Aquat Ecol (2014) 48:297–312
DOI 10.1007/s10452-014-9484-1

A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices

Jing Zhao · Jie Cao · Siquan Tian · Yong Chen · Shouyu Zhang · Zhenhua Wang · Xijie Zhou

Received: 26 November 2013 / Accepted: 10 May 2014 / Published online: 17 May 2014

Abstract Various regression methods can be used to quantify the relationships between fish populations and their environment. Strong correlations often existing between environmental variables, however, can cause multicollinearity, resulting in overfitting in modeling. This study compares the performance of a regular generalized additive model (GAM) with raw environmental variables as explanatory variables (regular GAM) and a GAM based on principal component analysis (PCA-based GAM) in modeling the relationship between fish richness and diversity indices and environmental variables. The PCA-based GAM tended to perform better than the regular GAM in cross-validation tests, showing a higher prediction precision. The variables identified being significant in modeling differed between the two models, and differences between the two models were also found in the scope and range of predicted richness and diversity indices for demersal fish community. This implies that choices between these two statistical modeling approaches can lead to different ecological interpretations of the relationships between fish communities and their habitats. This study suggests that the PCA-based GAM is a better approach than the original GAM in quantifying the relationship between fish richness and diversity indices and environmental variables if the environmental variables are highly correlated.

Keywords Generalized additive model · Principle component analysis · Fish richness and diversity indices · Habitat · Ma'an Archipelago

Handling Editor: Thomas Mehner.

J. Zhao · S. Tian · S. Zhang (✉) · Z. Wang · X. Zhou
College of Marine Sciences, Shanghai Ocean University, 999 Huchenghuan Road, Lingang, Shanghai 201306, China
e-mail: epaperzsy@gmail.com

J. Cao · Y. Chen
School of Marine Sciences, University of Maine, Orono, ME 04469, USA

Introduction

Fish community structure plays an important role in the dynamics of marine ecosystems which often support valuable fisheries (Liu et al. 2013). Understanding of the dynamic relationship between fish communities and environment helps identify the key variables regulating fish communities and is important in the evaluation of potential impacts of environmental changes on fish population dynamics (Araújo et al. 2002; Fischer et al. 2013; Hoeinghaus et al. 2007; Macedo-Soares et al. 2012). Such knowledge is critical for the protection of species biodiversity and the development of appropriate conservation priorities (Liu et al. 2013).

Fish community structures are known to be influenced by many factors such as bottom type, temperature, depth, and salinity (Blaber and Blaber 1980;

Toepfer et al. 1998; Jaureguizar et al. 2004; Love and May 2007; Sternberg and Kennard 2013; Fortes et al. 2014). Fish diversity tends to be higher in a more complex habitat (Friedlander 2001). Because of close relationships between fish distribution and environmental variables, ecologists often develop various statistical models to study spatial and temporal dynamics of fish distributions along environmental gradients (Yee 2006), and environment variables are often used to explain the distribution and structure of fish community (Annoni et al. 1997; Akin et al. 2005).

Multivariate statistical methods such as canonical correspondence analysis (CCA), detrended CCA (DCCA), partial CCA (PCCA), and principal component analysis (PCA) are commonly used in studying habitat-fish community interactions because of the multivariate nature of environmental and fish community data (Ahmadi-Nedushan et al. 2006; Ellis et al. 2006; Barquín and Death 2009; Jordaan et al. 2010; Dubey et al. 2012; Palamara et al. 2012; Jyväsjärvi et al. 2013). Kleyer et al. (2012) discussed the adaptability of models at fish population and community levels, and suggested that choices of the methods needed to be data-dependent. Redundancy analysis and outlying mean index with generalized additive model (GAM) are considered to be good choices for modeling of fish community and habitats. Ahmadi-Nedushan et al. (2006) reviewed different statistical models widely used in analyzing relationships between species and habitats, and concluded that the distribution of species was influenced by multiple environmental drivers and that multivariable statistical models, such as ordinary multiple linear regression, logistic regression, generalized linear models (GLM), and GAM, were more suitable for modeling the relationship between fish species and environment. Because of complex relationships between fish community and environment, it is difficult to identify whether the relationship is linear. Thus, models such as GAM that allow for nonlinear responses may be more suitable for exploring the relationships between fish community and environments (Hastie and Tibshirani 1990; Leathwick et al. 2006; Ptacnik et al. 2008; Chang et al. 2010; Schmiing et al. 2013).

The relationship between the fish community and their environments is likely to be complex, and many environmental variables tend to be strongly correlated with each other (Pérez et al. 1998; Saraceno et al. 2005; Ribeiro et al. 2012), resulting in multicollinearity which often lead to model overfitting, low precision in predication, and great uncertainty in the selection of habitat variables in modeling. Correlation analyses and variance inflation factor (VIF) can be used to identify high multicollinearity between variables (Emery and Thomson 2001). To overcome the problem of multicollinearity in the explanatory environmental variables, several approaches can be used for the selection of appropriate explanatory variables in habitat modeling. These approaches include removal of predictor variances, stepwise method, ordinary least squares, residual and sequential regression, and ridge regression (Toepfer et al. 1998; Straka et al. 2012; Kroll and Song 2013). However, those methods of removal of predictor variables may miss some important environmental variables. Stepwise method is often used to identify main explanatory variables (Francis et al. 2005; Maggini et al. 2006). However, the inclusion of highly correlated variables in a stepwise modeling can introduce large uncertainty and reduces prediction precision (Annoni et al. 1997; Francis et al. 2005).

Multivariate statistical methods such as PCA were suggested for summarizing the environmental variables prior to their inclusions in habitat modeling (Afifi and Clark 1996; Buisson et al. 2008). The newly derived principal components (PCs) for the environmental variables are then used as the explanatory variables in the GAM analysis. Such a PCA-based GAM can remove correlations between the explanatory variables in GAM and can balance the need of including all important environmental variables while removing impacts of highly correlated environmental variables in habitat modeling. The new explanatory variables of GAM derived from PCA have better statistical property (i.e., uncorrelated explanatory variables; Bierman et al. 2011) and can capture most information inherent in the original data (Ahmadi-Nedushan et al. 2006). However, the use of PCs as the explanatory variables in GAMs may complicate model interpretations because biological meanings of PCs may not be as obvious as the original variables. No studies have been performed to compare the PCA-based GAM approach with a more traditional GAM method which uses environmental variables directly as the explanatory variables.

In this study, we applied a regular GAM with raw environmental variables as the explanatory variables (referred to as "regular GAM" in this study) to the data collected in a coastal marine ecosystem to develop a

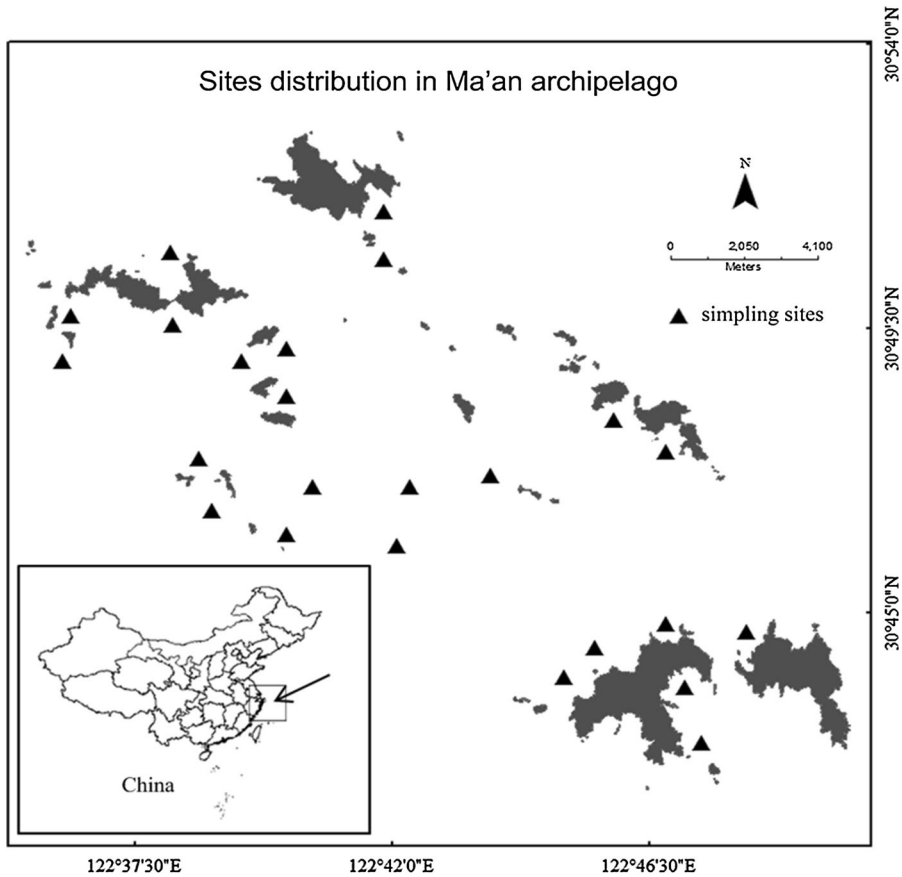

Fig. 1 The study area and sampling sites

habitat-fish community model. We then applied a PCA to analyze the same set of environmental data and used the resultant PCs as the explanatory variables in the subsequent GAM (referred to as "PCA-based GAM"). We compared differences in the quantification of spatial distribution of fish community using the regular GAM and PCA-based GAM, and discussed their ecological implications.

Materials and methods

Data collection

The study site, with a total area of 549 km^2, is located at Ma'an Archipelago, Zhejiang Province, China (Fig. 1). Water depth ranges from several meters to more than 30 m. The study area includes both natural and artificial habitats, such as rock reef, sand, mud, and mussels, and is characterized by spatial heterogeneity in the habitat. Main sediments in the center of the study area are mud and sand, and the surrounding area tends to be mud. Other sediments in the study area include rocks and sands.

The data of demersal fish community and environment variables used for modeling were collected monthly during 2009 from 24 sampling sites which were identified roughly through the stratification based on the sediment types (Fig. 1). The number of sampling sites in each type of the sediment was set approximately proportional to the area of the sediments. A stationary bottom gillnet was used for fishing (Wang et al. 2012). Two gillnets were set at the center of each selected site, one with a height of 1.8 m consisting of four panels of different mesh sizes (25, 34, 43, and 58 mm) with each panel being 15 m long (a total of 60 m long) and the other with a height of 2.4 m consisting of four panels of different mesh sizes (50, 60, 70, and 80 mm) with each panel being 30 m in length (a total length of 120 m). The gillnets were set for 24 h (mean 23.6 ± 2.5 h) at each sampling site to eliminate differences in sampling efficiency between

days and nights. Measurements of body length, weight, sex, and stomach of each sampled fish were completed within 24 h after catching the fish. The global positioning system (GPS) was used to record the sampling location, and conductivity, temperature, and depth sensor (CTD) was used to measure the environmental variables including depth, temperature, salinity, chlorophyll, turbidity, and oxygen.

The 2009 survey data were used to develop the model which was then used to predict the distribution of demersal fish community using the environmental data collected in June, August, October, and December in 2012, and in February and April in 2013. The study area was divided into 77 sampling grids with a size of $1' \times 1'$ longitude × latitude to collect the environmental variables for prediction. ArcGIS10 was used to interpolate and plot the distribution of predicted fish species richness and diversity indices.

GAM models

In this study, two indices, *Margalefs index* and *Shannon index*, were used to describe the structure of demersal fish community in the study area (Shannon and Weaver 1948; Margalef 1958). The *Margalefs index* (Margalef 1958) describes the richness of fish community, and the *Shannon index* (Shannon and Weaver 1948) quantifies the diversity of fish community. A Gaussian error distribution was used as the logic link function for both the original GAM and PCA-based GAM approaches. The original GAM for the *Margalefs index* or the *Shannon index* can be described as follows:

$$\begin{aligned} &Logit(Margalefs\ index\ or\ Shannon\ index) \\ &= s(Lon) + s(Lat) + s(temp) + s(sal) + s(chl) \\ &\quad + s(oxy) + s(dep) + s(turb) + month + type + \varepsilon \end{aligned} \tag{1}$$

where Logit is data transformation; s is spline smoother; Lon is longitude; Lat is latitude; temp is temperature; sal is salinity; chl is chlorophyll; oxy is oxygen; dep is depth; turb is turbidity; month ranges from January to December; and type is sediments, including rock reef, sand, mud, mud and sand, mussel, artificial reef, and aquaculture cage. The VIF was used to estimate the multicollinearity of environment variables, and the multicollinearity is considered severe when VIF values are higher than 10 (Neter et al. 1996; Graham 2003).

The numerical variables were standardized by subtracting their respective means and then divided by their respective standard deviations. For the PCA-based GAM approach, we first applied PCA to analyze all the standardized environmental variables except for sediment (and month) because the PCA may not work well with categorical variables. The new variables derived in the PCA are PCs which are the linear combinations of original environmental variables and tend to be uncorrelated from each other. The PCA-based GAM for the *Margalefs index* or the *Shannon index* can be written as follows:

$$\begin{aligned} &Logit(Margalefs\ index\ or\ Shannon\ index) \\ &= s(Lon) + s(Lat) + \sum_{i=1}^{n} s(comp\ i) + month \\ &\quad + type + \varepsilon \end{aligned} \tag{2}$$

where comp i are PCs, n is the number of PCs which are chosen in the PCA-based GAM. The GAM and PCA were conducted using the mgcv package (Wood 2011) in the R2.15.0 software (http://www.r-project.org/). The significance level was set at 0.05 in the modeling.

Model validation

In order to evaluate performance of the models, we conducted a cross-validation test, in which 80 % of the data were randomly selected for building the model which was then used to make the prediction for the remaining 20 % of the data. This subsampling process was repeated for 100 times for each cross-validation. The correlation coefficient between observed and predicted fish community indices (i.e., *Margalefs index* or *Shannon index*) for the 20 % of the data (i.e., test data) was calculated for each run. A simple linear regression model was fitted to the predicted index (P') and the observed index (P):

$$P = a + b(P') \tag{3}$$

where parameter a indicates systematic bias of the predicted indices, and b is a slope parameter. A value of parameter b not significantly different from 1 indicates that observed index and predicted index have similar spatial patterns (Chang et al. 2010).

Comparison of model performance

The main objective of this study was to compare performance of the two models, the regular GAM and

the PCA-based GAM, in predicting the *Margalefs index* and *Shannon index* which were used to quantify demersal fish community structure in the study area. We used the following two measures for the comparisons: (1) the regression parameters and correlations coefficient between observed and predicted values in the cross-validation; and (2) the proportion of the 100 simulation runs in which the same environmental variables were identified as significant for the regular GAM or PCA-based GAM. For the second measure, if the proportion of 100 runs in which an environmental variable was identified as significant is higher than 0.5, the variable was considered significant in influencing fish community. The first measure quantifies the predictive power of the model, and the second measure describes the consistency and stability of the models among the 100 simulation runs for which different sets of the data were subsampled.

Results

The correlation between environment variables

Strong correlations were found between the environmental variables (e.g., between temperature and salinity and between turbidity and chlorophyll) included in the study (Table 1). The VIF of numerical variables was ranged from 4 to 496, indicating high correlations between these variables. Five variables had VIF values greater than 10: chlorophyll (36), oxygen (11), turbidity (40), month (496), and habitat type (54). The correlations and VIFs showed the existence of strong multicollinearity if these environmental variables were used as the explanatory variables in GAMs. Temperature showed a clear seasonal change and was highest in August (Fig. 2). Meanwhile, salinity was lower during August to October but did not change widely during a year. Seasonal changes in oxygen showed a "V" pattern, with lower values occurring during June to October. Depth seemed to be related to habitat types. The artificial reef habitat tended to have a higher average depth, compared to the other habitats.

Principal component analysis

Together, the first three PCs explained 85.5 % (83.5–87.5 %) of variance on average over the 100 simulation runs with the first to third PCs explaining 40.1, 28.4, and 17.0 %, respectively. Seventy-six runs out of the 100 simulation runs of PCA showed that temperature had the highest positive loading and turbidity had the highest negative loading in the first PC (Fig. 3). Thus, the first PC mainly reflected temperature and turbidity. Salinity and chlorophyll had high loadings in the second PC, while water depth took the highest loading in the third PC. Moreover, the rocky reef habitat mainly had higher loadings of the second PC compared to the other habitats. The artificial reef habitat had a high and positive loading of the third PC (Fig. 4).

The general additive models

The deviance of demersal fish *Margalefs index* and *Shannon index* explained in the regular GAM and PCA-based GAM varied greatly over the 100 simulation runs. The average deviance of species richness index (i.e., *Margalefs index*) explained was 83.7 % (71.7–90 %) for the regular GAM model and 80.6 % (68.1–93.4 %) for the PCA-based GAM, and the Student's t test showed a significant difference between the two approaches ($p = 0.006$). The average deviance of species diversity index (i.e., *Shannon index*) explained was 83.3 % (73.9–90.5 %) and 80.5 % (68.9–92.1 %), respectively, for the regular GAM and PCA-based GAM, and the Student's t test showed a significant difference between the two methods ($p = 0.0017$). The average AIC of the regular GAM for the richness index was 3.1 (−86.8 to 45.8), and 22.9 (−39.7 to 49.8) for the PCA-based GAM. The average AIC of regular GAM for diversity index was −10.9 (−59.6 to 6.8) and −8.7 (−58.6 to 9.9) for PCA-based GAM.

Sediment and chlorophyll were identified as the most significant variables in most simulation runs (>50 %) for the regular GAM of richness index (Table 2). Month was the only variable identified as significant in influencing the diversity index for the regular GAM. The other environmental variables such as salinity and depth were found significant in some runs. In contrast, the month and sediment variables were identified as significant variables in most simulation runs for the PCA-based GAM for both the richness and diversity indices (Table 2).

Model validation

The average of correlation coefficients between the observed and predicted indices for the test data over

Table 1 The correlation coefficients between the environmental variables

	Temperature	Salinity	Chlorophyll	Turbidity	Oxygen	Depth	Month	Type
Temperature	–	**0.0003**	0.1032	**0.0076**	**<2.20e^{-16}**	0.5362	**2.358e^{-07}**	0.7144
Salinity	−0.4000	–	0.1297	0.2505	0.0522	0.4493	**0.0008**	0.4862
Chlorophyll	−0.1860	−0.1731	–	**2.20e^{-16}**	**0.0128**	0.3639	0.0726	0.9355
Turbidity	−0.2999	−0.1317	0.8223	–	**0.019**0	0.3149	0.0770	0.7615
Oxygen	−0.7985	0.2207	0.2807	0.2651	–	0.0839	**5.802e^{-09}**	0.2251
Depth	0.0711	−0.0869	−0.1042	0.1153	−0.1969	–	0.4410	**0.0104**
Month	0.5460	−0.3710	−0.2040	−0.2010	−0.6010	0.0885	–	0.4225
Type	−0.0420	0.0800	−0.0093	−0.0350	0.1390	−0.2890	−0.0920	–

Values in the lower part of the table are correlation coefficients, and values in the upper part are p values with bold font being significant at $p < 0.05$

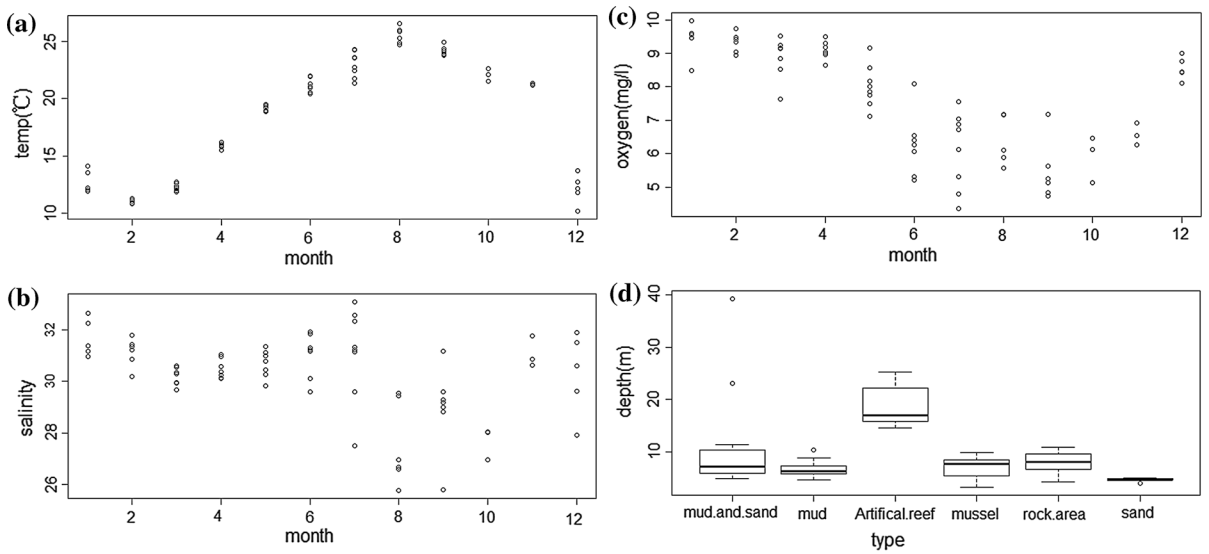

Fig. 2 The distribution of environmental variables in time and space. **a** Temperature changed in time; **b** salinity changed in time; **c** oxygen changed in time; **d** depth changed in space

the 100 simulation runs was greater than 0.5 for both the regular and PCA-based GAMs (Fig. 5). The predicted richness index values obtained from the regular GAM had an average correlation coefficient value of 0.53 over the 100 simulation runs (−0.247 to 0.932). The mean correlation coefficient for the richness index values for the PCA-based GAM was 0.60 (−0.067 to 0.905), which was significantly higher than the average value for the regular GAM (Student's t test, $p < 0.001$). Predicted and observed diversity index values had average correlation coefficients of 0.53 (−0.201 to 0.901) and 0.65 (0.01 to 0.936), respectively, for the regular GAM and PCA-based GAM. Thus, for both the richness and diversity indices, the average correlation coefficients of predicted and observed values for the PCA-based GAM were greater than those for the regular GAM (Student's t test, $p < 0.001$).

Different results were observed between the regular and PCA-based GAMs for the same set of data. The intercepts of the regression models of predicted and observed test data in the cross-validation were positive in the most simulation runs, implying the existence of system biases between observed and predicted indices (Table 3). The regular GAM tended to have large values of intercepts than the PCA-based GAM

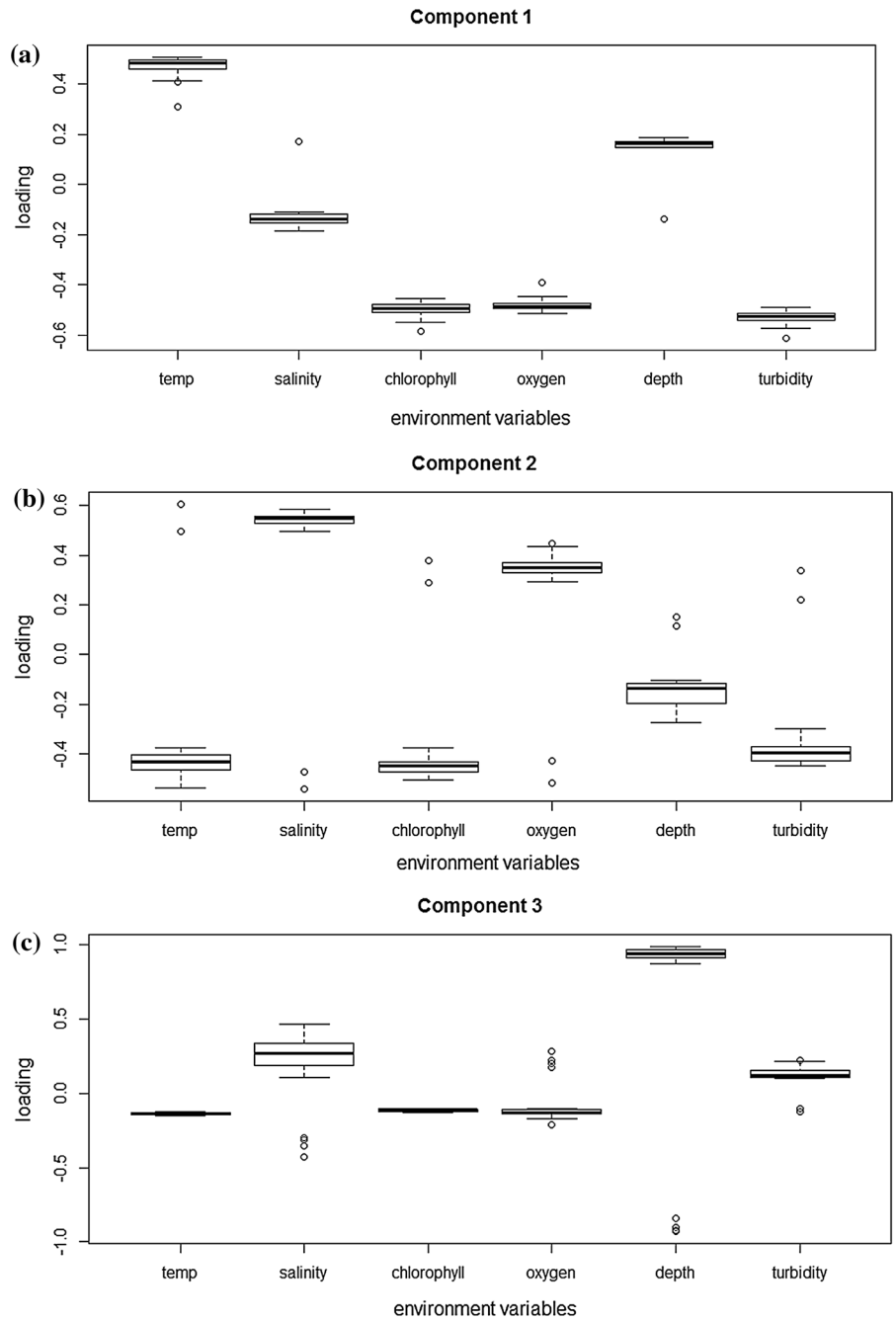

Fig. 3 Box plot of loadings of environmental variables in each principal component over the 100 simulation runs. **a** Loadings of environmental variables in first principal component; **b** loadings of environmental variables in second principal component; **c** loadings of environmental variables in third principal component

(Table 3). The slope coefficient was closer to 1 for the PCA-based GAM, implying that the PCA-based GAM predicted richness and diversity indices more consistently than the regular GAM. We concluded that the PCA-based GAM performed better than the regular GAM in the cross-validation.

Comparison of the two models in habitat predictions

The predicted distributions of species richness and diversity indices tended to differ between the regular GAM and PCA-based GAM (Figs. 6, 7) for all the

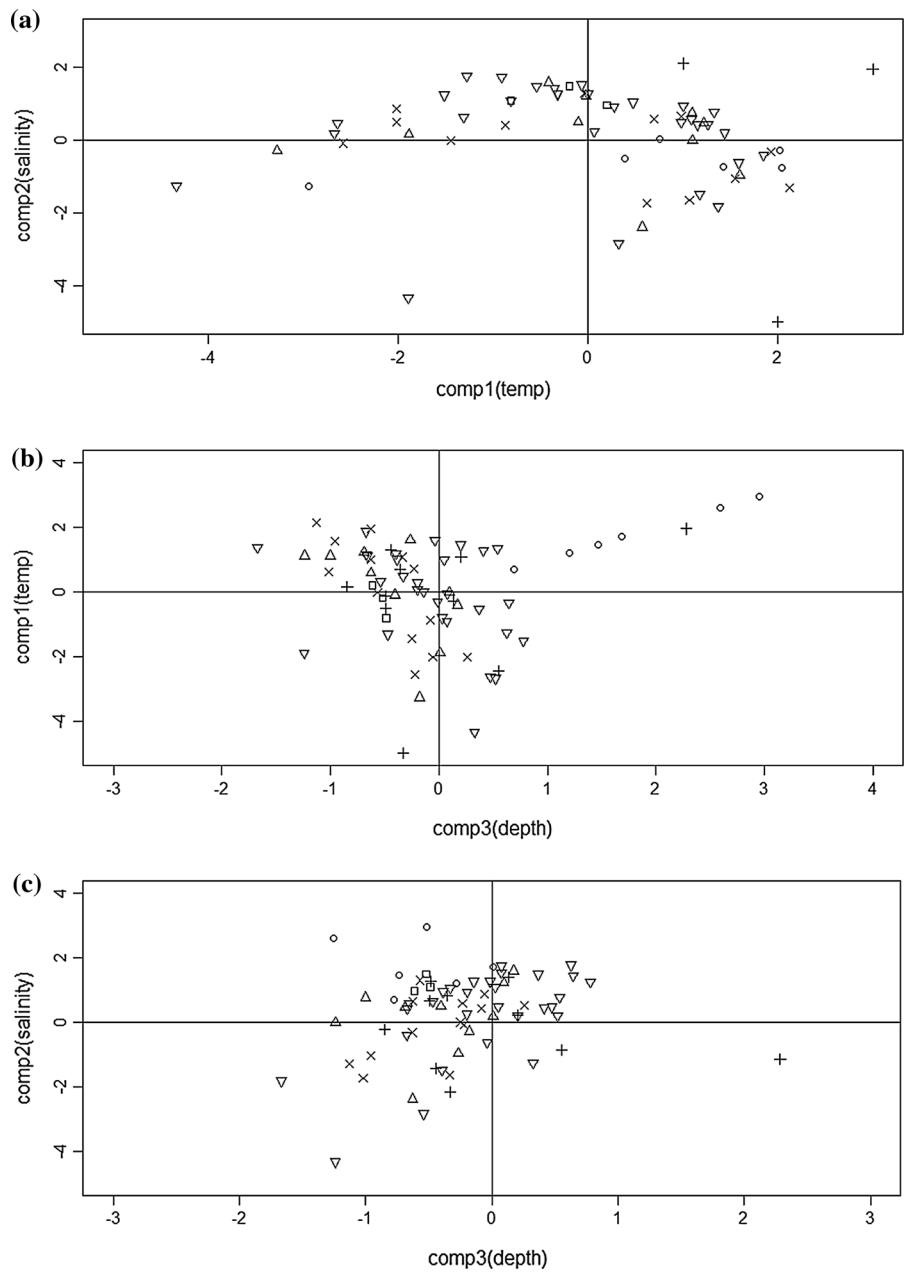

Fig. 4 The distribution of PCs in space, where *circle* artificial reef, *triangle* mussel, *plus* mud and sand, *multiplication* mud, *square* sand, *inverted triangle* Rocky reef, Comp1(temp) means the temperature has the highest loading in the first principal component (PC), Comp2(salinity) means the salinity has the highest loading in the PC, Comp3(depth) means the depth has the highest loading in the third PC. **a** The distribution of first PC in space; **b** the distribution of second PC in space; **c** the distribution of third PC in space

months, except for the richness index in October (Fig. 6a, b). The PCA-based GAM identified the high richness index in August located in the northwestern study area and around the island in the southeastern area; however, the high richness index was mainly concentrated in the center of study area according to the regular GAM (Fig. 6d). The difference in distribution also existed in the diversity index prediction for August and October (Fig. 7). Moreover, the ranges of richness index and diversity index predicted by the PCA-based GAM tended to be wider than the values predicted by the regular GAM. For example, the richness index ranged from 1.9 to 2.9 in August for the regular GAM (Fig. 6d), but from 1.1 to 3.3 for the PCA-based GAM (Fig. 6c). The same trend could also been found for the diversity index.

Table 2 Proportion of the simulation runs in which the factor identified as significant factors

Factors	Richness regular GAM		Diversity regular GAM		Factors	Richness PCA-based GAM		Diversity PCA-based GAM	
	Total (%)	Correlation >0.5 (%)	Total (%)	Correlation >0.5 (%)		Total (%)	Correlation >0.5 (%)	Total (%)	Correlation >0.5 (%)
Month 1	31	30.2	10	9.2	Month 1	6	6.6	2	2.4
Month 2	**63**	**57.1**	**99**	**98.5**	Month 2	**68**	**72.4**	**95**	**95.2**
Month 3	12	7.9	3	1.5	Month 3	4	2.6	0	0
Month 4	34	38.1	37	32.3	Month 4	5	3.9	**65**	**63.9**
Month 5	27	28.6	4	3.1	Month 5	3	1.3	1	1.2
Month 6	16	9.5	5	4.6	Month 6	7	3.9	2	2.4
Month 7	20	12.7	1	0	Month 7	7	3.9	3	2.4
Month 8	28	19.0	12	12.3	Month 8	**56**	**57.9**	0	0
Month 9	32	22.2	48	49.2	Month 9	**70**	**71.1**	36	37.3
Month 10	23	19.0	21	27.7	Month 10	31	25	43	42.2
Month 11	29	30.0	34	40	Month 11	37	34.2	**56**	**54.2**
Month 12	14	6.3	9	3.1	Month 12	0	0	1	1.2
Artificial reef	36	31.7	13	16.9	Artificial reef	17	18.4	12	13.30
Mussel	**69**	**61.9**	49	49.2	Mussel	**73**	**76.3**	**80**	**80.7**
Cage	**60**	**54**	31	36.9	Cage	**88**	**90.8**	**82**	**83.1**
Mud	**67**	**57.1**	32	30.8	Mud	29	25	25	22.9
Sand	39	36.5	11	13.8	Sand	18	21.0	0	0
Mud and sand	**58**	**52.3**	15	26.9	Mud and sand	47	47.4	1	0
Rock	**79**	**76.2**	49	49.2	Rock	**76**	**80.3**	**67**	**65.1**
Temp	7	6.3	1	0	Component 1	29	19.7	31	27.7
Lon	44	38.1	47	44.6	Component 2	7	6.6	16	18.7
Lat	18	11.1	19	13.8	Component 3	2	0	5	6.0
Salinity	43	36.5	11	6.2	Lon	17	1.3	17	14.5
Chlorophyll	**54**	**52.3**	47	41.5	Lat	3	19.7	15	10.8
Oxygen	5	1.6	7	7.7					
Depth	31	30.2	39	44.6					
Turbidity	18	11.2	21	23.1					

Proportion of the simulation runs in which the factor identified as significant factors was in bold print. The "Correlation >0.5" is the proportion of variables as significant factors when the correlation between the predicted and observed indices is higher than 0.5

Discussion

Performance of the two GAM models was compared in their quantifications of relationships between fish community indices and environmental variables. This study suggests that PCA-based GAM can reduce the uncertainty introduced by the existence of strong correlations of environmental variables in a GAM and improve the performance of the model in predicting spatial distribution of the fish community index. The PCA-based GAM showed the improved predictive power of the habitat models compared to the regular GAM in this study. Although the deviance explained by the regular GAM was slightly higher than that by the PCA-based GAM, the predicted precision was higher and biases were lower for the PCA-based GAM in the cross-validation. We considered the results derived from the cross-validation were more subjective in reflecting the predictive power of the model because the test data were not used in the model development.

The regular and PCA-based GAMs yielded different spatial distributions for the fish richness and diversity indices. Because of the space limitation, only

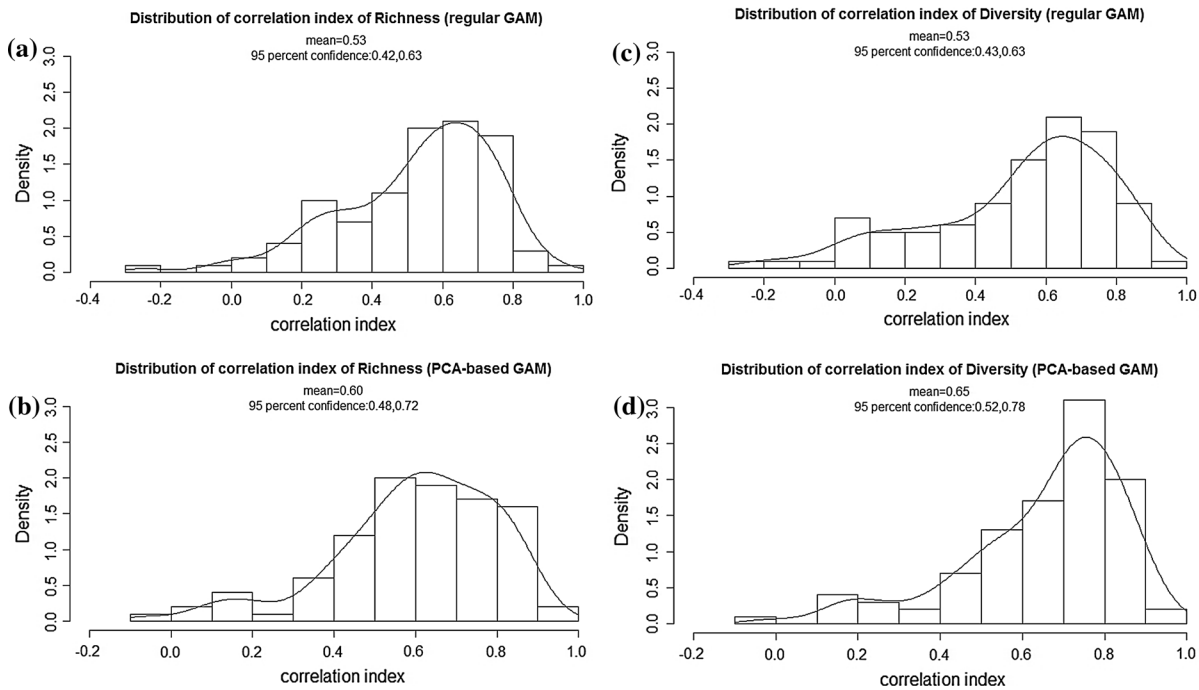

Fig. 5 Probability distribution of correlation coefficients between observed and predicted richness and diversity indices for the test data in the cross-validation over the 100 simulation runs. **a** The results of regular GAM for the richness index; **b** the result of GAM based on principal component analysis (PCA-based GAM) for the richness index; **c** the result of regular GAM for the diversity index; and **d** the result of PCA-based GAM for the diversity index

Table 3 The coefficients for the regression models of observed and predicted indices for the test data in the cross-validation

	Richness index		Diversity index	
	Regular GAM	PCA-based GAM	Regular GAM	PCA-based GAM
Correlation coefficient				
Estimates	0.53	0.82	0.48	0.89
p value	0.07	**0.002**	0.1163	**0.0003**
Intercept (a)				
Estimates	1.28	0.80	0.5885	0.26
p value	0.05	0.19	0.420	0.63
Slope (b)				
Estimates	0.39	0.59	0.7248	0.79
p value (H_0:$b = 0$)	0.07	**0.02**	0.116	**0.03**

Significant test at the $p < 0.05$ level is identified in bold print

the distributions for maximum and minimum fish richness and diversity indices (i.e., August and October, respectively) were shown (Figs. 6, 7). Although the overall distributional patterns were similar between the two GAMs in the richness indices predicted, large differences occurred on fine scales. For example, the area with low richness indices was smaller in August for the regular GAM than for the PCA-based GAM (Fig. 6). Differences in the prediction of fish richness and diversity indices may result in the development and adaptation of different management regulations (Link et al. 2011; Zarkami et al. 2012; Maloney et al. 2013). For example, a low predicted fish diversity may lead to the exclusion of some areas from being included in marine protection areas because marine protection area tends to include areas with high biodiversity (Edgar et al. 2008). Based on the PCA-based GAM, areas with high fish diversity appeared to be around rocky areas in August (Fig. 6). This is consistent with previous studies which showed that rocky areas tended to have higher fish diversity because the kelp in this area bloomed in August which made rocky areas attractive for many fish species (Wang et al. 2011, 2012). This was not obvious for the results derived from the regular GAM. Thus, the use of

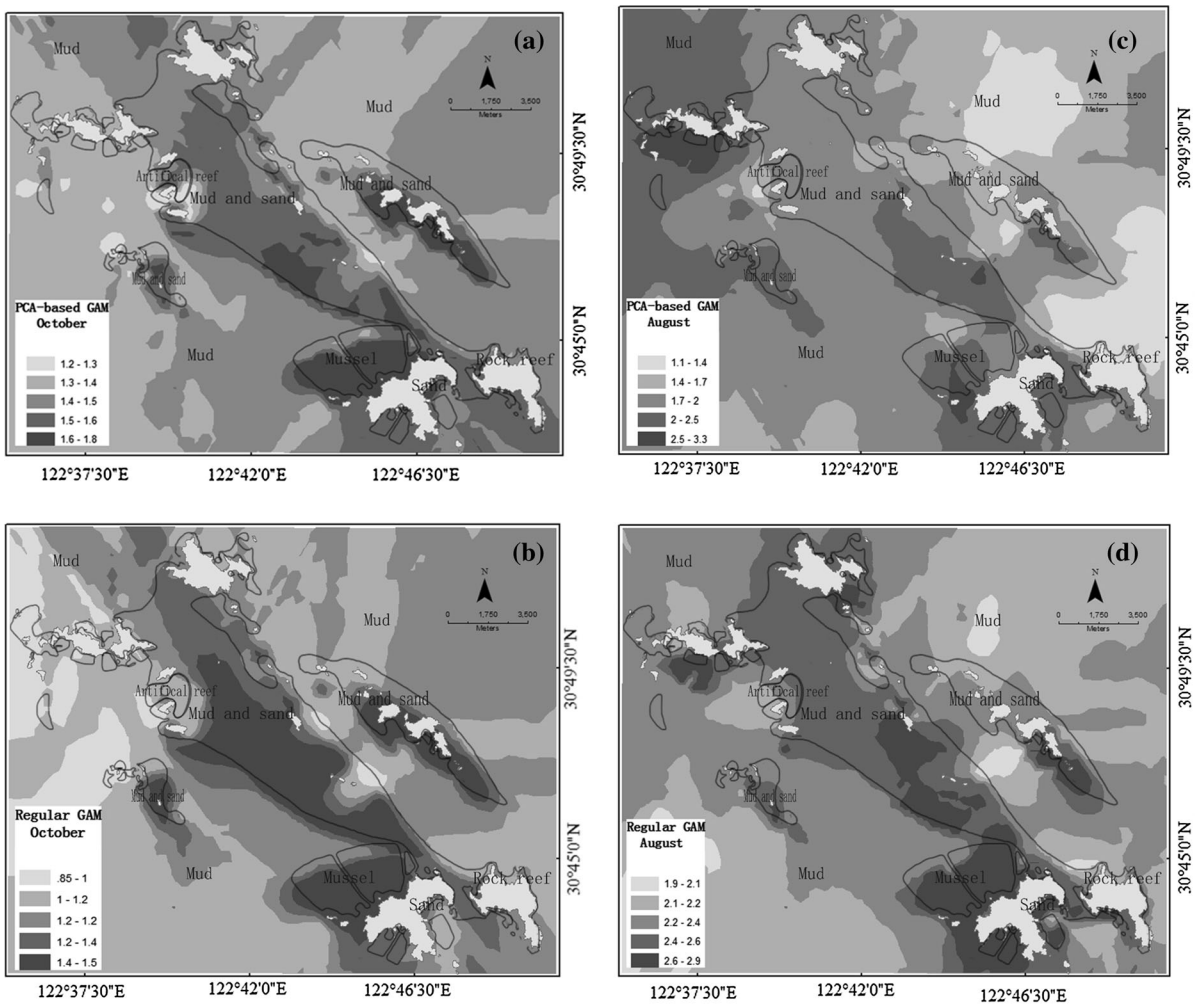

Fig. 6 The distribution of richness index of demersal fish community from different models. **a** Richness distribution in October (PCA-based GAM); **b** richness distribution in October (regular GAM); **c** richness distribution in August (PCA-based GAM); and **d** richness distribution in August (regular GAM)

the regular and PCA-based GAMs may have different impacts on resource conservation and fisheries management.

Although the PCA-based GAM tended to yield better fish richness and diversity indices in this study, it is often difficult to choose an appropriate statistical model because many methods are often available for addressing one question. As Graham (2003) studied, methods like PCA only help reduce biases and make analyses more repeatable, but the explanatory variables are still correlated by nature. Therefore, an optimal model should be chosen not only based on its statistical properties but also based on its appropriateness of ecological implications. The variables that appeared significant in a high proportion of simulation runs were chosen as the main significant variables in this study. This study suggests that month and habitat type were the two most important variables for both fish richness index and diversity index (Table 2). However, the two GAMs considered their importance differently. The influence of month was strengthened, but the influence of environmental variables was reduced in the PCA-based GAM compared to the regular GAM. The previous studies in the same sites showed that fish community had an obvious seasonal change especially in rocky areas (Wang et al. 2011, 2012). Comparing the results derived from this study and previous studies on the influence of month and

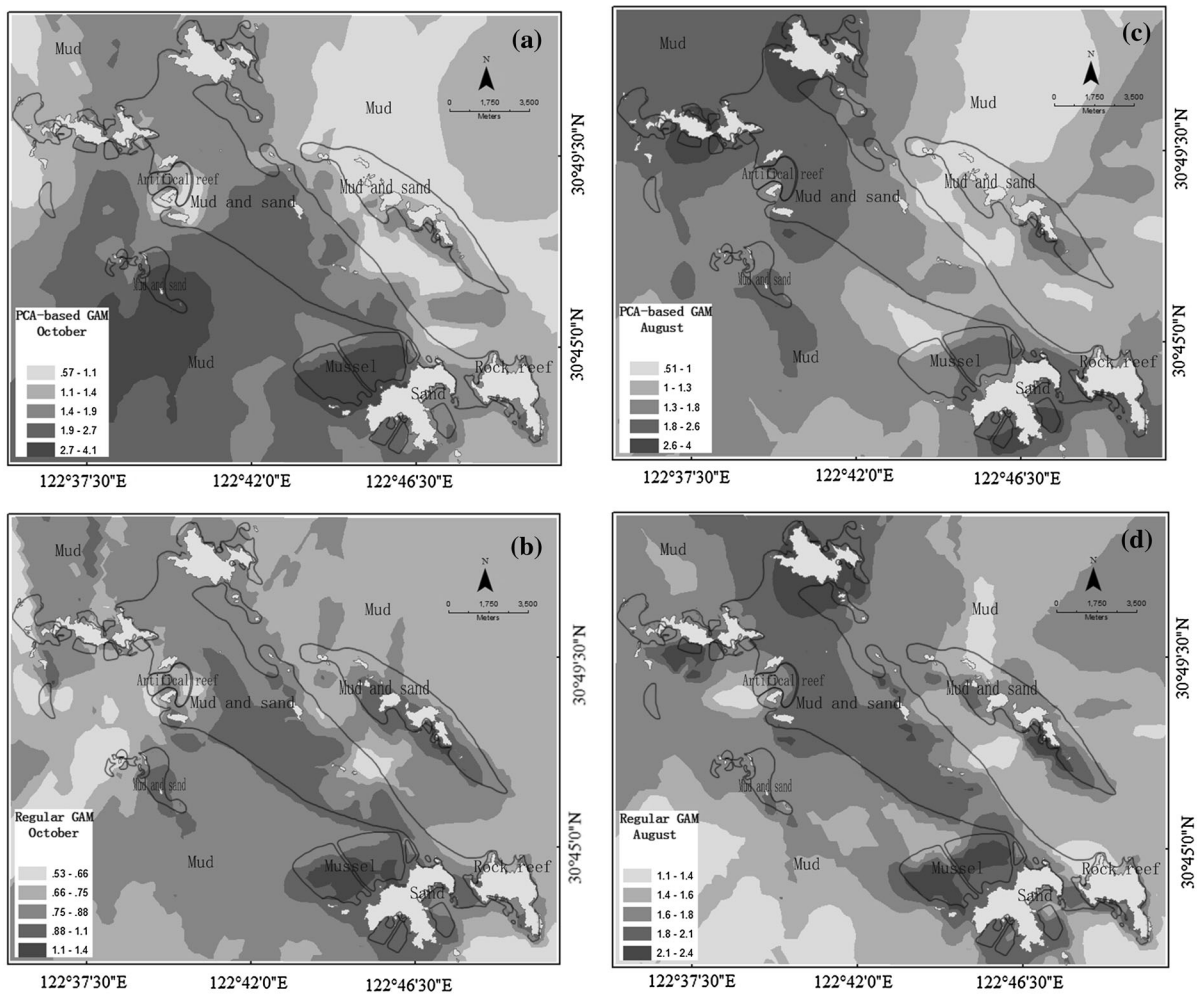

Fig. 7 The distribution of diversity index of demersal fish community from different models. **a** Diversity distribution in October (PCA-based GAM); **b** diversity distribution in October (regular GAM); **c** diversity distribution in August (PCA-based GAM); and **d** diversity distribution in August (regular GAM)

habitat, PCA-based GAM tended to yield results more consistently compared to the regular GAM in exploring the relationship between environment and fish richness and diversity indices.

Different sediments may result in different fish community assemblages (Jenkins and Wheatley 1998; Lopez-Lopez et al. 2011). This study showed sediments could have great impacts on fish community. A large spatial heterogeneity exists in the distribution of natural sediments such as rock reef, sand, mud, and their mixtures in the study area (Wang et al. 2012). Artificial sediments associated with aquaculture activities, such as mussel aquaculture, artificial reef, and net cage aquaculture, also exist in this study area (Wang et al. 2012). Fish community structures differ greatly among these habitats (Saraceno et al. 2005; Ribeiro et al. 2012; Wang et al. 2012). A good understanding of the roles of sediments is important for resources conservation. For example, rock areas in this study support high species diversity because of the existing of kelp bed. Thus, maintaining kelp bed is critical to preserve fish resources (Terawaki et al. 2003; Wang et al. 2011).

Other environmental variables in addition to sediment also play a major role in regulating fish community distribution (Lopez-Lopez et al. 2011). Depth, salinity, and temperature were three environmental variables commonly considered in habitat

modeling (Jacob et al. 1998; Marshall and Elliott 1998; Lefkaditou et al. 2008). For example, Bulger et al. (1993) found salinity was an important variable influences fish movement and distribution. Jaureguizar et al. (2004) considered that salinity was the main factor influencing fish spatial distribution. Turbidity affects fish distribution through food supplies and providence of refuges (Johnston et al. 2007). Other factors such as oxygen or chlorophyll also influence fish distribution directly or indirectly (Maes et al. 2007; Agboola et al. 2013). Thus, we tried to include as many potential explanatory variables as possible to build up our habitat models. Understanding the effect of essential fish habitat provides a legal basis for the creation of marine-protected areas (Rieser 2000). However, as shown in our study, different approaches might yield different results in identifying critical environmental factors. Therefore, it is important to compare different methods and interpret the results cautiously.

Several environmental variables were considered in this study. However, other variables such as fishing and hydrodynamic forces, which also can influence the distribution of demersal fish community, were not included in this study (Steele 1996; MacKenzie and KiØrboe 2000; Marchetti and Moyle 2001; Burrows 2012). If additional variables could be included, the model might perform better.

Many ecological studies tend to have relatively small sample sizes because of financial/time restrictions (Altekruse et al. 2003; Varela et al. 2011). Thus, each sample can potentially affect the results greatly. To consider such an issue, we used a cross-validation to test the performance of the model, for which 80 % of the data were randomly selected for model building, while the remaining 20 % data were used for model testing. This random sampling process was repeated over the 100 simulation runs. For each simulation run, because of random sampling, the data sampled for model building are likely to have a narrower range compared to the full data set. Thus, the models developed in some simulation runs may be used beyond their ranges in the predictions. The interpolation may introduce large errors in the predication (Agostini et al. 2008). For instance, when we look into the subsample used for building model in some simulation runs, the maximum depth was reduced from 35 to 26 m. Thus, the fish community index in the area where depth is deeper than 26 m had to be predicted beyond the depth ranges. Thus, given the sampling restriction of an ecological study, the cross-validation approach used in this study can more realistically mimic what we have to experience in habitat modeling, and the results are thus more realistic compared to the regular summary statistics for model goodness of fit.

The variables identified as significant for a given model varied among simulation runs because data were subsampled for model building. This may result in some potential problems in model fitting. For example, it is difficult to understand the response curve between fish richness and diversity indices and environmental variables. Response curves in most simulation runs between fish community and chlorophyll were similar with two obvious peaks at 20 and 80 mg/L, but response curves in a few simulation runs only had one obvious peak at 20 mg/L. This may indicate that the information of the second peak was excluded in modeling as a result of random sampling, suggesting that the results might be influenced by subsampling.

The *Margalefs index* and *Shannon index* were often used to quantify spatial assemblages of fish species. Although these two indices are not able to distinguish changes in specific species, compositions of fish communities and two fish communities of different species compositions can have similar values of these two indices, and they can describe an overall change in species richness and diversity, which serves the purpose of this study which is the comparison of the two different approaches to modeling the relationships between fish community and environmental variables.

The relationships between fish community and environment variables tend to be complex, nonlinear, and not easy to understand. Usually, interaction terms of environmental variables need to be considered in habitat modeling (Saraceno et al. 2005; Ribeiro et al. 2012). For simplicity, no interaction terms were considered in this study. The potential impacts of interaction terms on PCA-based GAM need to be evaluated. However, this should not affect the comparisons between the regular GAM and PCA-based GAM because both the models were subject to the same data and variables with no interaction terms considered.

Conclusion

In summary, this study shows that if environmental variables are highly correlated, the PCA-based GAM

is a more appropriate approach for statistical habitat modeling. Given the high likelihood of the existence of these data in an ecological study, we recommend that PCA-based GAM be considered in habitat modeling. However, the attention should be paid to the interpretation of results.

Acknowledgments Financial support for this study was provided by the National Basic Research Program of China (No. 2011CB111608), National Natural Science Foundation of China (No. 41176110), Shanghai Ocean University College of Marine Sciences and International Center for Marine Sciences. The data analysis done at the University of Maine is partially supported by the Maine Sea Grant College Program. We would like to thank K. Wang, Q. M. Chen, Q. Xu, X. Zhao, and L. R. Chen for their assistance in the field. We graciously acknowledge J. Lin and J. Ding for their support of dealing with environment data and part of figures.

References

Afifi AA, Clark V (1996) Computer-aided multivariate analysis. Chapman and Hall/CRC, New York

Agboola JI, Uchimiya M, Kudo I, Osawa M, Kido K (2013) Seasonality and environmental drivers of biological productivity on the western Hokkaido coast, Ishikari Bay, Japan. Estuar Coast Shelf Sci 127:12–13

Agostini VN, Hendrix AN, Hollowed AB, Wilson CD, Pierce SD, Francis RC (2008) Climate-ocean variability and Pacific hake: a geostatistical modeling approach. J Mar Syst 71:237–248

Ahmadi-Nedushan B, St-Hilaire A, Bérubé M, Robichaud É, Thiémonge N, Bobée B (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. River Res Appl 22:503–523

Akin S, Buhan E, Winemiller KO, Yilmaz H (2005) Fish assemblage structure of Koycegiz Lagoon-Estuary, Turkey: spatial and temporal distribution patterns in relation to environmental variation. Estuar Coast Shelf Sci 64:671–684

Altekruse SF, Elvinger F, Wang Y, Ye K (2003) A model to estimate the optimal sample size for microbiological surveys. Appl Environ Microbiol 69:6174–6178

Annoni P, Saccardo I, Gentili G, Guzzi L (1997) A multivariate model to relate hydrological, chemical and biological parameters to salmonoid biomass in Italian Alpine rivers. Fish Manag Ecol 4:439–452

Araújo FG, Azevedo MCCD, Silva MDA, Pessanha ALM, Gomes ID, Cruz-Filho AGD (2002) Environmental influences on the demersal fish assemblages in the Sepetiba Bay, Brazil. Estuaries 25:441–450

Barquín J, Death RG (2009) Physical and chemical differences in karst springs of Cantabria, northern Spain: do invertebrate communities correspond? Aquat Ecol 43(2):445–455

Bierman P, Lewis M, Ostendorf B, Tanner J (2011) A review of methods for analysing spatial and temporal patterns in coastal water quality. Ecol Indic 11:103–114

Blaber SJM, Blaber TG (1980) Factors affecting the distribution of juvenile estuarine and inshore fish. J Fish Biol 17:143–162

Buisson L, Blanc L, Grenouillet G (2008) Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors. Ecol Freshw Fish 17:244–257

Bulger AJ, Hayden BP, Monaco ME, Nelson DM, McCormick-Ray MG (1993) Biologically-based estuarine salinity zones derived from a multivariate analysis. Estuaries 16:311–322

Burrows MT (2012) Influences of wave fetch, tidal flow and ocean colour on subtidal rocky communities. Mar Ecol Prog Ser 445:193–207

Chang JH, Chen Y, Holland D, Grabowski J (2010) Estimating spatial distribution of American lobster *Homarus americanus* using habitat variables. Mar Ecol Prog Ser 420:145–156

Dubey VK, Sarkar UK, Pandey A, Sani R, Lakra WS (2012) The influence of habitat on the spatial variation in fish assemblage composition in an unimpacted tropical river of Ganga basin, India. Aquat Ecol 46:165–174

Edgar GJ, Langhammer PF, Allen G, Brooks TM, Brodie J, Crosse W, Silva ND, Fishpool LDC, Foster MN, Knox DH, Mccosker JE, Mcmanus R, Millar AJK, Mugo R (2008) Key biodiversity areas as globally significant target sites for the conservation of marine biological diversity. Aquat Conserv 18:969–983

Ellis RN, Kroonenberg PM, Harch BD, Basford KE (2006) Non-linear principal components analysis: an alternative method for finding patterns in environmental data. Environmentrics 17:1–11

Emery WJ, Thomson RE (2001) Data analysis methods in physical oceanography. Elsevier Science, Amsterdam, p 654

Fischer JR, Krogman RM, Quist MC (2013) Influences of native and non-native benthivorous fishes on aquatic ecosystem degradation. Hydrobiologia 711:187–199

Fortes WLS, Almeida-Silva PH, Prestrelo L, Monterio-Neto C (2014) Patterns of fish and crustacean community structure in a coastal lagoon system, Rio de Janeiro, Brazil. Mar Biol Res 10:111–122

Francis MP, Morrison MA, Leathwick J, Walsh C, Middleton C (2005) Predictive models of small fish presence and abundance in northern New Zealand harbours. Estuar Coast Shelf Sci 64:419–435

Friedlander AM (2001) Essential fish habitat and the effective design of marine reserves: application for marine ornamental fishes. Aquar Sci Conserv 3:135–150

Graham MH (2003) Confronting multicollinearity in ecological multiple regression. Ecology 84(11):2809–2815

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, London, p 335

Hoeinghaus DJ, Winemiller KO, Birnbaum JS (2007) Local and regional determinants of stream fish assemblage structure: inferences based on taxonomic vs. functional groups. J Biogeogr 34:324–338

Jacob W, McClatchie S, Probert PK, Hurst RJ (1998) Demersal fish assemblages off southern New Zealand in relation to depth and temperature. Deep Sea Res 45:2119–2155

Jaureguizar AJ, Menni R, Guerrero R, Lasta C (2004) Environmental factors structuring fish communities of the Río de la Plata estuary. Fish Res 66:195–211

Jenkins GP, Wheatley MJ (1998) The influence of habitat structure on nearshore fish assemblages in a southern Australian embayment: comparison of shallow seagrass, reef-algal and unvegetated sand habitats, with emphasis on their importance to recruitment. J Exp Mar Biol Ecol 221:147–172

Johnston R, Sheaves M, Molony B (2007) Are distributions of fishes in tropical estuaries influenced by turbidity over small spatial scales? J Fish Biol 71:657–671

Jordaan A, Chen Y, Townsend DW, Sherman S (2010) Identification of ecological structure and species relationships along an oceanographic gradient in the Gulf of Maine using multivariate analysis with bootstrapping. Can J Fish Aquat Sci 67:701–719

Jyväsjärvi J, Boros G, Jones RI, Hämäläinen H (2013) The importance of sedimenting organic matter, relative to oxygen and temperature, in structuring lake profundal macroinvertebrate assemblages. Hydrobiologia 709:55–72

Kleyer M, Dray S, Bello FD, Lepš J, Pakeman RJ, Strauss B, Thuiller W, Lavorel S (2012) Assessing species and community functional responses to environmental gradients: which multivariate methods? J Veg Sci 23:805–821

Kroll CN, Song P (2013) Impact of multicollinearity on small sample hydrologic regression models. Water Resour Res 49:3756–3769

Leathwick JR, Elith J, Hastie T (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecol Model 199:188–196

Lefkaditou E, Politou CY, Palialexis A, Dokos J, Cosmopoulos P, Valavanis VD (2008) Influences of environmental variability on the population structure and distribution patterns of the short-fin squid *Illex coindetii* (Cephalopoda: Ommastrephidae) in the Eastern Ionian Sea. Hydrobiologia 612:71–90

Link JS, Nye JA, Hare JA (2011) Guidelines for incorporating fish distribution shifts into a fisheries management context. Fish Fish 12:461–469

Liu C, White M, Newell G, Griffioen P (2013) Species distribution modeling for conservation planning in Victoria, Australia. Ecol Model 249:68–74

Lopez-Lopez L, Preciado I, Velasco F, Olaso I, Gutiérrez-Zabala JL (2011) Resource partitioning amongst five coexisting species of gurnards (Scorpaeniforme: Triglidae): role of trophic and habitat segregation. J Sea Res 66:58–68

Love JW, May EB (2007) Relationships between fish assemblage structure and selected environmental factors in Maryland's Coastal Bays. Northeast Nat 14:251–268

Macedo-Soares LCP, Freire AS, Muelbert JH (2012) Small-scale spatial and temporal variability of larval fish assemblages at an isolated oceanic island. Mar Ecol Prog Ser 444:207–222

MacKenzie BR, KiØrboe T (2000) Larval fish feeding and turbulence: a case for the downside. Limnol Oceanogr 45:1–10

Maes J, Stevens M, Breine J (2007) Modelling the migration opportunities of diadromous fish species along a gradient of dissolved oxygen concentration in a European tidal watershed. Estuar Coast Shelf Sci 75:151–162

Maggini R, Lehmann A, Zimmermann NE, Guisan A (2006) Improving generalized regression analysis for the spatial prediction of forest communities. J Biogeogr 33: 1729–1749

Maloney KO, Weller DE, Michaelson DE, Ciccotto PJ (2013) Species distribution models of freshwater stream fishes in Maryland and their implications for management. Environ Model Assess 18:1–12

Marchetti MP, Moyle PB (2001) Effects of flow regime on fish assemblages in a regulated California stream. Ecol Appl 11:530–539

Margalef DR (1958) Information theory in ecology. Gen Syst 3:36–71

Marshall S, Elliott M (1998) Environmental influences on the fish assemblage of the Humber estuary, U.K. Estuar Coast Shelf Sci 46:175–184

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago

Palamara L, Manderson J, Kohut J, Oliver MJ, Gray S, Goff J (2012) Improving habitat models by incorporating pelagic measurements from coastal ocean observatories. Mar Ecol Prog Ser 447:15–30

Pérez FF, Ríos AF, Castro CG, Fraga F (1998) Mixing analysis of nutrients, oxygen and dissolved inorganic carbon in the upper and middle North Atlantic Ocean east of the Azores. J Mar Syst 16:219–233

Ptacnik R, Lepistö L, Willén E, Brettum P, Andersen T, Rekolainen S, Solheim AL, Carvalho L (2008) Quantitative responses of lake phytoplankton to eutrophication in Northern Europe. Aquat Ecol 42:227–236

Ribeiro J, Carvalho GM, Goncalves JMS, Erzini K (2012) Fish assemblages of shallow intertidal habitats of the Ria Formosa lagoon (South Portugal): influence of habitat and season. Mar Ecol Prog Ser 446:259–273

Rieser A (2000) Essential fish habitat as a basis for marine protected areas in the U.S. exclusive economic zone. Bull Mar Sci 66:889–899

Saraceno M, Provost C, Piola AR (2005) On the relationship between satellite-retrieved surface temperature fronts and chlorophyll *a* in the western South Atlantic. J Geophys Res 110:1–16

Schmiing M, Afonso P, Tempera F, Santos RS (2013) Predictive habitat modeling of reef fishes with contrasting trophic ecologies. Mar Ecol Prog Ser 474:201–216

Shannon EC, Weaver W (1948) The mathematical theory of communication. Urbana University of Illinois Press, Illinois

Steele MA (1996) Effects of predators on reef fishes: separating cage artifacts from effects of predation. J Exp Mar Biol Ecol 198:249–267

Sternberg D, Kennard MJ (2013) Environmental, spatial and phylogenetic determinants of fish life-history traits and functional composition of Australian rivers. Freshw Biol 58:1767–1778

Straka M, Syrovátka V, Helešic J (2012) Temporal and spatial macroinvertebrate variance compared: crucial role of CPOM in a headwater stream. Hydrobiologia 686:119–134

Terawaki T, Yoshikawa K, Yoshida G, Uchimura M, Iseki K (2003) Ecology and restoration techniques for *Sargassum* beds in the Seto inland sea, Japan. Mar Pollut Bull 47:198–201

Toepfer CS, Williams LR, Martinez AD, Fisher WL (1998) Fish and habitat heterogeneity in four streams in the central

Oklahoma/Texas plains ecoregion. Proc Okla Acad Sci 78:41–48

Varela Z, Fernández JA, Aboal JR, Real C, Carballeira A (2011) Determination of the optimal size of area to be sampled by use of the moss biomonitoring technique. J Atmos Chem 65:37–48

Wang L, Zhang SY, Wang ZH, Wang K, Lin J (2011) Constitution of fish assemblages in three nearshore habitats and the effect of benthic macroalgae on fish assemblages in Gouqi Island. J Fish China 35:1037–1049 (in Chinese)

Wang ZH, Zhang SY, Chen QM, Xu M, Wang K (2012) Fish community ecology in rocky reef habitat of Ma'an Archipelago. I. Species composition and diversity. Biodivers Sci 20:41–50 (in Chinese)

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc B 73:3–36

Yee TW (2006) Constrained additive ordination. Ecology 87:203–213

Zarkami R, Sadeghi R, Goethals P (2012) Use of fish distribution modeling for river management. Ecol Model 230:44–49