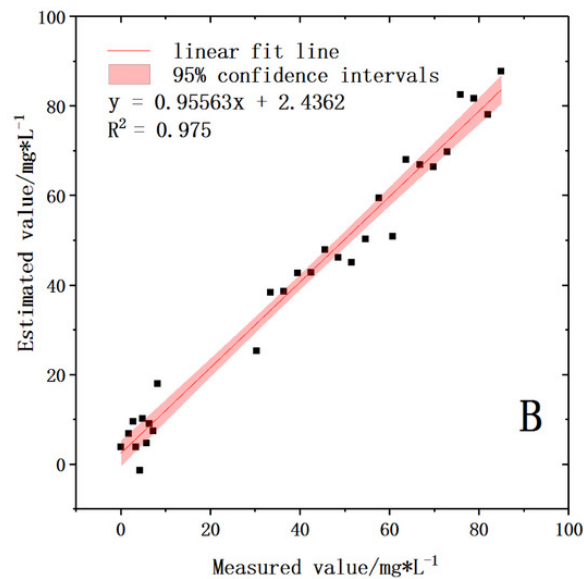
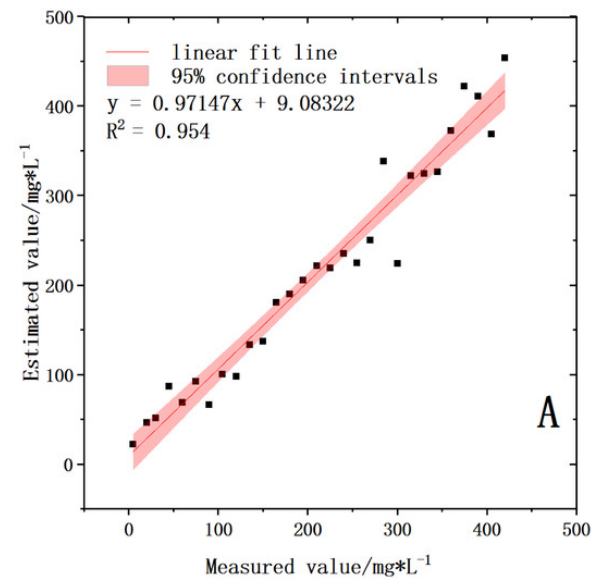



Figure 6

validation of water quality parameters based on the best model

(A) COD with GC-PLSR model (B) BOD with GC-ELM model (C) NH_3-N with GC-ELM model (D)

TDS with SPA-ELM model (E) TA with SPA-ELM model (F) TH with SPA-ELM model

Main water quality parameters

Parameters	Water inlet	Anoxic tank	Aerobic tank	Sedimentation tank	Outlet of water	Experiment methods
NH$_3$-N(mg/L)	34.853	1.723	1.499	0	0	According to Nessler's reagent spectrophotometer with the amount of visible light spectrophotometer 722 N for determining NH$_3$-N
Total alkalinity(mg/L)	251.70	147.02	148.20	101.15	103.50	According to acid base indicator titration method
Total hardness(mmol/L)	1.09	1.13	1.13	1.17	1.11	According to the EDTA titration method (GB11914-1989)
Total dissolved substance(mg/L)	351	323	317	344	343	According to Gravimetric method (GB T5750.5-2006)
COD(mg/L)	425	140	134	23	20	According to the dichromate method (GB11914-1989) using a standard COD digestion apparatus(K-100) to determine COD
BOD(mg/L)	86	8.5	13	3.1	6.2	According to the dilution and inoculation method (HJ505-2009) with a constant temperature incubator (HWS-150 type) for determining the content of BOD5

Maximum gray correlation degree and band intervals of water quality parameters content with standard normal variable reflectance

Water quality parameters	Sensitive band numbers	Maximum GCD	Maximum GCD band intervals(nm)
COD	49	0.7949	820~830
BOD	50	0.7974	820~830
NH_3-N	46	0.7973	820~830
TA	93	0.7878	820~830
TDS	381	0.802	815~825/766
TH	601	0.8504	815~825

The maximum VIP scores of water quality parameters with standard normal variable reflectance

Water quality parameters	Sensitive band numbers	Maximum VIP scores	Maximum VIP band interval(nm)l	Maximum VIP band(nm)
COD	753	1.634	465~475	466
BOD	770	1.439	464~474	762
NH$_3$-N	768	1.397	990~999	999
TA	709	2.466	460~470	464
TDS	497	4.275	460~470	463
TH	543	4.893	460~470	463

The Set pair analysis (SPA) scores of water quality parameter with standard normal variable reflectance

Water quality parameters	Sensitive band numbers	Maximum SPA	Maximum SPA band interval(nm)	Maximum SPA band(nm)
COD	750	1.598	465~475	466
BOD	767	1.409	762~763	762
NH$_3$-N	765	1.364	990~999	999
TA	696	2.409	460~470	464
TDS	280	1.928	460~470	463
TH	223	1.944	460~470	463

PLSR model results based on full band, GC, SPA and VIP screening band

Water quality parameters	Wavelength selection methods	Main factor numbers	Modeling set	Validation set	Relative percent deviation	Robustness
			R²c	R²p	RPD	Robust
COD	GC	6	0.970	0.954	4.690	0.984
	VIP	4	0.938	0.917	3.238	0.978
	SPA	4	0.938	0.917	3.237	0.978
	All	9	0.992	0.956	4.532	0.964
BOD	GC	5	0.976	0.965	5.335	0.988
	VIP	10	0.992	0.951	4.424	0.958
	SPA	4	0.958	0.930	3.348	0.971
	All	7	0.990	0.960	4.998	0.970
NH₃-N	GC	5	0.972	0.970	5.894	0.998
	VIP	7	0.970	0.935	3.977	0.965
	SPA	7	0.969	0.935	3.974	0.965
	All	8	0.985	0.962	5.228	0.977
TDS	GC	8	0.876	0.772	2.049	0.880
	VIP	3	0.767	0.758	1.892	0.988
	SPA	5	0.820	0.793	2.150	0.964
	All	5	0.862	0.791	2.126	0.918
TA	GC	5	0.917	0.890	2.963	0.971
	VIP	4	0.883	0.927	3.578	1.049
	SPA	4	0.883	0.927	3.588	1.050
	All	7	0.956	0.921	3.206	0.963
TH	GC	12	0.978	0.900	2.871	0.920
	VIP	4	0.779	0.733	1.764	0.941
	SPA	3	0.794	0.708	1.701	0.892
	All	5	0.820	0.817	2.228	0.996

ELM model results based on full band, GC, SPA and VIP screening band

Water quality parameters	Wavelength selection methods	Main factor numbers	Modeling set	Validation set	Relative percent deviation	Robustness
			R^2_c	R^2_p	RPD	Robust
COD	GC	6	0.970	0.954	4.690	0.984
	VIP	4	0.938	0.917	3.238	0.978
	SPA	4	0.938	0.917	3.237	0.978
	All	9	0.992	0.956	4.532	0.964
BOD	GC	5	0.976	0.965	5.335	0.988
	VIP	10	0.992	0.951	4.424	0.958
	SPA	4	0.958	0.930	3.348	0.971
	All	7	0.990	0.960	4.998	0.970
NH_3-N	GC	5	0.972	0.970	5.894	0.998
	VIP	7	0.970	0.935	3.977	0.965
	SPA	7	0.969	0.935	3.974	0.965
	All	8	0.985	0.962	5.228	0.977
TDS	GC	8	0.876	0.772	2.049	0.880
	VIP	3	0.767	0.758	1.892	0.988
	SPA	5	0.820	0.793	2.150	0.964
	All	5	0.862	0.791	2.126	0.918
TA	GC	5	0.917	0.890	2.963	0.971
	VIP	4	0.883	0.927	3.578	1.049
	SPA	4	0.883	0.927	3.588	1.050
	All	7	0.956	0.921	3.206	0.963
TH	GC	12	0.978	0.900	2.871	0.920
	VIP	4	0.779	0.733	1.764	0.941
	SPA	3	0.794	0.708	1.701	0.892
	All	5	0.820	0.817	2.228	0.996