Modeling the ecological status response of rivers to multiple stressors using machine learning: a comparison of environmental DNA metabarcoding and morphological data

Juntao Fan [a], Shuping Wang [a], Hong Li [b, c], Zhenguang Yan [a, *], Yizhang Zhang [a, d], Xin Zheng [a], Pengyuan Wang [a]

[a] *State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing, 100012, China*

[b] *Lancaster Environment Centre, Lancaster University, LA1 4YQ, UK*

[c] *UK Centre for Ecology & Hydrology, MacLean Building, Wallingford OX108 BB, UK*

[d] *Chinese Research Academy of Environmental Sciences Tianjin Branch, Tianjin, 300457, China*

* Corresponding author.

E-mail address: zgyan@craes.org.cn (Z. Yan).

ABSTRACT

Understanding the ecological status response of rivers to multiple stressors is a precondition for river restoration and management. However, this requires the collection of appropriate data, including environmental variables and the status of aquatic organisms, and analysis via a suitable model that captures the nonlinear relationships between ecological status and various stressors. The morphological approach has been the standard data collection method employed for establishing the status of aquatic organisms. However, this approach is very laborious and restricted to a specific set of organisms. Recently, an environmental DNA (eDNA) metabarcoding data approach has been developed that is far more efficient than the morphological approach and potentially applicable to an unlimited set of organisms. However, it remains unclear how well eDNA metabarcoding data reflects the impacts of environmental stressors on aquatic ecosystems compared with morphological data, which is essential for clarifying the potential applications of eDNA metabarcoding data in the ecological monitoring and management of rivers. The present work addresses this issue by modeling organism diversity based on three indices with respect to multiple environmental variables in both the catchment and reach scales. This is done by corresponding support vector machine (SVM) models constructed from eDNA metabarcoding and morphological data on 24 sampling locations in the Taizi River basin, China. According to the mean absolute percent error (MAPE) between the measured diversity index values and the index values predicted by the SVM models, the SVM models constructed from eDNA metabarcoding data (MAPE = 3.87) provide more accurate predictions than the SVM models constructed from morphological data (MAPE = 28.36), revealing that the eDNA metabarcoding data better reflects environmental conditions. In addition, the sensitivity of SVM model predictions of the ecological indices for both catchment-scale and reach-scale stressors is evaluated, and the stressors having the greatest impact on the ecological status of rivers are identified. The results demonstrate that the ecological status of rivers is more sensitive to environmental stressors at the reach scale than to stressors at the catchment scale. Therefore, our study is helpful in exploring the potential applications of eDNA metabarcoding data and SVM modeling in the ecological monitoring and management of rivers.

Keywords

Machine learning; Modeling; Environmental DNA; Biomonitoring; Freshwater ecosystem

1. Introduction

River ecosystems are impacted by multiple environmental variables at both the catchment scale and reach scale simultaneously, and any of these variables lying outside of their normal range can become a stressor. These natural and anthropogenic stressors always interact and are directly or indirectly impacting ecological status (Mori et al., 2019; Romero et al., 2018). For example, catchment scale stressors, such as increased impervious land use by humans, altere physical and chemical conditions of rivers such as increased nutrition through hydrological processes, affecting the structure and function of aquatic ecosystems (Bernhardt et al., 2012; Von Schiller et al., 2017). Here, aquatic communities play an important role in supporting ecosystem services, stability, and biodiversity, and their status can reflect the long-term cumulative effects of environmental stressors on aquatic ecosystems (Franzo and Del Negro, 2019). Therefore, biomonitoring is essential for assessing the impacts of human disturbance at the multiple scales of river basins. The standard approach that has been applied to river biomonitoring involves the sorting and morphological identification of aquatic communities, which is time-consuming and demands a high degree of taxonomic expertise (Pawlowski et al., 2018). However, the high-throughput amplicon sequencing of environmental DNA (eDNA) has recently provided a viable option for biomonitoring, which purified from substrates such as soil or water contains DNA fragments originating from organisms present in that environment (Cordier et al., 2017; Jarman et al., 2018; Mize et al., 2019; Visco et al., 2015). Moreover, a number of previous studies have shown that eDNA metabarcoding data can provide an accurate indication of environmental changes. For example, the relative abundance of operational taxonomic units (OTUs) indicative of plankton was demonstrated to have a significant negative correlation with river nutrient levels (Li et al., 2018a). The foraminifera diversity inferred from eDNA metabarcoding data was found to have a significant positive correlation with the biodiversity in the benthic zone impacted by fish farming activities (He et al., 2019), and the distance from a wellhead in the ocean (Laroche et al., 2016). Benthic macroinvertebrates diversity inferred from eDNA metabarcoding data were also used to assess the freshwater quality (Fernandez et al., 2018; Hering et al., 2018). In addition, previous studies have shown that, compared with morphological classification, eDNA metabarcoding is a relatively simple and affordable method for assessing biodiversity on a large temporal and spatial scale without the need for time-consuming microscopy analysis by experts (He et al., 2019; Ji et al., 2013). Taxonomic classification based on eDNA metabarcoding is usually more accurate than morphological identification, particularly for species with similar morphology and species with poor life cycle characteristics (He et al., 2019; Humbert et al., 2010). Furthermore, eDNA metabarcoding data can be easily reanalyzed to make it suitable for review by third parties (Ji et al., 2013). However, it remains unclear how well eDNA metabarcoding data reflects the impacts of environmental stressors on aquatic ecosystems in comparison with morphological identification data. Clarifying this issue will illuminate potential applications of eDNA technology in the monitoring and management of aquatic ecosystems.

Understanding the response of river ecosystems to multiple stressors and identifying important stressors are prerequisites for conducting effective river restoration and management (Meissner et al., 2019; Zhang, 2019). Developing this understanding requires the analysis of biomonitoring data via a suitable model that captures the relationships between the status of ecosystems and various stressors. However, the interactions of multiple stressors produce a combined effect that can be equal to (additive), greater than (synergistic), or less than (antagonistic) the sum of each single effect (Piggott et al., 2015). Indeed, the response of aquatic ecosystems to multiple stressors is typically nonlinear, which greatly complicates the development of accurate models (Jones et al., 2017). The modeling of nonlinear responses can be conducted using various methods, including mathematical/physical models, statistical models, and data-driven models (Al-Mukhtar, 2019; Choubin et al., 2018; Park et al., 2015). However, the complexity of relationships between ecological status and multiple stressors limits the application of mathematical/physical models, and statistical models also suffer from disadvantages, such as poor generalizability due to relatively small sample sizes (Cui and Gong, 2018; Varoquaux, 2018). The development of machine learning (ML) over the past few years has provided a new approach for quantifying these nonlinear relationships (Torija and Ruiz, 2015). At present, ML models have been widely used in the prediction of environmental or ecological indicators. For example, a Bayesian belief network (BBN) was applied to model the combined effects of land use change and climate change on the status of macroinvertebrates and fish in freshwater bodies (Olson, 2018). In addition, artificial neural networks (ANNs), the support vector machine (SVM) and generalized regression neural network, were used for predicting chlorophyll-a concentrations in freshwater, and the results demonstrated that these data-driven ML methods achieved better prediction performance than conventional statistical methods (Marvuglia et al., 2015; Park et al., 2015). The SVM method is particularly advantageous for modeling nonlinear response relationships because the SVM is good for solving high-dimensional and nonlinear problems, while avoiding the difficulties associated with determining the network structure and local minima of the solutions, and provides good generalizability and relatively good prediction performance under small sample size conditions (Vapnik, 1999). These advantages have made SVM outperform other ML methods, e.g., standard ANNs, random forest (RF) classifiers, and boosted trees (BT) classification, in the prediction of soil organic carbon, clay content, and pH (Rossel and Behrens, 2010; Were et al., 2015) and chlorophyll-a (Park et al., 2015) in some regions. Therefore, the SVM is well suited for modeling the relationships between the ecological status of rivers and multiple stressors.

The present study compares the ability of eDNA metabarcoding data and morphological identification data to reflect the nonlinear impact of multiple environmental stressors on aquatic ecosystems by employing both sets of data in SVM models corresponding to three ecological indices (i.e. observed species, Shannon Wiener index, and Simpson index), which were commonly used in biodiversity assessment inferred from eDNA metabarcoding or morphological data. As such, the present work helps to explore the potential applications of eDNA technology in the monitoring and management of aquatic ecosystems. In addition, the sensitivity of SVM model predictions of the ecological indices to individual catchment-scale and reach-scale stressors is evaluated, and the stressors having the greatest impact on the ecological status of rivers are identified.

2. Materials and methods

2.1. Study area

The study area was the upstream area of the Taizi River basin (122°23’E–122°53’E, 40°28’N–41°39’N) in northeastern China. The location and characteristics of the study area are illustrated in Fig. 1. A previous study demonstrated that the ecological status of the Taizi River in this area was relatively good because the majority of the land in the upstream area was covered by forests, and the intensity of human activities was relatively low (Fan et al., 2015). The primary aquatic organisms of the Taizi River, particularly those species most sensitive to environmental stressors, such as clean-type fishes (*Lampetra morii* and *Odontobutis Obscurus*) and macrobenthos (*Epeorus melli* and *Cambaroides dauricus*), are mainly distributed in the upper reaches of the river. All of these organisms play an important role in maintaining the health of the aquatic ecosystem. However, the urbanization process in the region and the acceleration of human activities in recent years, such as agriculture and mining, have resulted in water shortages, the deterioration of water quality, habitat damage, loss of biodiversity, and the reduction of ecological functions.

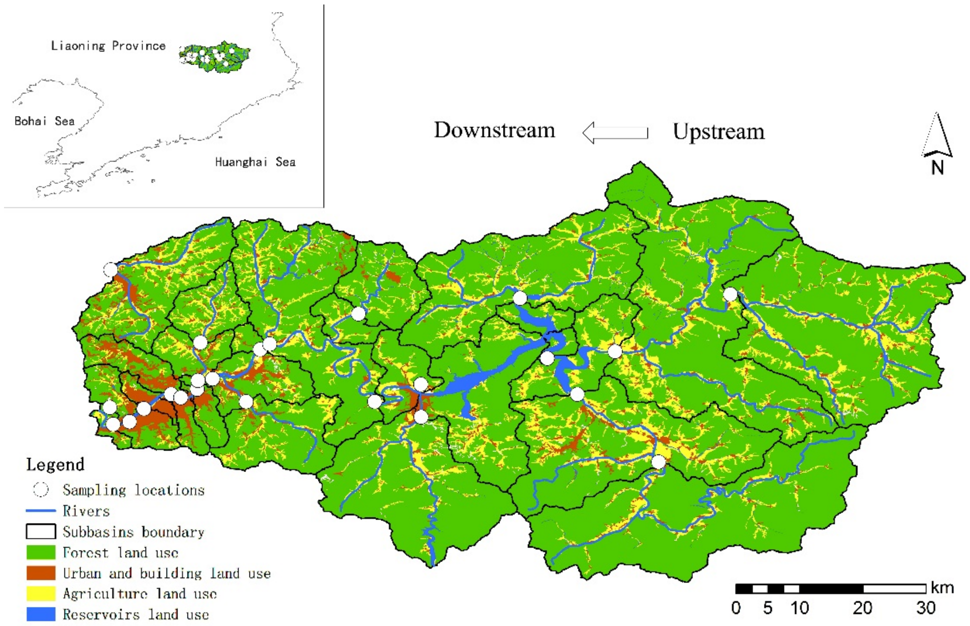

Fig. 1. Map of the study area incorporating the upper area of the Taizi River basin at the time of sample collection in October, 2018. The 24 sampling sites and different types of land use in the sub-basins are indicated, and the location of the study area relative to the national boundary of China is shown in the inset.

2.2. Ecological and environmental data collection

The 24 sites sampled during October 2018 (Fig. 1) were located in the mainstem and tributaries of the upstream area of the Taizi River basin. Surface water was sampled using sterile bottles. One liter per site was used for eDNA metabarcoding analysis.

Three independent extractions of 300 mL were obtained from each one-liter water sample within 6 h after sampling by filtering across a Millipore 0.22 μm hydrophilic nylon membrane. The membrane discs containing captured eDNA were placed in 5.0 mL centrifugal tubes, and were instantly frozen and stored at −20°C until DNA extraction. For morphological identification, phytoplankton samples were collected at each sampling site by dragging a nylon mesh with a pore size of 64 μm under the water surface for about 2 min. The water sample concentrated in the drip tube of the net was collected in a 50 mL sample bottle and fixed using Lugol's solution.

Environmental variables considered include catchment-scale variables (i.e., land use data) and reach-scale variables (i.e., physicochemical parameters). Land use data were extracted from an analysis of Spot Image data obtained with a 2.5 m resolution. The proportion of land use types (i.e., forest, agriculture, urban, and industrial) was determined for the region of the catchment upstream of each sampling site contributing to the sample characteristics and for a 250 m impact zone adjacent to the studied river segment. Ten physicochemical indicators were selected, including electrical conductivity (EC), dissolved oxygen (DO), pH, biological oxygen demand over 5 days (BOD_5), permanganate index (COD_{Mn}), total phosphorus (TP), ammonia nitrogen concentration (NH_3-N), total nitrogen (TN), suspended sediment (SS), and volatile phenol (VP). The work of (Fan et al., 2015) and Chinese Quality Standards for Surface Water Resources (Ministry of Water Resources, 1994) established thresholds not to be exceed to assure high ecological status for these physicochemical parameters. These are given as follows: EC = 400 μs/cm, DO = 7.5 mg/L, BOD_5 = 3 mg/L, COD_{Mn} = 2 mg/L, NH_3-N = 0.15 mg/L, TN = 0.2 mg/L (which was only considered in lake or reservoir samples), TP = 0.02 mg/L, VP = 0.002 mg/L, SS = 20 mg/L, pH = 6.5~8.5.

2.3. eDNA metabarcoding and morphological identification

Phytoplankton is the target taxonomic group of eDNA metabarcoding and morphological identification. Total eDNA was extracted using the cetyl trimethylammonium bromide (CTAB) method combined with the Zymo DNA Clean & Concentrator kit (Zymo Research Corp, Irvine, USA) (Yuan et al., 2015). The concentration of eDNA was determined using a NanoDrop One microvolume ultraviolet-visible (UV-vis) spectrophotometer (Thermo Fisher Scientific, Carlsbad, USA). The eDNA was used as templates for the polymerase chain reaction (PCR) method with 18S rRNA gene primers 18SV9F (5'-CCCTGCCNTTTGTACACAC-3') and 18SV9FR (5'-CCTTCNGCAGGTTCACCTAC-3') (Amaral-Zettler et al., 2009; De Vargas et al., 2015). The 18S rRNA gene primers were used because phytoplankton diversity including the cryptic diversity in environmental samples can be indicated by sequencing of 18S rRNA gene, and the SILVA datasets offered the 18S primer opportunity to assess distribution patterns of phytoplankton species (Treusch et al., 2012). The purified PCR products were added with 8-base sequence tags corresponding to each sample. High throughput sequencing was conducted using a MiSeq sequencing platform (Illumina, San Diego, USA). All low-quality sequencing data points with adaptors, ambiguous bases, low complexity, and those having average quality scores less than 20 were discarded using the UPARSE pipeline (Edgar, 2013). The OTUs were determined at the ≥97% identity level (Edgar, 2013). Taxonomic annotation analysis was performed using the Qiime2 pipeline (Caporaso et al., 2010) with respect to the SILVA-119 reference database. The remaining high-quality data were transformed to relative proportions before conducting subsequent statistical analysis.

For morphological identification, samples were concentrated and precipitated, and the sample volume was adjusted to 20–50 mL. The concentrated sample was then shaken uniformly, and 0.1 mL of the sample was immediately placed in a counting box for morphological identification. The phytoplankton taxa in each sample were identified under a 10 × 40 microscope. However, if a high concentration of diatoms were observed, the sample was sealed and identified under a 10 × 100 microscope. The specimens were identified to species level through microscopy and taxonomic experts consultation. The reference used to identify phytoplankton is the Freshwater Alage of China – Systematics, Taxonomy and Ecology (Hu and Wei, 2006).

2.4. SVM model development

The ecological status of the samples was evaluated according to the obtained eDNA metabarcoding and morphology identification data based on three widely used ecological indices, i.e., observed species (Kefford et al., 2011), Shannon Wiener index (Strong, 2016), and Simpson index (Keylock, 2005). The abbreviations and ecological significance of each of these indices are listed in Table 1. The values for these ecological indices obtained from the eDNA metabarcoding and morphology identification data were employed as the response/dependent variables in their respective SVM models. The catchment-scale variables and reach-scale variables were input to the respective SVM models as the independent variables.

Table 1

List of ecological indices with abbreviations and ecological significance.

Ecological index/Response variables	Abbreviations in eDNA metabarcoding	Abbreviations in morphological identification	Ecological significance
Observed species	Species _E	Species _M	Number of species or OTU observed.
Shannon Wiener index	Shannon _E	Shannon _M	The species/OTUs richness and evenness of the community, but predominantly sensitive to richness.

			Richness increases with increasing index value.
Simpson index	Simpson _E	Simpson_M	The species/OTUs richness and evenness of the community, but predominantly sensitive to evenness. Evenness increases with increasing index value.

The SVM was applied for nonlinear regression analysis to establish the response of the ecological indices to the multiple environmental variables. Here, the input data were mapped initially into a higher-dimensional feature space via a kernel function (i.e., a linear kernel, polynomial kernel, radial basis kernel, and Gaussian kernel), and then linear regression was performed in the high-dimensional feature space to obtain the nonlinear regression effect in the original space (Balfer and Bajorath, 2015; Bouboulis et al., 2015). The specific kernel function applied was selected by cross-validation (Piette and Moore, 2018).

The regression performance of the SVM depends on the appropriate selection of parameter values, including cost (c), epsilon (ε), and gamma (γ), where both c and ε are employed to establish the penalty coefficient, which represents the error tolerance of the regression analysis, and γ determines the distribution of the data after it is mapped to the new feature space. Here, the number of support vectors decreases with increasing γ, which affects the speed of training and prediction. The values of these parameters are optimized using a loop traversal algorithm (Cherkassky and Ma, 2004). Normalization was applied to all independent variables to ensure that the indicator values were comparable.

The generalization ability of the model was verified by 8-fold cross validation, where the dataset was divided into 8 subsets, and each subset was employed as the testing set once, while the remaining 7 subsets were used as the training set. Accordingly, this process was repeated 8 times. The prediction error of each model was evaluated based on the mean absolute percent error (MAPE), which is calculated for *n* samples as follows:

$$\mathrm{MAPE} = \sum_{t=1}^{n} \left| \frac{Observed_t - Predicted_t}{Observed_t} \right| \times \frac{100}{n} \qquad (1)$$

where *Observed*$_t$ is the observed value and *Predicted*$_t$ is the predicted value. Then, the model with the smallest MAPE value was selected as the optimal model.

Sensitivity analysis was applied to determine the environmental variables that most greatly influenced the model predictions of the ecological indices. This was conducted using the one-factor-at-a-time (OAT) approach. Here, the MAPE values of the model predictions were obtained with one environmental variable omitted at a time, while the other environmental variables were held constant. Then, the impact of each environmental variable on the model prediction was evaluated according to the absolute value of the difference between the MAPE obtained with and without that variable, which is denoted herein as ΔMAPE. Accordingly, the sensitivity of the ecological index predictions to an environmental variable increases with increasing ΔMAPE.

3. Results

3.1. Environmental conditions in catchment and reach scales

All the environmental variables have become stressors, which are marked with "+" in Table 2. Spatial analysis showed that almost all sites were under the selected catchment-scale stressors, and the downstream sites (e.g., s19, s15 and s22) were under more reach-scale stressors than the upstream sites (Table 2). We note that the proportion of forest land use in the catchment scale (0.268–0.910) is greater than that in the 250 m buffer zone (0.092–0.566). However, the proportion of agriculture land use in the 250 m buffer zone (maximum value of 0.596) is greater than that in the catchment scale (maximum value of 0.265), which indicates that agricultural disturbance is greater in the riparian zone than at the catchment scale, while urban and industrial disturbances have opposite behaviors. Table 2 also indicates that, TN and VP were the reach-scale variables with the highest number of sites exceeding the thresholds.

Table 2

List of environmental variables included in the modeling and spatial distribution of sites with corresponding stressors. Stressors, i.e., catchment-scale variables impacted by any artificial land use types (i.e., agriculture, urban and industrial land use), and reach-scale variables with values less than or greater than the threshold values representing high environmental status established by the work of (Fan et al., 2015) and Chinese quality standards for surface water resources (Ministry of Water Resources, 1994), are marked with "+".

Environmental variables		Abbreviations （Units）	Ranges	Sites with corresponding stressors
Catchment-scale variables				
Forest land use (catchment scale)	+	F_cat （proportion）	0.268–0.910	All sites
Forest land use (250 m buffer zone)	+	F_buf (proportion)	0.092–0.566	All sites
Agriculture land use (catchment scale)	+	A_cat (proportion)	0.018–0.265	All sites
Agriculture land use (250 m buffer zone)	+	A_buf (proportion)	0.000–0.596	All sites except s19, s21, s16, s20
Urban and industrial land use (catchment scale)	+	U_cat (proportion)	0.016–0.646	All sites
Urban and industrial land use (250 m buffer zone)	+	U_buf (proportion)	0.011–0.520	All sites
Reach-scale variables				
Electrical conductivity	+	EC （μs/cm）	142.47–655.33	s19, s13, s14, s21, s15, s17,

				s20, s22
Dissolved oxygen	+	DO (mg/L)	7.02–14.26	s22
pH	+	pH	7.84–8.98	s19, s12, s10, s02, s20
Permanganate index	+	COD_{Mn} (mg/L)	0.48–5.72	All sites except s10, s03, s11
Five-day biochemical oxygen demand	+	BOD_5 (mg/L)	0.75–8.41	s04, s14, s15, s16, s18, s24
Ammonia nitrogen	+	NH_3-N (mg/L)	0.12–3.87	All sites except s05, s03, s07, s06, s01, s02
Total nitrogen	+	TN (mg/L)	1.55–6.75	All sites
Total phosphorus	+	TP (mg/L)	0.004–0.223	s19, s04, s14, s06, s21, s23, s15, s09, s16, s24, s20, s22
Suspended sediment	+	SS (mg/L)	1.56–35.33	s15, s01, s08
Volatile phenol	+	VP (mg/L)	0.004–0.112	All sites

3.2. Ecological status derived from eDNA metabarcoding and morphological data

A total of 67 18S rRNA gene libraries were analyzed according to the methodology presented in Subsection 2.3, which resulted in a total of 2,305,498 high-quality sequences, and a total of 6,635 OTUs. The number of OTUs in each sample was distributed between 477–2,661 (Table S1). The result of taxonomic group distribution of OTUs showed that approximately 83% eukaryotic sequences were annotated as phytoplankton (Fig. 2), which confirmed that the phytoplankton can be indicated by sequencing of 18S rRNA gene. Therefore, the eDNA metabarcoding data and morphological data are comparable in this study.

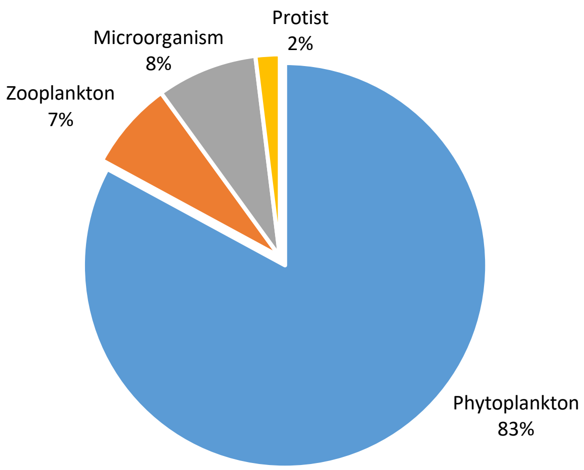

Fig. 2. Percentage of the sequences assigned to each of taxonomic groups.

An analysis of the relative abundances of the top 15 orders and families of organisms for the three replications of the 67 samples were shown in Fig. S1 and S2, respectively. However, approximately 70% of sequences cannot be assigned to genus level because the limitation of reference information in the SILVA database. Analysis of top 15 families of organisms indicated that the Mediophyceae, Ochromonadales and Chlorodendrales accounted for approximately 17.5%, 9.9% and 5.4% of all taxa, respectively. Analysis of variance (ANOVA) results indicated that no significant differences were observed for the relative family abundances among the sample replications ($p > 0.05$).

The OTU compositions of the different samples were analyzed according to beta diversity to reflect differences between samples using principal component analysis (PCA). Here, PCA uses variance decomposition to reflect the differences between multiple sets of data on a two-dimensional coordinate graph, where the coordinate axes are two eigenvalues that reflect the variance to the greatest extent. As such, samples with similar compositions were clustered in the PCA graph, as shown in Fig. 3A based on the sampling locations illustrated in Fig. 3B, which also showed the Shannon_E values for the individual sampling locations and the land use types of the study area.

The results indicated that significant differences exist between the sampling sites of upstream tributaries (e.g., s3, s2 and s1) and the sampling sites of the middle and lower mainstem, while differences were also observed between the urban (e.g., s22, s21, and s19) and mountainous sections (e.g., s6, s2, and s5) of the mainstem. However, the some sites were impacted by the reservoir located in the mainstem of upstream (e.g. s04 and s08). The spatial distribution of Shannon_E values presented the same pattern, where the Shannon_E value tended to gradually decrease with increasing disturbance from human activity from the upstream to the downstream regions, as reflected by increasing urban and industrial land use.

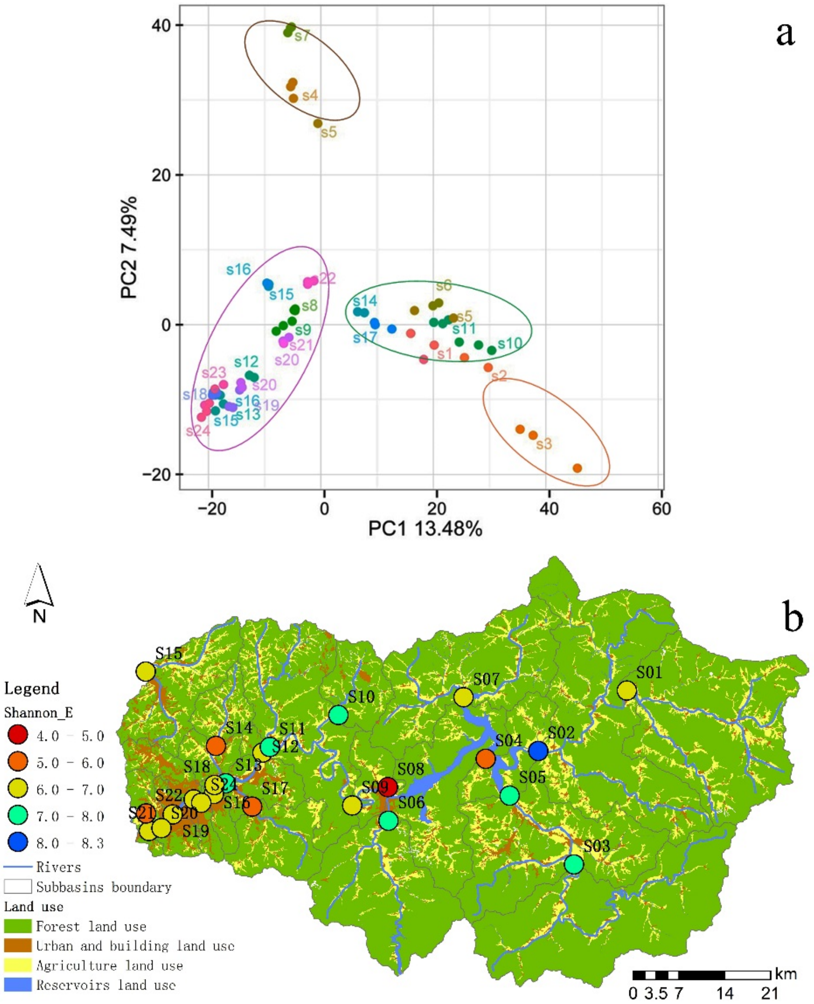

Fig. 3. (a) Principal component analysis graph for all samples based on the beta

diversity derived from eDNA metabarcoding data and (b) the spatial distribution of ecological status based on the Shannon Wiener index.

The diversity values measured according to the observed species, Shannon Wiener index, and Simpson index derived from eDNA metabarcoding and morphological identification data were normalized and compared, and the results were given in Fig. 4A, B, and C, respectively, for sample locations s01–s24. The results in Fig. 4A indicated that in most sites, the observed species values obtained based on eDNA metabarcoding were higher than the values based onmorphological data, because OTUs contained a greater number of taxa information. This difference decreased in the Shannon Wiener and Simpson index values, which demonstrated that the data obtained by the two methods reflect similar richness and evenness characteristics of community composition in most sampling sites (Fig. 4B and C). Fig. 4A also showed that 8 sites out of 24 were higher for morphological data than eDNA metabarcoding data, and most of these sites are located in the downstream of study area (e.g. s15, s18, s19 and s24), where a large number of *Cyclotella meneghiniana* were detected in morphological data. *Cyclotella meneghiniana* is a typical indicator of water pollution (Duong et al., 2008). This was proved by Fig. 4B and C, which showed that the Shannon and Simpson indices derived from morphological data were relatively low at these downstream sites. However, the ecological indices derived from eDNA data showed better consistency at these sites, which indicated that the difference between eDNA metabarcoding and morphological data may become larger in polluted river sections.

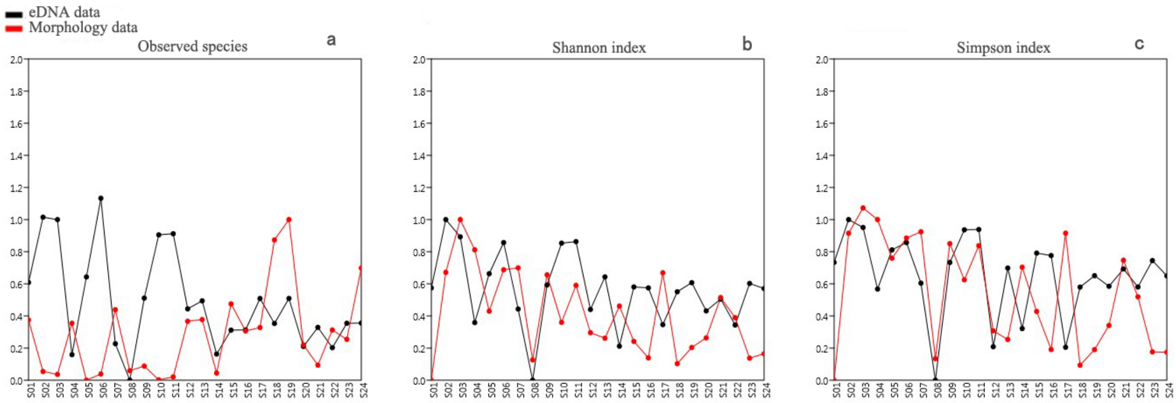

Fig. 4. Comparison of the three ecological index values derived from eDNA metabarcoding and morphology identification data.

3.2. Predictive performances and sensitivity analysis of SVM models

After optimizing the model parameters ($c = 10000$, $\varepsilon = 0.2$, and $\gamma = 0.025$) according to the methodology presented in Subsection 2.4, the nonlinear regression analysis results obtained by the SVM models for the three indices (Species_E, Shannon_E, Simpson_E) derived from eDNA metabarcoding data and the three indices (Species_M, Shannon_M, Simpson_M) derived from morphological identification data are presented in Fig. 5. The results indicated that, with the exception of Species_M (squared correlation coefficient $R^2 = 0.66$), the SVM models achieved good prediction performance, with R^2 values that were all greater than 0.80.

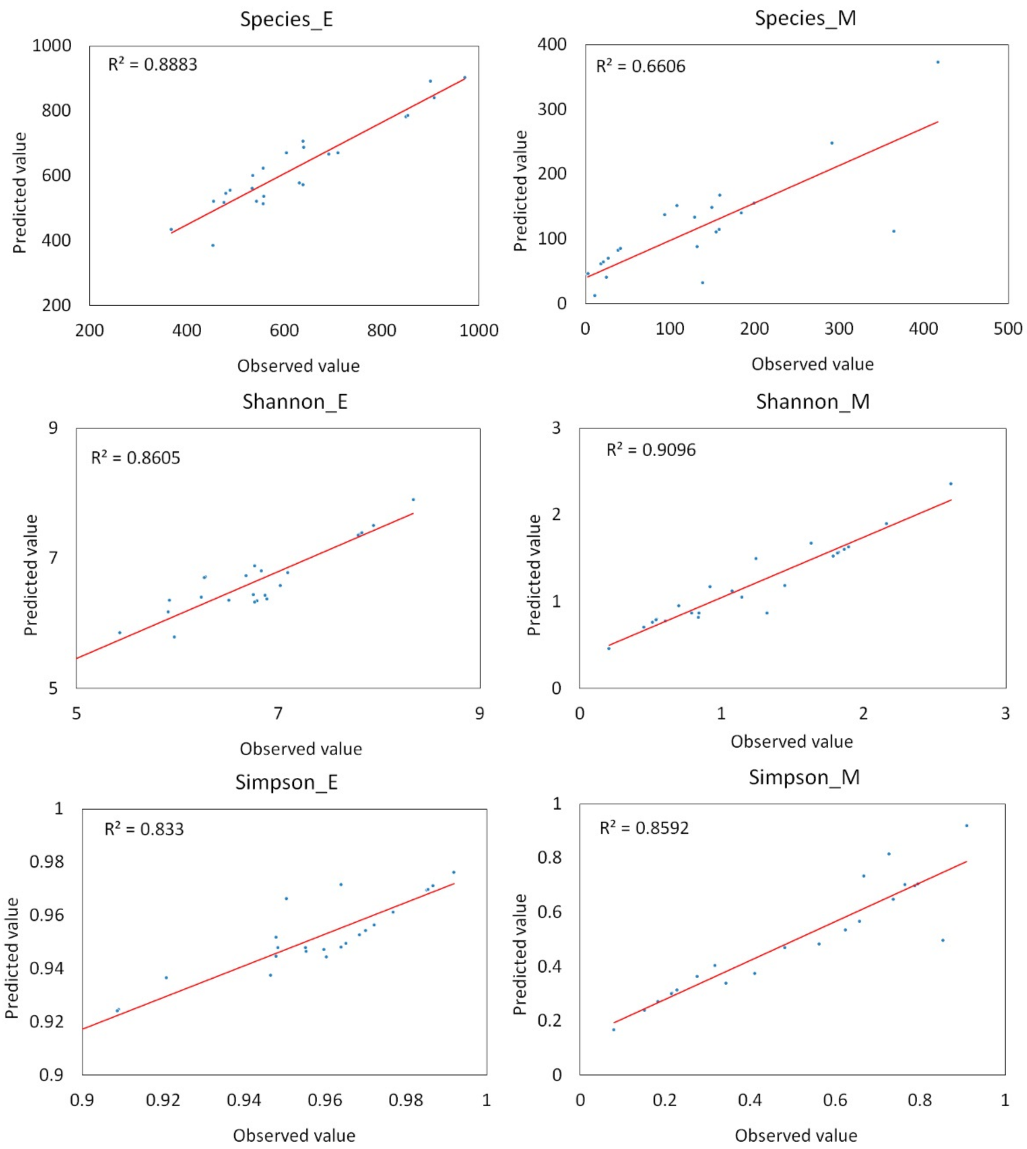

Fig. 5. Nonlinear regression fitting plots of the support vector machine (SVM) models for the measured values and predicted values of the three ecological indices.

The minimum values of MAPE for all samples (MAPE_ALL) and the minimum values of MAPE for the test samples (MAPE_TEST) obtained by 8-fold cross-validation indicated the accuracy of different models (Table 3). The results indicated that the MAPE_ALL values of the three most accurate SVM models obtained from eDNA metabarcoding data were in the order of Species_E > Shannon_E > Simpson_E, and the MAPE_ALL values of the three most accurate SVM models obtained from morphology identification data exhibited an equivalent pattern. Nevertheless, the SVM models constructed from the eDNA metabarcoding data had MAPE values that were much smaller than those of the models constructed from the morphological identification data whether based on MAPE_ALL or MAPE_TEST values. This indicated that the models constructed from eDNA metabarcoding data were more accurate than those constructed from the morphological identification data.

Table 3

Results of model selection using 8-fold cross-validation for each ecological index given in terms of the minimum values of MAPE for all samples (MAPE_ALL), and the minimum values of MAPE for the test samples (MAPE_TEST).

Ecological index	MAPE_ALL	MAPE_TEST
Index derived from eDNA metabarcoding data		
Species _E	9.06	6.72
Shannon_E	5.14	4.14
Simpson_E	1.33	0.75
Index derived from morphology identification data		
Species _M	183.96	49.57
Shannon_M	25.61	15.50
Simpson_M	25.37	20.00

The sensitivity of each ecological index to multiple stressors were varying (Table 4). For Species_E, the largest value of ΔMAPE = 1.12 was obtained for SS, indicating that the Species_E prediction was most sensitive to this variable. For Shannon_E, the largest value of ΔMAPE = 0.47 was obtained for SS, indicating that the Shannon_E prediction was most sensitive to this variable. For Simpson_E, the largest value of ΔMAPE = 0.05 was obtained for DO, indicating that the Simpson_E prediction was most sensitive to this variable. Likewise, we can determine that the Species_M prediction was most sensitive to DO (ΔMAPE = 21.79), the Shannon_M prediction was most sensitive to VP (ΔMAPE = 2.17), and the Simpson_M prediction was most sensitive to VP (ΔMAPE = 2.13). We also note from Table 4 that the magnitudes of the ΔMAPE values for the ecological indices obtained from DNA metabarcoding data are much smaller than those obtained from morphological identification data.

Table 4

Results of sensitivity analysis based on the change in MAPE values (ΔMAPE) for all samples with respect to the individual environmental variables.

Environmental variables	Species_E	Shannon_E	Simpson_E	Species_M	Shannon_M	Simpson_M
	ΔMAPE					
Catchment-scale variables						
F_cat	0.13	0.04	0	3.19	0.16	0.02
F_buf	0.62	0.44	0.03	15.47	1.59	0.18
A_cat	0.62	0.09	0	3.69	0.78	0.31
A_buf	0	0.07	0.01	4.67	0.55	0.07
U_cat	0.1	0.11	0.04	4.87	0.07	0.42
U_buf	0.29	0.01	0	4.2	0.62	0.16
Reach-scale variables						
EC	0.36	0.02	0.03	1.34	0.15	0.14
DO	0.16	0.35	0.05	21.79	0.06	0.29
pH	0.15	0.08	0.03	1.84	0.68	0.28
COD_{Mn}	0.14	0.04	0	8.27	0.13	0.47
BOD_5	0.24	0.16	0.02	0.68	0.11	1.39
NH_3-N	0.22	0.02	0.03	2.01	0.32	0.8
TN	0.34	0.19	0.04	0.63	1.12	0.64

TP	0.28	0.07	0.01	0.8	0.05	0.47
SS	1.12	0.47	0.02	14.22	0.72	0.34
VP	0.35	0.12	0.02	11.99	2.17	2.13

The variations in the ΔMAPE values for the ecological indices obtained from DNA metabarcoding and morphological identification data are more clearly shown in Fig. 6A. We note that, among all six ecological indices, DO, SS, and VP are the three environmental variables in the reach scale that most greatly affect the index value predictions. These are followed by F_buf, the variable in the catchment scale. A comparison of the average ΔMAPE values obtained for the environmental variables shown in Fig. 6B indicate that, with the exception of Shannon_M, the environmental variables at the reach scale have a greater impact on the ecological indices than those at the catchment scale.

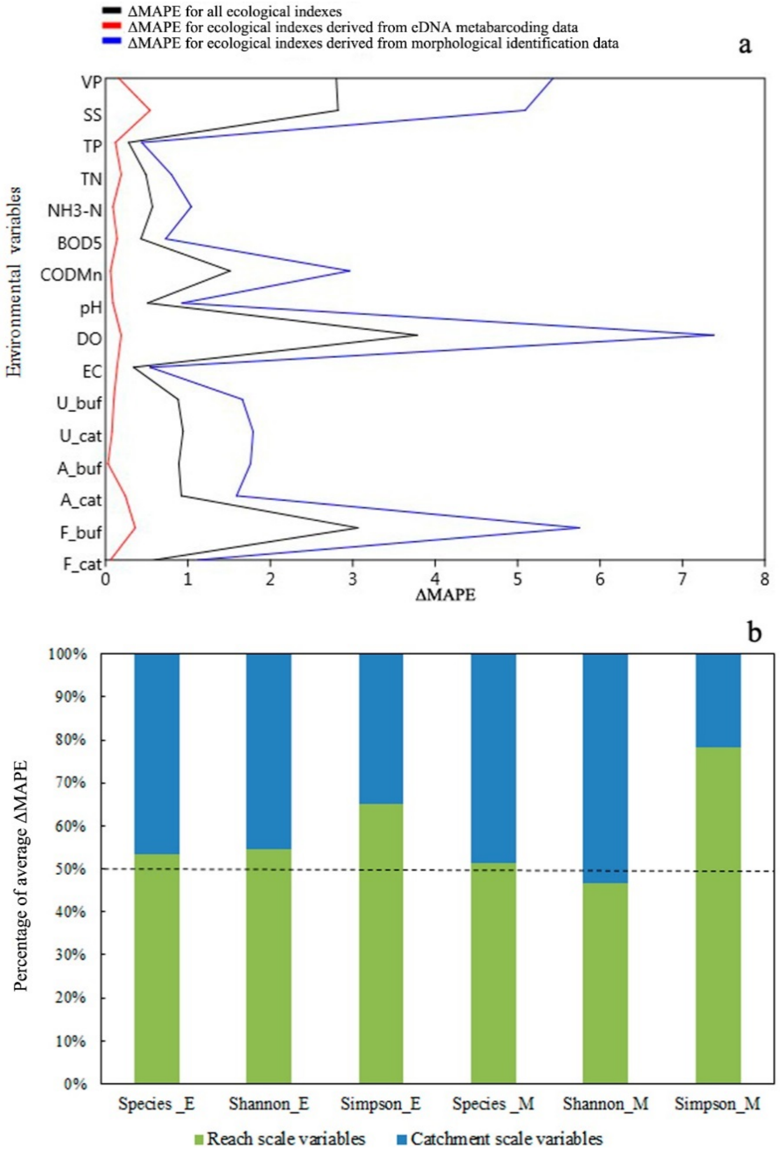

Fig. 6. (a) ΔMAPE values for each environmental variable in the sensitivity analysis and (b) a comparison of sensitivities between catchment and reach scale environmental variables.

4. Discussion

4.1. SVM model development and validation

The SVM models increase our understanding of the non-linear relationships between ecological status and multiple stressors on the one hand and the sensitivity of the ecological status to each stressor on the other. More importantly, the MAPE and high

R^2 values obtained by the SVM models demonstrate quantitatively that eDNA metabarcoding data provide modeling results that were more indicative of environmental degradation compared with morphological identification data. However, we must note that OTUs contained more taxa information than species, which may increase the uncertainty of the model comparison. However, the greater the number of OTUs does not necessarily mean the better model performance, because a larger data may also bring noise for modeling (Lu et al., 2018). In many biodiversity surveys and assessments, the concept of OUTs diversity has been roughly equated with the concept of species diversity (Caron and Hu, 2019), because OUTs use 3% sequence difference to distinguish species, which is an accepted standard in molecular biology techniques (Schloss and Handelsman, 2005). Previous study also showed that eDNA metabarcoding and morphological macroinvertebrate metrics are positively correlated and indicate the same key gradients in stream condition (Emilson et al., 2017).

Furthermore, this kind of uncertainty can be reduced by using some ecological indices (e.g. the Simpson and Shannon-Weiner indices), which represent the relative diversity of taxa. These indices have all been normalized before modeling, and the results showed that the normalized values of these ecological indices are relatively consistent in most sampling sites (Fig. 4). Therefore, the uncertainty of the model due to different classification levels can be reduced. In addition, our results were obtained with a relatively small training dataset, and increasing the number of samples or applying a larger sampling area can lead to the process of refining our predictive models. This is supported by a previous study, which has shown that the accuracy and stability of predictions increased exponentially with increasing sample size regardless of the type of ML algorithm adopted (Cui and Gong, 2018).

4.2. Ecological response derived from eDNA metabarcoding and morphological identification data to multiple stressors

Although the ecological indices obtained from morphological identification data exhibited good response relationships with multi-scale stressors, the models constructed from eDNA metabarcoding data provided better accuracy, as shown in Table 3. This is because the effective eDNA sequencing information includes a large number of intact and fragmentary organisms, and even includes the DNA information of many historically existing organisms. This is supported by a previous study, which found that the DNA information of some species may exist in water for up to one month after the removal of DNA release sources (Li et al., 2018b). In addition, it has been shown that eDNA metabarcoding data provide more integral information regarding biology, including the taxa and even the potential bioindicators of pollution, for example, the OTUs that dominate eDNA datasets in high mercury concentration do not need to be assigned taxonomically, which are typically overlooked in morphological identification (Frontalini et al., 2018). In conclusion, the biodiversity information contained in eDNA data is massive, and the large volume of data available may alleviate model prediction uncertainties caused by sample size limitations to some extent.

It is worth noting that eDNA metabarcoding data are not able to provide some information available from morphological identification data. For example, eDNA metabarcoding data provides no information regarding the morphological deformations of target organisms, which are often found in highly polluted environments, and are commonly used as evidence for heavy metal pollution (Yanko et al., 1998). Therefore, eDNA metabarcoding data cannot replace morphological analysis when studying the response of a particular species, but methods to detect change population of one or multiple organisms to environmental stressors by eDNA metabarcoding are developing, such as screening for functional genes, which may enable the eDNA metabarcoding to assess toxicological information (Zhang et al., 2019). This will widen the application of eDNA metabarcoding in environmental sciences.

4.3. Sensitivity differences with respect to catchment and reach scale stressors

The sensitivity analysis results indicated that DO, SS, VP, and F_buf have the greatest impact on the model predictions of diversity indices, and that environmental variables at the reach scale are more influential than that at the catchment scale, as shown in Table 4 and Fig. 6. This greater sensitivity of ecological status to reach-scale stressors can be explained by noting that disturbances in land use at the catchment scale affect the aquatic ecological status by generating non-point source pollutants, such as fertilizers, pesticides, and sewage irrigation, that enter water bodies, resulting in increased nutrition, bacteria, toxicity, and harmful substances, which means that the changes at the reach scale affect the ecological status of rivers directly (Meador and Goldstein, 2003; Piggott et al., 2012).

The DO is directly decreased under these degradating conditions (Mineau et al., 2015) Moreover, DO has been shown to be a key variable impacting the status of many aquatic species because it can affect the tolerance limit of organisms (Marshall and Elliott, 1998). In addition, sites with the highest DO level have also been shown to have the highest aquatic species diversity (Wilhm and McClintock, 1978). A previous study has also demonstrated that SS is critical to phytoplankton communities because phytoplankton growth requires photosynthesis, and light intensity in the photic zone has a significant negative correlation with SS (Van Duin et al., 2001). Finally, we note that urban and industrial land use in the urban section of the study area increased significantly since the work of (Fan et al., 2015), and this can be expected to have released toxic chemicals from industrial pollution, such as VP, into water bodies. In this regard, the photosynthetic activity parameters of algae have been shown to have a negative dose-response relationship to phenol toxicity (Kottuparambil et al., 2014). Therefore, VP represents another critical environmental variable impacting the status of many aquatic species.

5. Conclusion

The present study compared the ability of eDNA metabarcoding data and morphological identification data to reflect the nonlinear impact of multiple environmental stressors on aquatic ecosystems by employing both sets of data in SVM models corresponding to three ecological indices (i.e. observed species, Shannon Wiener index, and Simpson index). Analysis of the environmental variables at the catchment and reach scales of the study area indicated that most of the variables exceeded their natural thresholds at some of the sampling sites, and became a complex of simultaneously interacting stressors affecting the ecological status of the river. The SVM models constructed from eDNA metabarcoding data (MAPE = 3.87) provided more accurate predictions than the SVM models constructed from morphological identification data (MAPE = 28.36), revealing that the eDNA metabarcoding data better reflected ecological conditions. As such, the present work helps to explore the potential applications of eDNA technology in the monitoring and management of aquatic ecosystems. In addition, the sensitivity of SVM model predictions of aquatic ecosystem diversity to catchment-scale and reach-scale stressors was evaluated, and the stressors having the greatest impact on the ecological status of rivers were identified. These results indicated that the model predictions were more sensitive to the environmental variables at the reach scale than those at the catchment scale. In addition, DO, SS, VP, and F_buf were found to be the most influential variables impacting the ecological status of the river.

Acknowledgments

The authors thank Dr. Xuwang Yin and Dr. Jianing Lin for their taxonomic experts consultation on morphological identification. The authors also thank the anonymous referees for their valuable comments regarding the manuscript.

Funding

This work was financially supported by the Fundamental Research Funds for Central Public Welfare Scientific Research Institutes of China (Grant No. 2019YSKY-021), and the Major Science and Technology Program for Water Pollution Control and Treatment (Grant No. 2017ZX07301002-01).

References

Al-Mukhtar, M., 2019. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. Environmental Monitoring and Assessment 191 (11), 673. https://doi.org/10.1007/s10661-019-7821-5.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W. and Huse, S.M., 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4 (7), e6372. https://doi.org/10.1371/journal.pone.0006372.

Balfer, J. and Bajorath, J., 2015. Systematic artifacts in support vector regression-based compound potency prediction revealed by statistical and activity landscape analysis. PLoS One 10 (3), e0119301. https://doi.org/10.1371/journal.pone.0119301.

Bernhardt, E.S., Lutz, B.D., King, R.S., Fay, J.P., Carter, C.E., Helton, A.M., Campagna, D. and Amos, J., 2012. How many mountains can we mine? Assessing the regional degradation of central Appalachian rivers by surface coal mining. Environmental Science & Technology 46 (15), 8115–8122. https://doi.org/10.1021/es301144q.

Bouboulis, P., Theodoridis, S., Mavroforakis, C. and Evaggelatou-Dalla, L., 2015. Complex support vector machines for regression and quaternary classification. IEEE Transactions on Neural Networks and Learning Systems 26 (6), 1260–1274. https://doi.org/10.1109/TNNLS.2014.2336679.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7 (5), 335–336. https://doi.org/10.1038/nmeth.f.303.

Caron, D.A. and Hu, S.K., 2019. Are we overestimating protistan diversity in nature? Trends in Microbiology 27 (3), 197–205. https://doi.org/10.1016/j.tim.2018.10.009.

Cherkassky, V. and Ma, Y.Q., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17 (1), 113–126. https://doi.org/10.1016/S0893-6080(03)00169-2.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F. and Klove, B., 2018. River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Science of the Total Environment 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen,

T. and Pawlowski, J., 2017. Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. Environmental Science & Technology 51 (16), 9118–9126. https://doi.org/10.1021/acs.est.7b01518.

Cui, Z.X. and Gong, G.L., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage 178, 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N. and Probert, I., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348 (6237), 1261605. https://doi.org/10.1126/science.1261605.

Duong, T. T., Morin, S., Herlory, O., Feurtet-Mazel, A., Coste, M., 2008. Seasonal effects of cadmium accumulation in periphytic diatom communities of freshwater biofilms. Aquatic Toxicology 90, 19–28. https://doi.org/ 10.1016/j.aquatox.2008.07.012.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods 10 (10), 996–998. https://doi.org/10.1038/nmeth.2604.

Emilson, C.E., Thompson, D.G., Venier, L.A., Porter, T.M., Swystun, T., Chartrand, D., Capell, S. and Hajibabaei, M., 2017. DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. Scientific Reports 7, 1–11. https://doi.org/10.1038/s41598-017-13157-x.

Fan, J.T., Semenzin, E., Meng, W., Giubilato, E., Zhang, Y., Critto, A., Zabeo, A., Zhou, Y., Ding, S. and Wan, J., 2015. Ecological status classification of the Taizi River Basin, China: a comparison of integrated risk assessment approaches. Environmental

Science and Pollution Research 22 (19), 14738–14754. https://doi.org/10.1007/s11356-015-4629-x.

Fernandez, S., Rodriguez, S., Martinez, J.L., Borrell, Y.J., Ardura, A. and Garcia-Vazquez, E., 2018. Evaluating freshwater macroinvertebrates from eDNA metabarcoding: a river Nalón case study. PLoS One 13 (8), e0201741. https://doi.org/10.1371/journal.pone.0201741.

Franzo, A. and Del Negro, P., 2019. Functional diversity of free-living nematodes in river lagoons: can biological traits analysis (BTA) integrate traditional taxonomic-based approaches as a monitoring tool? Marine Environmental Research 145, 164–176. https://doi.org/10.1016/j.marenvres.2019.02.015.

Frontalini, F., Greco, M., Di Bella, L., Lejzerowicz, F., Reo, E., Caruso, A., Cosentino, C., Maccotta, A., Scopelliti, G. and Nardelli, M.P., 2018. Assessing the effect of mercury pollution on cultured benthic foraminifera community using morphological and eDNA metabarcoding approaches. Marine Pollution Bulletin 129 (2), 512–524. https://doi.org/10.1016/j.marpolbul.2017.10.022.

He, X., Sutherland, T.F., Pawlowski, J. and Abbott, C.L., 2019. Responses of foraminifera communities to aquaculture-derived organic enrichment as revealed by environmental DNA metabarcoding. Molecular Ecology 28 (5), 1138–1153. https://doi.org/10.1111/mec.15007.

Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B. and Kahlert, M., 2018. Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. Water Research 138, 192–205. https://doi.org/10.1016/j.watres.2018.03.003.

Hu, H.J. and Wei, Y.X. (2006) The Freshwater Algae of China, Systematics, Taxonomy and Ecology, Science Press, Beijing.

Humbert, J.-F., Quiblier, C. and Gugger, M., 2010. Molecular approaches for monitoring potentially toxic marine and freshwater phytoplankton species. Analytical and Bioanalytical Chemistry 397 (5), 1723–1732. https://doi.org/10.1007/s00216-010-3642-7.

Jarman, S.N., Berry, O. and Bunce, M., 2018. The value of environmental DNA biobanking for long-term biomonitoring. Nature Ecology & Evolution 2 (8), 1192–1193. https://doi.org/10.1038/s41559-018-0614-3.

Ji, Y.Q., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P. and Edwards, F.A., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16 (10), 1245–1257. https://doi.org/10.1111/ele.12162.

Jones, F.C., Plewes, R., Murison, L., MacDougall, M.J., Sinclair, S., Davies, C., Bailey, J.L., Richardson, M. and Gunn, J., 2017. Random forests as cumulative effects models: a case study of lakes and rivers in Muskoka, Canada. Journal of Environmental Management 201, 407–424. https://doi.org/10.1016/j.jenvman.2017.06.011.

Kefford, B.J., Marchant, R., Schäfer, R.B., Metzeling, L., Dunlop, J.E., Choy, S.C. and Goonan, P., 2011. The definition of species richness used by species sensitivity distributions approximates observed effects of salinity on stream macroinvertebrates. Environmental Pollution 159 (1), 302–310. https://doi.org/10.1016/j.envpol.2010.08.025.

Keylock, C.J., 2005. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. Oikos 109 (1), 203–207. https://doi.org/10.1111/j.0030-1299.2005.13735.x.

Kottuparambil, S., Kim, Y.J., Choi, H., Kim, M.S., Park, A., Park, J., Shin, W. and Han, T., 2014. A rapid phenol toxicity test based on photosynthesis and movement of the freshwater flagellate, Euglena agilis Carter. Aquatic Toxicology 155, 9–14. https://doi.org/10.1016/j.aquatox.2014.05.014.

Laroche, O., Wood, S.A., Tremblay, L.A., Ellis, J.I., Lejzerowicz, F., Pawlowski, J., Lear, G., Atalah, J. and Pochon, X., 2016. First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. Marine Environmental Research 120, 225–235. https://doi.org/10.1016/j.marenvres.2016.08.009.

Li, F., Peng, Y., Fang, W., Altermatt, F., Xie, Y., Yang, J. and Zhang, X., 2018a. Application of environmental DNA metabarcoding for predicting anthropogenic pollution in rivers. Environmental Science & Technology 52 (20), 11708–11719. https://doi.org/10.1021/acs.est.8b03869.

Li, M., Shan, X.J., Wang, W.J., Dai, F.Q., Lu, D. and Wu, H.H., 2018b. Study on the retention time of environmental DNA of fenneropenaeus chinensis in water. Progress in Fishery Sciences, https://doi.org/10.19663/j.issn2095-9869.20180906005.

Lu, X.J., Ming, L., Liu, W.B. and Li, H.X., 2018. Probabilistic regularized extreme learning machine for robust modeling of noise data. IEEE Transactions on Cybernetics 48 (8), 2368–2377. https://doi.org/10.1109/TCYB.2017.2738060.

Marshall, S. and Elliott, M., 1998. Environmental influences on the fish assemblage of the Humber estuary, UK. Estuarine, Coastal and Shelf Science 46 (2), 175–184. https://doi.org/10.1006/ecss.1997.0268.

Marvuglia, A., Kanevski, M. and Benetto, E., 2015. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. Environment International 83, 72–85.

https://doi.org/10.1016/j.envint.2015.05.011.

Meador, M.R. and Goldstein, R.M., 2003. Assessing water quality at large geographic scales: relations among land use, water physicochemistry, riparian condition, and fish community structure. Environmental Management 31 (4), 0504–0517. https://doi.org/10.1007/s00267-002-2805-5.

Meissner, T., Sures, B. and Feld, C.K., 2019. Multiple stressors and the role of hydrology on benthic invertebrates in mountainous streams. Science of the Total Environment 663, 841–851. https://doi.org/10.1016/j.scitotenv.2019.01.288.

Mineau, M.M., Wollheim, W.M. and Stewart, R.J., 2015. An index to characterize the spatial distribution of land use within watersheds and implications for river network nutrient removal and export. Geophysical Research Letters 42 (16), 6688–6695.

Ministry of Water Resources, Quality Standards for Surface Water Resources. http://www.mwr.gov.cn/zwgk/zfxxgkml/201301/t20130124_964289.html. (1994, accessed 24 January 2013)

Mize, E.L., Erickson, R.A., Merkes, C.M., Berndt, N., Bockrath, K., Credico, J., Grueneis, N., Merry, J., Mosel, K. and Tuttle-Lau, M., 2019. Refinement of eDNA as an early monitoring tool at the landscape-level: Study design considerations. Ecological Applications 29 (6), e01951. https://doi.org/10.1002/eap.1951.

Mori, N., Debeljak, B., Skerjanec, M., Simcic, T., Kanduc, T. and Brancelj, A., 2019. Modelling the effects of multiple stressors on respiration and microbial biomass in the hyporheic zone using decision trees. Water Research 149, 9–20. https://doi.org/10.1016/j.watres.2018.10.093.

Olson, J.R., 2018. Predicting combined effects of land use and climate change on river and stream salinity. Philosophical Transactions of the Royal Society B: Biological Sciences 374 (1764), 20180005. https://doi.org/doi:10.1098/rstb.2018.0005.

Park, Y., Cho, K.H., Park, J., Cha, S.M. and Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Science of the Total Environment 502, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T. and Domaizon, I., 2018. The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. Science of the Total Environment 637, 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002.

Piette, E.R. and Moore, J.H., 2018. Improving machine learning reproducibility in genetic association studies with proportional instance cross validation (PICV). BioData Mining 11, 6. https://doi.org/10.1186/s13040-018-0167-7.

Piggott, J.J., Lange, K., Townsend, C.R. and Matthaei, C.D., 2012. Multiple stressors in agricultural streams: a mesocosm study of interactions among raised water temperature, sediment addition and nutrient enrichment. PLoS One 7 (11), e49873. https://doi.org/10.1371/journal.pone.0049873.

Piggott, J.J., Townsend, C.R. and Matthaei, C.D., 2015. Reconceptualizing synergism and antagonism among multiple stressors. Ecology and Evolution 5 (7), 1538–1547. https://doi.org/10.1002/ece3.1465.

Romero, F., Sabater, S., Timoner, X. and Acuña, V., 2018. Multistressor effects on river biofilms under global change conditions. Science of the Total Environment 627, 1–10. https://doi.org/10.1016/j.scitotenv.2018.01.161.

Rossel, R.A.V. and Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54. https://doi.org/0.1016/j.geoderma.2009.12.025.

Schloss, P.D. and Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology 71 (3), 1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005.

Strong, W.L., 2016. Biased richness and evenness relationships within Shannon–Wiener index values. Ecological Indicators 67, 703–713. https://doi.org/10.1016/j.ecolind.2016.03.043.

Torija, A.J. and Ruiz, D.P., 2015. A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods. Science of the Total Environment 505, 680–693.

Treusch, A. H., Demirhilton, E., Vergin, K. L., Worden, A. Z., Carlson, C. A., Donatz, M. G., ... & Giovannoni, S. J., 2012. Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. The ISME Journal, 6(3), 481-492. https://doi.org/10.1038/ismej.2011.117

Van Duin, E.H.S., Blom, G., Los, F.J., Maffione, R., Zimmerman, R., Cerco, C.F., Dortch, M. and Best, E.P.H., 2001. Modeling underwater light climate in relation to sedimentation, resuspension, water quality and autotrophic growth. Hydrobiologia 444 (1–3), 25–42. https://doi.org/10.1023/A:1017512614680.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Transactions on Neural Networks 10 (5), 988–999. https://doi.org/10.1109/72.788640.

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180 (Part A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.

Visco, J.A., Apotheloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L. and Pawlowski, J., 2015. Environmental monitoring: inferring the diatom index from next-generation sequencing data. Environmental Science & Technology 49 (13), 7597–7605. https://doi.org/10.1021/es506158m.

Von Schiller, D., Acuna, V., Aristi, I., Arroita, M., Basaguren, A., Bellin, A., Boyero, L., Butturini, A., Ginebreda, A. and Kalogianni, E., 2017. River ecosystem processes: a synthesis of approaches, criteria of use and sensitivity to environmental stressors. Science of the Total Environment 596, 465–480. https://doi.org/10.1016/j.scitotenv.2017.04.081.

Were, K., Bui, D.T., Dick, O.B. and Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028.

Wilhm, J. and McClintock, N., 1978. Dissolved oxygen concentration and diversity of benthic macroinvertebrates in an artificially destratified lake. Hydrobiologia 57 (2), 163–166. https://doi.org/10.1007/BF00016460.

Yanko, V., Ahmad, M. and Kaminski, M., 1998. Morphological deformities of benthic foraminiferal tests in response to pollution by heavy metals: implications for pollution monitoring. The Journal of Foraminiferal Research 28 (3), 177–200.

Yuan, J., Li, M. and Lin, S., 2015. An improved DNA extraction method for efficient and quantitative recovery of phytoplankton diversity in natural assemblages. PLoS One 10 (7), e0133060. https://doi.org/10.1371/journal.pone.0133060.

Zhang, L., Calvo-Bado, L., Murray, A.K., Amos, G.C.A., Hawkey, P.M., Wellington, E.M. and Gaze, W.H., 2019. Novel clinically relevant antibiotic resistance genes associated with sewage sludge and industrial waste streams revealed by functional metagenomic screening. Environment International 132, 105120. https://doi.org/10.1016/j.envint.2019.105120.

Zhang, X., 2019. Environmental DNA shaping a new era of ecotoxicological research. Environmental Science & Technology 53 (10), 5605–5612. https://doi.org/10.1021/acs.est.8b06631.