RESEARCH ARTICLE

Modelling the spatial distribution of three *Portunidae* crabs in Haizhou Bay, China

Jing Luan[1], Chongliang Zhang[1], Binduo Xu[1], Ying Xue[1], Yiping Ren[1,2]*

1 College of fisheries, Ocean University of China, Qingdao, China, **2** Laboratory for Marine Fisheries Science and Food Production Processes, Pilot National Laboratory for Marine Science and Technology (Qingdao), Aoshanwei Town, Jimo, Qingdao, China

* renyip@ouc.edu.cn

Abstract

Crab species are economically and ecologically important in coastal ecosystems, and their spatial distributions are pivotal for conservation and fisheries management. This study was focused on modelling the spatial distributions of three *Portunidae* crabs (*Charybdis bimaculata*, *Charybdis japonica*, and *Portunus trituberculatus*) in Haizhou Bay, China. We applied three analytical approaches (Generalized additive model (GAM), random forest (RF), and artificial neural network (ANN)) to spring and fall bottom trawl survey data (2011, 2013–2016) to develop and compare species distribution models (SDMs). Model predictability was evaluated using cross-validation based on the observed species distribution. Results showed that sea bottom temperature (SBT), sea bottom salinity (SBS), and sediment type were the most important factors affecting crab distributions. The relative importance of candidate variables was not consistent among species, season, or model. In general, we found ANNs to have less stability than both RFs and GAMs. GAMs overall yielded the least complex response curve structure. *C*. *japonica* was more pronounced in southwestern portion of Haizhou Bay, and *C*. *bimaculata* tended to stay in offshore areas. *P*. *trituberculatus* was the least region-specific and exhibited substantial annual variations in abundance. The comparison of multiple SDMs was informative to understand species responses to environmental factors and predict species distributions. This study contributes to better understanding the environmental niches of crabs and demonstrates best practices for the application of SDMs for management and conservation planning.

OPEN ACCESS

Citation: Luan J, Zhang C, Xu B, Xue Y, Ren Y (2018) Modelling the spatial distribution of three *Portunidae* crabs in Haizhou Bay, China. PLoS ONE 13(11): e0207457. https://doi.org/10.1371/journal.pone.0207457

Editor: Suzannah Rutherford, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: March 9, 2018

Accepted: October 31, 2018

Published: November 14, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the China Postdoctoral Science Foundation (Grant No. 2017M622284) to CZ and the Fundamental Research Funds for the Central Universities (Grant No. 201562030) to BX. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Many fish populations have decreased in abundance and shifted distributions due to marine pollution, climate changes and over-exploitation [1–3]. In many marine ecosystems the declines of large predatory species have coincided with increase of small size, short-lived crustacean, including shrimps and crabs [4]. Moreover, the emerging economic values of crustacean species tend to be large and provide ample supports for local, small-scale fisheries [5–6]. For example, an increase of *Portunidae* contributed substantially to crab fisheries in the Yellow Sea over recent decades. Three crabs in the *Portunidae* family: *Charybdis bimaculata*, *Charybdis japonica*, and *Portunus trituberculatus*, are ecologically and economically valuable along the coastal area of China [7, 8]. Among them, *P. trituberculatus* has a larger body size, relative longer life span and are more migratory than *C. bimaculata* and *C. japonica* [9, 10]. Due to the functional natatorial legs of Portunids, "swimming crabs", they have higher mobility than most other benthic crustaceans [11]. Consequent to their enhanced mobility, characterizing Portunid distribution is difficult modelling tasks. Unfortunately, despite their regional importance, there has been few studies to characterize the distribution and phenology of these species.

The spatial distributions of the crabs are influenced by environmental factors. For instance, temperature and salinity may influence overwintering of migratory crabs [12, 13], and in some special life history stage crabs might be more sensitive to salinity [7, 12]. Additionally, ranges of suitable temperature and substrate are strongly associated with their habitat preferences [14–16]. It has also been shown that dissolved oxygen influences the recruitment mortality of crabs, and therefore can serve as an important bottom-up driver of population dynamic [17]. Thereby, spatial distributional modelling studies that consider abiotic mechanisms are necessary to understand the environmental niches of crabs and assist in management and conservation planning.

Regarding spatial prediction, species distribution models (SDMs) are commonly used in ecology and biodiversity studies to predict species' potential distribution [18]. For crabs, SDMs have been applied for estimating distributions of crabs living in different habitats (e.g. estuarine, intertidal zone and mangrove area), monitoring crab invasions [19, 20], forecasting fishing grounds [21, 22], reflecting modifications of typical habitats, identifying habitat suitability [12], and standardization of catch-per-unit-effort (CPUE) [23]. A wide range of statistical techniques are commonly used for SDMs, ranging from regression-based methods (linear regression, generalized additive models, and multivariate discriminant analysis [24, 25]), to non-parametric methods [26], such as machine learning (ML). It is established that although regression-based methods are straightforward to interpret, they are limited in their abilities to handle complicated relationships [24]. In many applications, ML can identify complex relationships flexibly and outperform regression-based methods in predictive capability [24]. In particular, previous ML studies on development of SDMs, using boosted regression trees (BRT), random forest (RF), maximum entropy (MaxEnt) and genetic algorithm, showed promise for predicting the native ranges of crabs [19, 20, 22] and many other marine species [22, 25, 27]. However, it should be noted that the models' predictive performances depend not only on their algorithms [24], but also on study goals, spatial scales, sample sizes [28], distribution patterns [24, 29, 30], species characteristics [31], and the form of species responses to environmental changes [32]. Assessing the reliability of those models for various species is of great concern for their application. Therefore, a comparison of different MLs to traditional regression methods for multiple species is need to validate the practical application of SDMs.

In this study, we conducted a bottom-trawl crab survey and collected relevant environmental variables including temperature, salinity, depth, and sediment type. SDMs of three *Portunidae* crabs were developed using three modelling methods, including one traditional generalized additive model (GAM), and two ML approaches, random forest (RF) and artificial neural network (ANN) [33, 34, 35]. In order to understand these crabs' spatial distributions and to assess the reliability of their distribution models, we identified the effects of environmental factors on three crabs and compared the performances among these models regarding fitting capability, predictability, and model stability. We paid special consideration to the effects of modelling method, species, and seasonality on model predictability. Finally, the distribution maps of three crabs were predicted using the developed models to support current regional fisheries management.

Materials and methods

Data collection

The biomass data of the three crab species were collected in Haizhou Bay, China (34˚25′ −35˚35′N, 119˚25′−121˚5′E), an open bay on the south-western Yellow Sea. (No specific permissions were required for the surveys, as the survey area was located in a typical fishing ground, in which there were no national parks or other protected area of wildlife. The surveys did not involve endangered or protected species). Haizhou Bay was a historically important fishing ground and served as a spawning and feeding habitat for many species in the 1980s [36]. Nonetheless, the ecosystem structure changed over past decades as a result of climate changes and increasing fishing pressure [37]. We conducted bottom trawl surveys in spring and fall of 2011, 2013, 2014, 2015 and 2016. A stratified random sampling design was used, in which the survey area was divided into five strata based on water depth (from 3.77m to 39.86m) and latitude (Fig 1). A total of 24 sampling sites in 2011 and 18 sites in the following years were chosen in each survey, covering the whole area of Haizhou Bay. We used otter trawl vessels of 162 kW and trawl nets with the cod-end mesh size of 17mm and width of 25m. The trawl was hauled for about 1h at the speed of 2–3 knots in each site, standardized to 1h haul with 2 knot vessel speed (i.e. CPUE of kn*h). The logarithmic transformed relative catch was used as the response variable to reduce data heterogeneity and to avoid the undue effect of outliers [38–40].

The predictive variables had two categories, environmental variables including sea surface salinity (SSS), sea surface temperature (SST), sea bottom salinity (SBS), sea bottom temperature (SBT), water depth and sediment type, and the spatio-temporal variables including geographical positions (i.e. longitude and latitude) and survey years. Temperature, salinity and depth were recorded using a CTD system (XR-420) in each sample site. The sediment types included sand, sandy silt, sand-silt-clay according to Shepard's nomenclature of sediments [41] and were treated as factors in the analysis. Data were provided by the College of Environmental Science and Engineering, Ocean University of China (unpublished data). As there were substantial variations in species abundance among years due to the population dynamics rather than distribution patterns, survey years were included in the SDMs as a factor to adjust the fluctuation of relative abundance. Considering the differences in habitat and distributional pattern among seasons, and particularly the migration of *P*. *trituberculatus*, the spring and autumn data were treated separately in models and subsequent analyses (Table 1). That is, we assumed a relatively consistent distributional patterns within the same seasons and built the models for each season individually. The model for *P*. *trituberculatus* was omitted for spring due to its low prevalence as a result of seasonal migration.

We used variation inflation factor (VIF) to examine the collinearity between predictive variables before model construction [42]. The VIF value of variable that was higher than 3 implied substantial correlation with other variables [42], thus were omitted.

Statistical methods

GAM, RF and ANN were used to develop a set of species distribution models. Among these statistical methods, GAM was one of the most widely used methods in SDMs, whereas RF and ANN had many strengths over the traditional regression-based methods [24, 29, 43], such as efficient recognition of data patterns, independence of particular functional relationships, free-assumption of data properties, and the ability to accommodate interactions among variables without a priori specification [27].

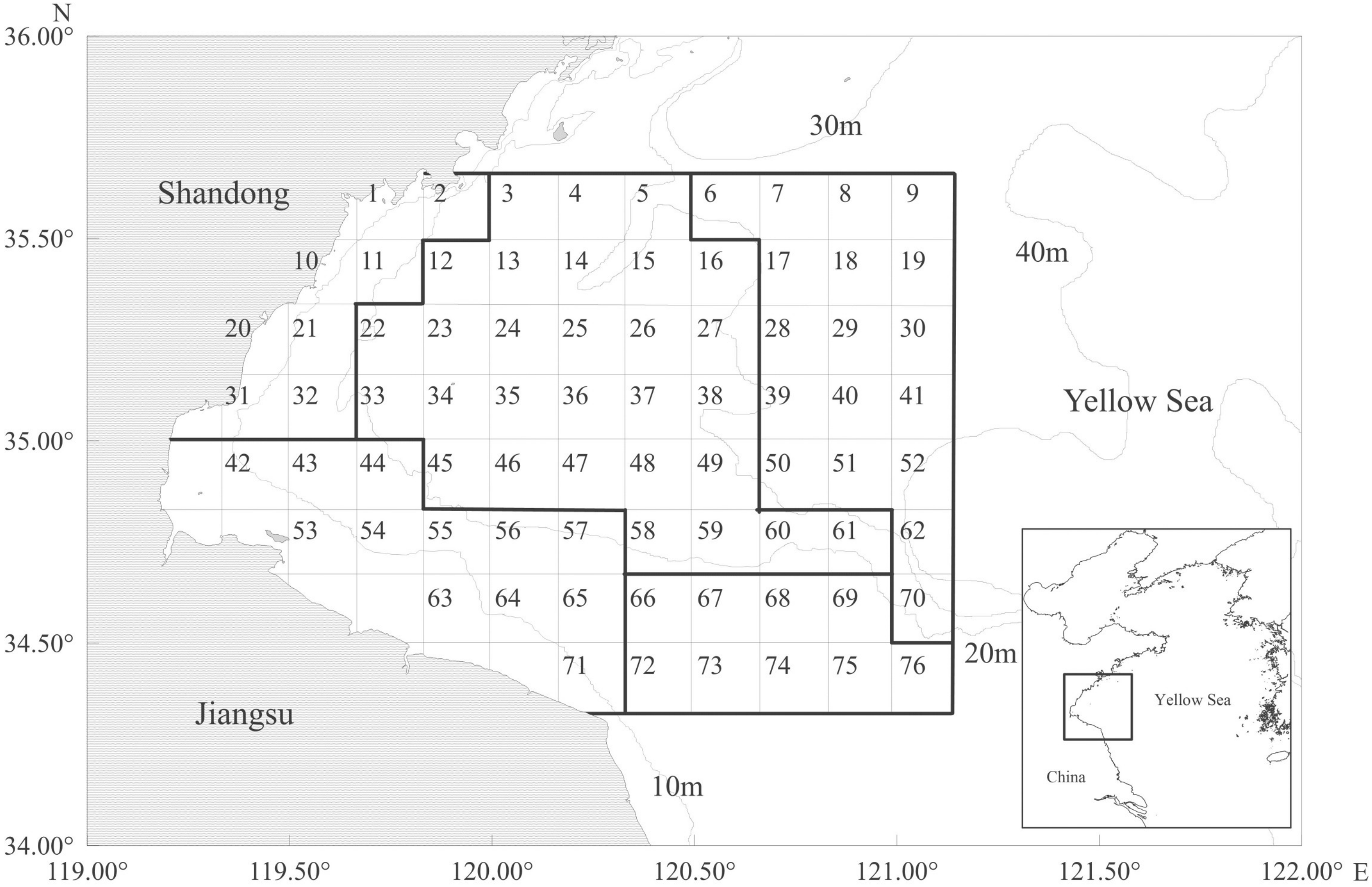

Fig 1. Map of survey stations for 2011, 2013–2016 in Haizhou Bay and adjacent waters.

https://doi.org/10.1371/journal.pone.0207457.g001

Generalized additive model. Generalized additive model (GAM) is a non-parametric extension of generalized linear model (GLM) [44]:

$$g(Y) = \alpha + \sum_{i=1}^{n} f_i(x_i) \tag{1}$$

Table 1. The description of data attributions used in the SDMs.

Environmental variable	Spring	Fall
year	2011, 2013, 2014, 2015, 2016	same as spring
sea surface salinity (SSS)	28.69–31.96	27.54–31.89
sea surface temperature (SST)	10.87˚C-19.12˚C	17.76˚C-25.85˚C
sea bottom salinity (SBS)	28.36–32.02	21.75–31.94
sea bottom temperature (SBT)	9.07˚C-17.95˚C	17.77˚C-25.89˚C
depth	6.37m-36.64m	3.77m-39.86m
sediment	sand, sandy silt, sand-silt-clay	same as spring
longitude	119.42˚E-121.08˚E	same as spring
latitude	34.42˚N-35.58˚N	same as spring

https://doi.org/10.1371/journal.pone.0207457.t001

Where g() is the monotonic link function that establishes a relationship between the mean of the response variable and predictive variables, f_i is a 'smoothed' function of explanatory variables, which enables to flexibly describe non-linear relationships [34, 45]. α is the intercept, and n is the number of explanatory variables.

Random forest. Random forest (RF) is an ensemble learning approach that generates multiple regression or decision trees [46, 47]. RF often shows satisfactory performance on prediction and gains increasing attention in a wide range of research areas. This method is typically implemented with the following steps [48]:

(i) Draw n_{tree} bootstrapped samples of the training dataset from the original data. (ii) Build multiple classification or regression trees with the bootstrap samples, in which each node of the unpruned tree is split by sampling m_{try} variables randomly and the best split is chosen automatically. (iii) Aggregate these units of tree information to attain the output.

In our study, the number of trees (n_{tree}) was set to 2000, and we trained models with different m_{try} values and chose the optimal $m_{try} = 2$ when RF performed best.

Artificial neural network. Artificial neural network (ANN), inspired by the structure and activity of human brain, is a powerful tool for ecological issues that are difficult to be recognized or predicted by traditional statistical methods [27]. There exist many types of ANNs, but a common type and the one used in this study is specified as one hidden-layer with a feed-forward network trained by a back-propagation algorithm [49]. Specifically, the network is constituted by three layers of neurons: an input layer at which predictive variables are received, a hidden layer with complex connections, and an output layer with one or more neurons to make predictions. The number of neurons in the hidden layer is determined by minimizing the tradeoff between bias and variance [50]. Here, our study selected 5 hidden neurons in the network according to the performance of training models. The connection weights between neurons of different layers were adjusted to minimize the prediction error when training ANNs [25]. The models were implemented using the R packages *mgcv*, *randomForest*, and *nnet*, respectively.

Model development and evaluation

Predictive variables were examined in the process of model development. The significant variables were selected using a stepwise variable selection procedure, which started with a null model and added one more predictive variable to the present model at each time step. For GAM, Akaike Information Criterion (AIC) [51] and Chi-square test among nested models [21] were used in variable selection for GAM, and the percentage of variance explained by the model ("variance explained") was used for RF. The contribution of each variable to the final model was measured by the 'percent deviation explained' and the IncMSE value (i.e. the changes of mean square errors) in GAM and RF, respectively [46, 47]. For ANN, Garson's algorithm [52] modified by Goh [53] was used to select predictive variables and determine their relative contributions [54]. In addition, variance explained (VE) was used as the common measure to compare the fitting capability among different models:

$$\mathrm{VE} = \left(1 - \frac{\mathrm{Var(residual)}}{\mathrm{Var(y)}}\right) \times 100\% \quad (2)$$

Where Var(residual) denoted the residual variance, and Var(y) denoted the variance of original data.

Sensitivity analysis was used to visualize the relationships between predictive variables and predictions, for which we changed each variable across its range while fixing the levels of other variables [54]. Since the relationships produced by ANN depended on the initial values and were not constant, we produced 100 response curves for each predictive variable in ANN to illustrate the variations, while other modelling methods produced one response curve for each predictive variable.

The cross-validation approach was used to evaluate the predictive performances of the models [55]. For each iteration (n = 100), the original dataset was randomly partitioned into 80% observations as training set for model building and 20% observations as test set for model validation [56, 57]. We used two metrics, the relative root-mean-square-error (RRE) and the coefficient of determination (R^2), to evaluate the accuracy and precision of model prediction [27, 43, 58]. The degree of model overfitting was indicated by the difference of R^2 between model fitting and model validation. The RRE measured the deviation of observed values and predictions, for which a smaller value implied improved predictability.

$$RRE = \frac{\sqrt{\frac{\sum_{i=1}^{n} (O_i - P_i)^2}{n}}}{O_{max} - O_{min}} \times 100\% \tag{3}$$

Here, n was the number of data points in the cross-validation, O_i was the observed values, P_i was the predicted values, O_{max} and O_{min} were the maximum and minimum values of observation.

Additionally, the standard errors of RRE and R^2 were estimated as the measures of model stability, i.e., the robustness of predictability on random datasets [18, 29]. Multi-way ANOVA was used to identify the relative importance and interactions of three factors (i.e. modelling method, species, and season) on the variation of predictability [29].

Distribution mapping

We used the Finite Volume Coastal Ocean Model (FVCOM) [59] to project the predictive variables over the whole area for mapping crab distributions. In this study, 64392 grid points were extracted from the FVCOM developed in Haizhou Bay (calibrated by College of Environmental Science and Engineering, Ocean University of China), including data of temperature and salinity by depth and time. The sediment types of these grid points were extracted from the same sediment map above. These environment data being estimated by the FVCOM were used to hindcast portunid crab distribution using the fitted models. The spatial and temporal variations of their potential distributions were shown and compared among the aforementioned modelling methods.

Results

Model fitting

VIF test suggested that SST showed multicollinearity with other variables. Thus, we used SBT in lieu of SST as a candidate predictive variable. Overall, sampling year, SBT, SBS, and sediment type were the most important factors affecting spatial distributions for different species and among survey seasons (Table 2). However, the fitted SDMs presented differences among species, modelling methods, and two seasons, regarding the retained predictive variables and their relative importance. ANNs included more variables in the fitted models than other approaches and showed much better explanatory performances, whereas GAMs included fewer variables and exhibited reduced performance for model fitting (Table 2).

We examined the sensitivity of predicted crabs' biomass to the environmental variables selected for different models. The response curves exhibited the conditional effects of one variable on the predictions when the levels of other variables were fixed (Fig 2). GAMs showed

Table 2. A Summary of fitted models for three crab species in spring and fall.

Species	Season	Model	Relative importance (%)	Variance explained (%)	The determination coefficient (R^2)	ΔR^2
C. bimaculata	Spring	GAM	depth(18.1)>year(7.1)>longitude(6.6)> SBT(2.4)	34.8	0.22	0.13
		RF	longitude(35.6)>year(24.3)>SBS(19.8)	67.3	0.28	0.5
		ANN	SBS(14.7)>SBT(11.5)>latitude>year(11.1)>sediment(7.7)	91.7	0.23	0.73
	Fall	GAM	year(12.2)>SBS(10.6)>SBT(3.6)	27.2	0.09	0.13
		RF	SBS(23.1)> sediment(19.0)>SBT(11.0)	58.7	0.11	0.61
		ANN	SBT(14.4)>year(13.4)>SBS(9.2)> sediment(6.1)>depth(6.0)	89.4	0.18	0.76
C. japonica	Spring	GAM	depth(20.9)>latitude(5.6) >year(5.3)> SSS(3.8)	36.2	0.18	0.25
		RF	Latitude(40.7)>SBT(31.4)>year(31.0)	70.5	0.20	0.59
		ANN	SBT(19.2)>depth(14.9)>year(11.2)> sediment(8.1)>latitude(4.6)	98.7	0.47	0.52
	Fall	GAM	SBS(11.8)>longitude(9.4)> SBT(1.0)	27.5	0.18	0.20
		RF	depth(36.9)>sediment(20.4)> SBS(12.8)	65.6	0.27	0.53
		ANN	SBT(16.3)>sediment(15.2)>latitude(14.5)>longitude(11.3)> year(6.4)	96.3	0.31	0.66
P. trituberculatus	Fall	GAM	year(35.7)>SBT(17.9)>sediment(1.0)	55.7	0.40	0.19
		RF	year(78.6)>SBT(22.4)	63.7	0.34	0.36
		ANN	year(14.9)>SBS(9.0)>SBT(8.1)> latitude(7.7)>sediment(7.3)	95.6	0.35	0.62

Note: Only the variables included in the optimal models were shown in the table. Relative importance referred to the contribution of predictive variables that retained in fitted models (GAM was based on 'deviation explained', RF was based on the percentage of IncMSE, ANN was based on Garson's algorithm). Variance explained by each model represented goodness-of-fit of model. The determination coefficient (R^2) represented predictive performance in the latter section of model predictability. ΔR^2 referred to the difference of R^2 between training models testing models.

https://doi.org/10.1371/journal.pone.0207457.t002

simple relationships between predictive variables and predicted biomass, whereas ANNs and RFs reflected complex relationships. The response curves in ANNs changed substantially among 100 repeats, and the reflected effects of predictive variables varied among modeling approaches. In particular, SBT show different effects on predicted distributions among models.

Model predictability

For cross-validation, RRE of all fitted models ranged from 28 to 60 and R^2 ranged from 0.08 to 0.47 (Fig 3). Among the three modelling approaches, no single method consistently outperformed others (Fig 3). GAMs and RFs provided better performances than ANNs on RRE, whereas ANNs exhibited higher R^2 (representing predictive performance). The model predictability also varied among species, i.e., *P. trituberculatus* exhibited consistently better performances, and *C. japonica was* slightly better predicted than *C. bimaculata*. The same modelling method exhibited different predictability between seasons when modelling the same crab species, especially in R^2 (Fig 3). Additionally, ANNs showed the largest difference between fitting capacity and predictability compared with other methods (Table 2).

Regarding model stability, the predictions of GAMs and RFs were more stable with lower standard errors in RRE and R^2 (Table 3) comparing with ANN. The stability of GAMs and RFs were similar in spring but largely different in fall. Given the above, RF showed the best predictive performance with lower RRE and higher stability, followed by ANN with the highest R^2.

ANOVA suggested all three factors were significant for predictability (Table 4). The modeling method showed a greater influence than species and season on RRE, whereas species was the

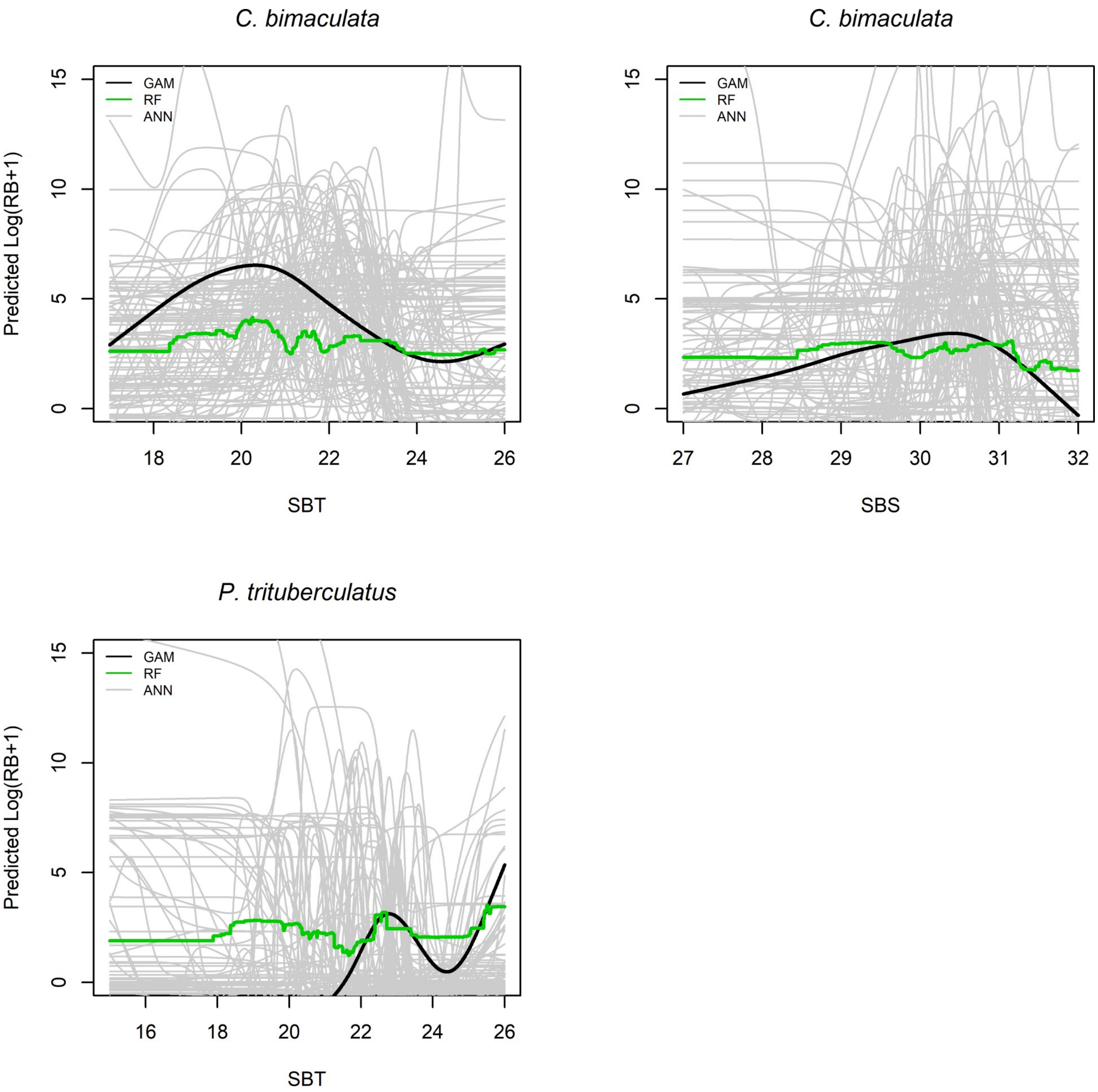

Fig 2. Contribution of influential environmental variables to the relative biomass (RB) of *C. bimaculata*, *P. trituberculatus* in fall. The results were derived from GAM, RF, and ANN, for which ANNs were examined with 100 repeats, and the other methods showed one curve.

https://doi.org/10.1371/journal.pone.0207457.g002

most influential factor for R^2. In addition, the interactions between method and both season and species were significant, that is, the modelling methods performed differently among seasons and species. The relatively weak effect of the method to season interaction suggested that the model predictability was relatively stable among seasons, whereas the large method to species interaction effect translated to unstable performances of modelling methods among species.

Mapping crab distributions

The distributions were mapped in each year using the most reliable models for each crab species i.e., RF for *C. bimaculata* and *C. japonica*, GAM for *P. trituberculatus*, respectively. The

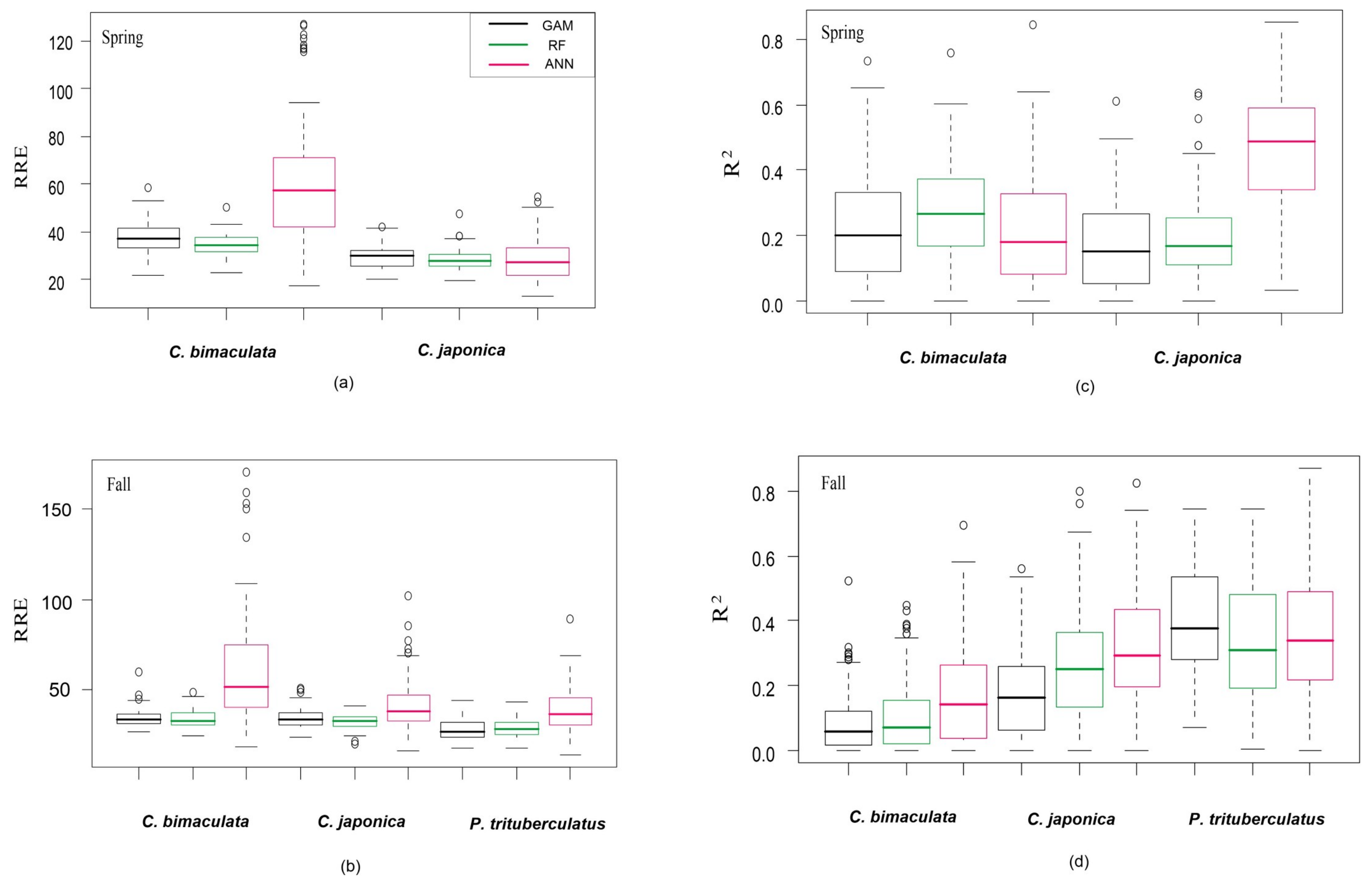

Fig 3. The predictability of crab distribution models measured by RRE and R^2. Each plot showed one metric for all species during spring or fall, and *P. trituberculatus* was absent from spring sampling. The dispersals of RRE and R^2 resulted from 100 times repeats in cross-validation.

https://doi.org/10.1371/journal.pone.0207457.g003

smaller biomass of *C. bimaculata* was predicted to be mostly in southwestern Haizhou Bay (Fig 4, results in fall as examples), whereas *C. japonica* was mainly located in the southwestern coastal waters (Fig 5). *P. trituberculatus* was predicted to distribute more evenly in the survey

Table 3. The stability of predictive capacity of crab distribution models measured by the standard errors of RRE and R^2 in spring and fall.

Species	Model	Spring		Fall	
		SE of RRE	SE of R^2	SE of RRE	SE of R^2
C. bimaculata	GAM	0.06	0.15	0.05	0.08
	RF	0.04	0.15	0.05	0.10
	ANN	0.24	0.18	0.30	0.16
C. japonica	GAM	0.05	0.14	0.05	0.13
	RF	0.04	0.13	0.04	0.17
	ANN	0.08	0.19	0.14	0.19
P. trituberculatus	GAM			0.06	0.16
	RF			0.05	0.19
	ANN			0.12	0.20

Note: The standard errors of RRE and R^2 were calculated by using 100 cross-validation results for each model.

https://doi.org/10.1371/journal.pone.0207457.t003

Table 4. The effect of influential factors on model predictive performance examined by ANOVA.

Criteria	Factors	SSE	Pr(>F)
RRE	method	5.730	< 0.001
	species	4.745	< 0.001
	season	0.317	< 0.001
	method:species	3.322	< 0.001
	method:season	0.223	< 0.001
R^2	method	2.50	< 0.001
	species	6.84	< 0.001
	season	1.59	< 0.001
	method:species	2.97	< 0.001
	method:season	0.14	0.069

Note: SSE referred to the sum of square errors of REE or R^2 attributed to each influential factor. (method: species) and (method: season) denoted the interactions between method and species or season, respectively.

https://doi.org/10.1371/journal.pone.0207457.t004

area (Fig 6). There was a substantial difference in the predicted density of *P. trituberculatus* between 2011 and the other years.

The predicted distribution maps were substantially different among modelling methods. Using *C. japonica* as an example, the results of GAM were similar to RF, showing higher biomass in the southwestern coastal waters in fall, whereas the results of ANN were substantially variable among survey years (S1 Fig). Considering seasonal differences in predictions, the maps showed that *C. japonica* tended to live nearshore, and the distributions were more stable in spring than that in fall (S2 Fig).

Discussion

Identifying reliable models for projecting species distributions is important for fisheries conservation, management, and spatial planning [60]. The present study showed a comprehensive framework for model assessment in regards to fitting performances, species response curves, predictive capacity, and model stability. For the three species, no method consistently outperformed others. Our results highlighted the advantages and shortcomings of the models. In particular, we found RF was the most reliable method with robust predictions. In addition, the predictive performances were more variable among species than among modelling methods, which was consistent with previous studies [24, 33, 51, 61], suggesting that individual traits of a species should be highlighted in the choice of appropriate methods. Based on our results, we recommend use of multiple modelling approaches to generate more robust predictions for fisheries management [60].

In this study, all three approaches showed substantially better performances with training data compared to that with test data, implying a risk of overfitting. This may be attributed to both the complex species response to environmental variables and the limited data availability [62, 63]. Among the three modelling methods, RFs provided the best predictability and stable predictions over years but had a lower R^2 compared to ANNs. Actually, the relative predictive capacity of ANNs and RFs varied greatly among studies with respect to different objectives and circumstances of their applications [35, 43, 64]. For instance, some studies suggested that RFs had advantages over ANNs in relation to avoiding overfitting [65] and simple adjustment to parameters [35], whereas ANNs could be adaptively trained to solve more complex ecological relationships [25, 27]. For modelling response curves, the simple patterns provided by GAM appeared to be more reasonable, whereas the complex relationships identified by ML methods

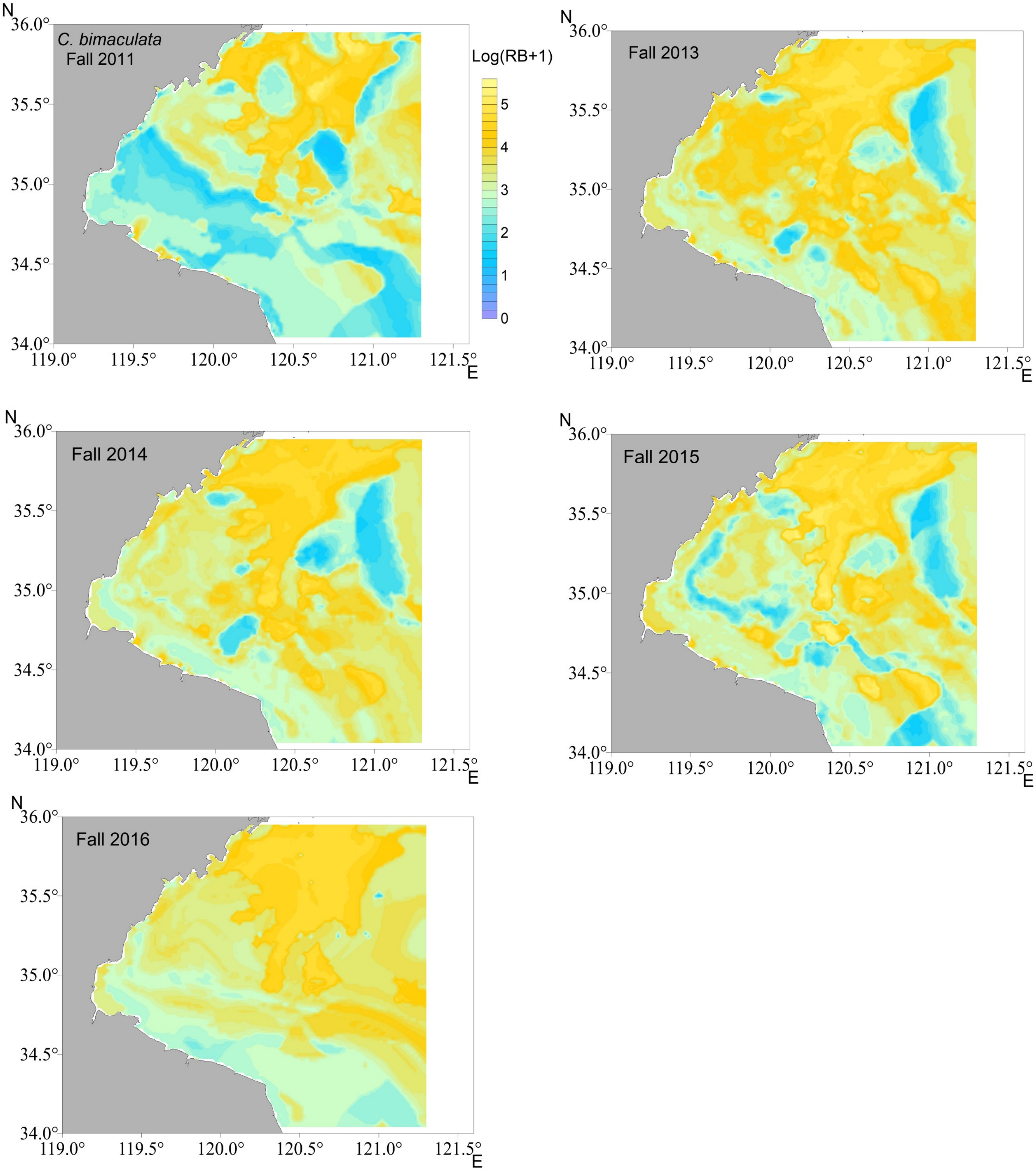

Fig 4. Spatial distribution of relative biomass (RB) for *C. bimaculata* in fall of each survey year predicted by random forest (RF) in Haizhou Bay.

https://doi.org/10.1371/journal.pone.0207457.g004

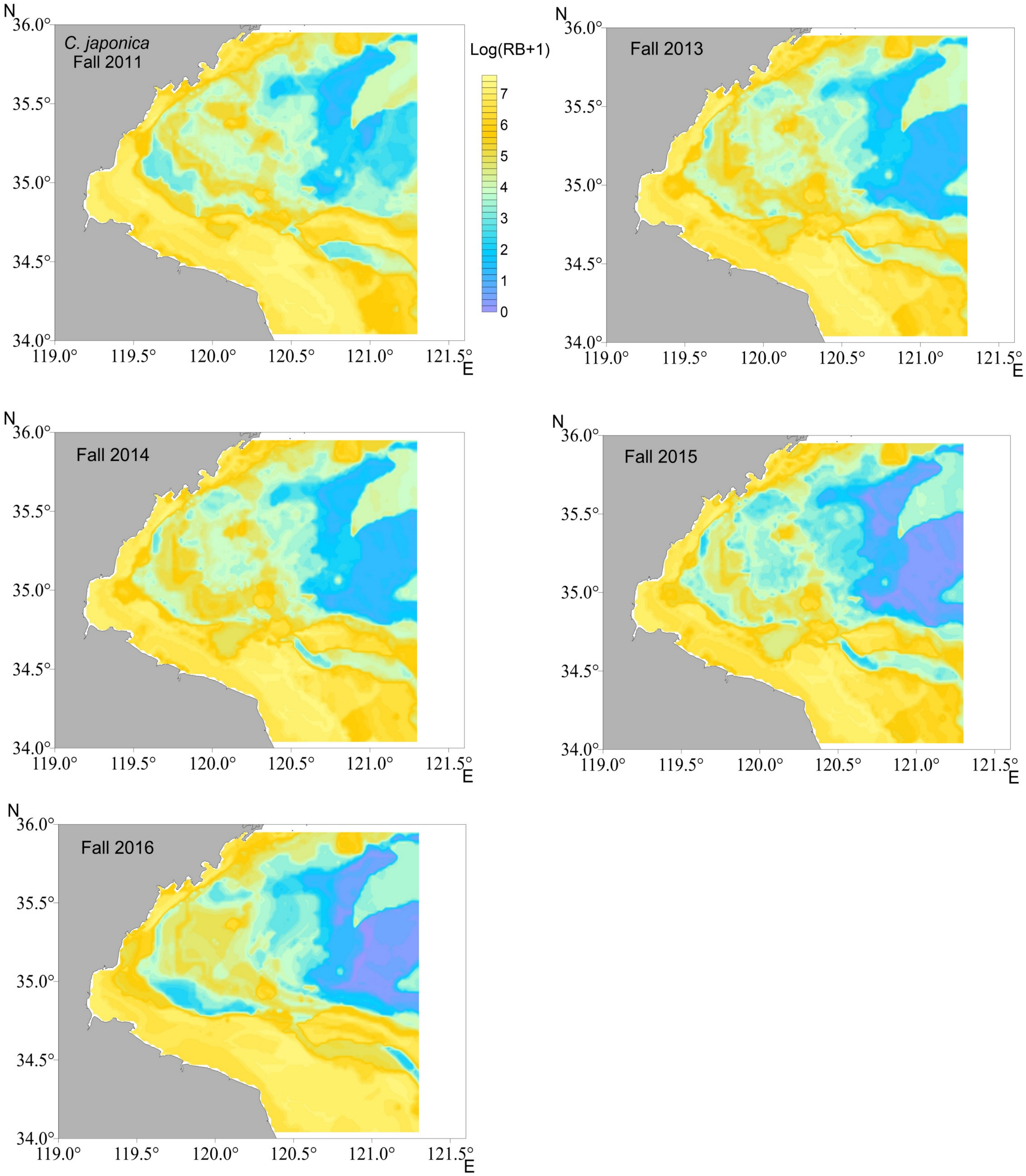

Fig 5. Spatial distribution of relative biomass (RB) for *C. japonica* in fall of each survey year predicted by random forest (RF) in Haizhou Bay.

https://doi.org/10.1371/journal.pone.0207457.g005

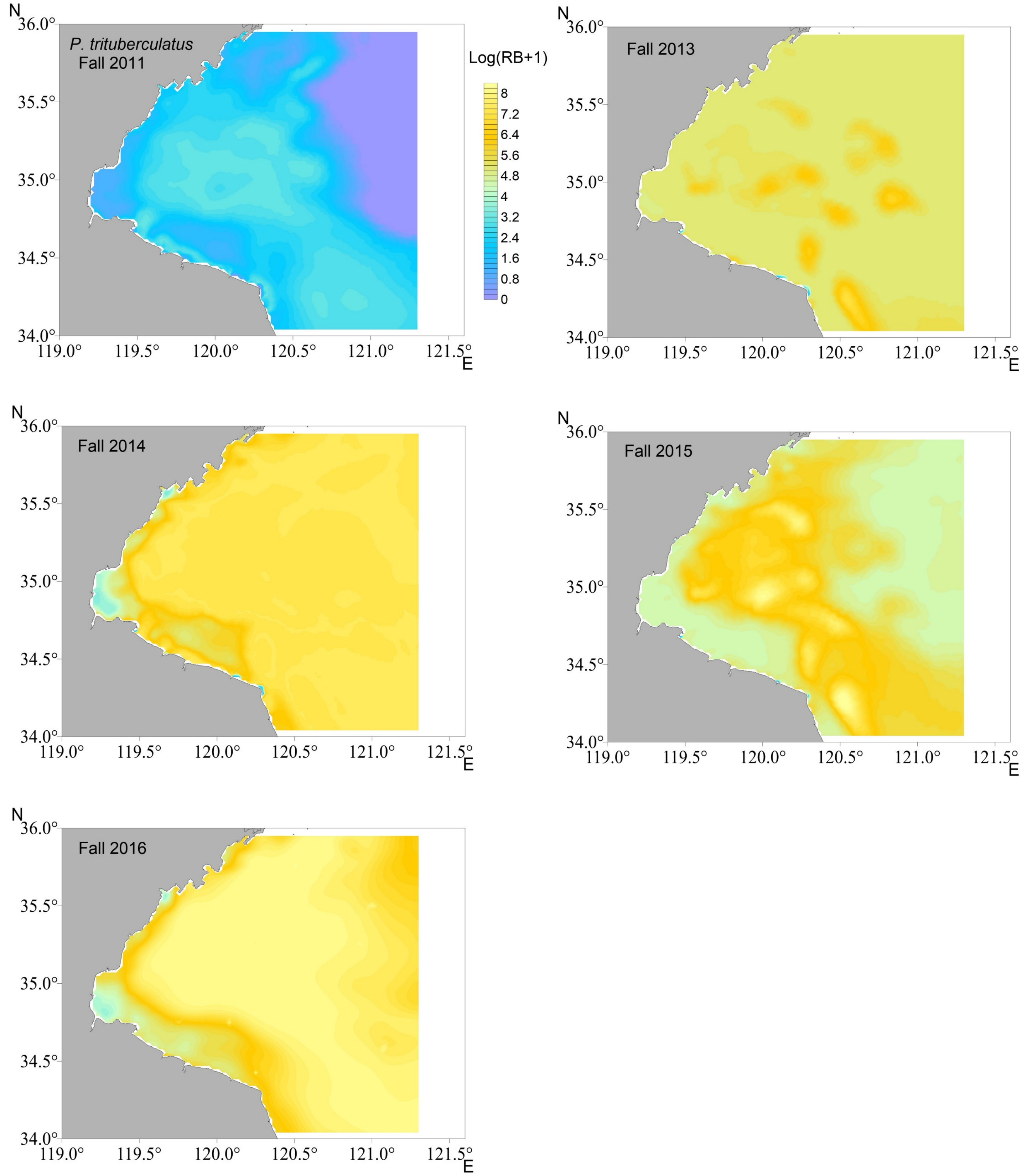

Fig 6. Spatial distribution of relative biomass (RB) for *P. trituberculatus* in fall of each survey year predicted by generalized additive model (GAM) in Haizhou Bay.

https://doi.org/10.1371/journal.pone.0207457.g006

did not necessarily mean they were unrealistic, because species-environment responses often tend to be complex, even after accounting for interactions between variables [21]. Given the synthetical evaluation of models, RFs showed better tradeoff among predictability and ecological interpretability and were more suitable for the crabs' fisheries management.

Examining the species responses to environmental factors was conducive to understanding physiological and behavioral characteristics of different species [66]. SBT and SBS, the key environmental variables for the three crabs, have been shown to play a decisive role in many short living species [13]. *C. japonica* showed a low biomass at the temperature range 10–13˚C, consistent with its preference of warm temperature. Likewise, SBS significantly influenced the distribution of *C. bimaculata*, indicating the preferred range of salinity 29–31. *P. trituberculatus* showed no optimum temperature range but instead exhibited more than one peak in the response curve (Fig 2). This result might be partly due to the ongoing southward migration of *P. trituberculatus* in the fall, which coincided with decreasing northern water temperatures. The latitude also showed large effects on crabs' distributions which might be related with habitat differences in the north and south of the bay. Moreover, there was no guarantee that the determinant variables were included in our analysis, such as dissolved oxygen [17], precipitation and food availability [22].

Although the three crabs are closely related in taxon, their corresponding SDMs showed substantial variations in predictive performances, which were consistent with previous studies [61, 67]. It should be noted that different biological and life history traits may influence model capacity to capture species-environment relationships [16, 30, 68]. The large body size of *P. trituberculatus* may enable the species to hold on preferable environmental conditions when there are environmental fluctuations [68], resulting in better predictions. On the other hand, *C. bimaculata* is characterized by small size and high prevalence, which may result in some individuals staying in less satisfactory habitats, and therefore explain lower model prediction power [68]. In addition, different spawning activities and range sizes of three crabs may also influence the model performances through making their environmental requirements difficult to be describe [69].

Annual variation of abundance in fall and the absence in spring led to substantial uncertainty in the fishery management of *P. trituberculatus*. As the population of *P. trituberculatus* have dramatically declined over the last few decades [70], the risk of uncertainty should be explicitly and carefully accounted for in the future fishery management strategy. On the other hand, the annual populations of the other two species in fall were more robust when using their best fitted models therefore fishing effort might accommodate to this pattern for improving fishing efficiency. However, the distribution maps of *C. japonica* predicted by suboptimal ANN in fall showed variations among years, this result alerted managers to combine multiple models to inform the stock assessment.

Several conclusions of this study were highlighted for future SDM practices and the management of crab fisheries. For example, the performances of the modelling methods were relatively stable among seasons but varied substantially among species, implying that seasonality might be less concerned when choosing suitable modelling techniques for species. Besides, the high SE of REE and R^2 suggested that the performances of ANN were not robust although they provided superior model fitting. The complex model structures implied that sufficiently large sample size of data should be desired in the use of ANN as well as other ML methods. In particular, *C. bimaculata* was characterized by wide tolerances in salinity and temperature [71] making it hard to be simulated. A larger sample size may benefit robust establishment of environmental requirements for this species. Long term climatic variations might be influential for species distributions, but were not incorporated due to the relatively short survey time series and improper resolutions [61]. Importantly, the SDMs in this study did not include consideration of biotic interaction and competitive exclusion [72, 73], which would be critical to correctly reflect realized ecological niches. These problems should also be considered in future studies.

Supporting information

S1 Excel. Relevant data used in model development. The unit of relative biomass is g/h.
(XLSX)

S1 Fig. Spatial distribution maps of relative biomass (RB) for *C. japonica* in fall of 2011, 2013 and 2014 predicted by three modelling methods (i.e. GAM, RF and ANN) in Haizhou Bay.
(EMF)

S2 Fig. Spatial distribution maps of relative biomass (RB) for *C. japonica* in two seasons of 2011 and 2013–2016 predicted by artificial neural network (ANN) in Haizhou Bay.
(EMF)

S1 Table. Summary of fitted GAMs for three crab species in spring and fall.
(DOCX)

Acknowledgments

We would like to thank the colleagues of the Fishery Ecosystem Monitoring and Assessment Laboratory of Ocean University of China for their help in sample collection and experimental assistance. We thank Dr. Robert Boenish from Environmental Defense Fund and two anonymous reviewers for their help in improving the manuscript of this study.

Author Contributions

Conceptualization: Binduo Xu, Ying Xue, Yiping Ren.

Data curation: Jing Luan, Chongliang Zhang, Binduo Xu.

Formal analysis: Jing Luan, Chongliang Zhang.

Funding acquisition: Chongliang Zhang, Binduo Xu, Yiping Ren.

Investigation: Jing Luan.

Methodology: Chongliang Zhang, Binduo Xu, Yiping Ren.

Project administration: Jing Luan.

Software: Jing Luan, Chongliang Zhang.

Supervision: Chongliang Zhang, Binduo Xu, Ying Xue.

Validation: Chongliang Zhang, Yiping Ren.

Writing – original draft: Jing Luan.

Writing – review & editing: Jing Luan, Yiping Ren.

References

1. Jackson JBC, Kirby MX, Berger WH, Bjorndal KA, Botsford LW, Bourque BJ, et al. Historical overfishing and the recent collapse of coastal ecosystems. Science. 2001; 293(5530): 629–637. https://doi.org/10.1126/science.1059199 PMID: 11474098

2. Pauly D, Christensen V, Guénette S, Pitcher TJ, Sumaila UR, Walters CJ, et al. Towards sustainability in world fisheries. Nature. 2002; 418(6898): 689–695. https://doi.org/10.1038/nature01017 PMID: 12167876
3. Doney SC, Ruckelshaus M, Duff JE, Barry JP, Chan F, English CA, et al. Climate change impacts on marine ecosystems. Annual Review of Marine Science. 2012; 4: 11–37. https://doi.org/10.1146/annurev-marine-041911-111611 PMID: 22457967
4. Heithaus MR, Frid A, Wirsing AJ, Worm B. Predicting ecological consequences of marine top predator declines. Trends in Ecology & Evolution. 2008; 23(4): 202–210.
5. Smith MT, Addison JT. Methods for stock assessment of crustacean fisheries. Fisheries Research. 2003; 65(1–3): 231–256.
6. Qi HM. Relationship of Crustaceans Community Structure and Resources Distribution with Environment Factors in the Jiaojiang Estuary. Ph.D. Thesis, Shanghai Ocean University. 2014.
7. Yu C, Song H, Yao G, Lu H. [Composition and distribution of economic crab species in the East China Sea]. Journal of Oceanology and Limnology in Chinese. 2006; 37(1): 53–60. Chinese.
8. Zhang B, Thang Q, Jin X. [Functional groups of communities and their major species at high trophic level in the Yellow Sea ecosystem]. Acta Ecologica Sinica in Chinese. 2009; 29(3): 1099–1111.
9. Ren LP, Qin Y, Li XC, Sun YN, Wang RX. Isolation and characterization of polymorphic microsatellite loci in the swimming crab Portunus trituberculatus (Portunidae). Genetics and Molecular Research. 2013; 12(4): 5911–5915. https://doi.org/10.4238/2013.November.22.19 PMID: 24301961
10. Yang G, Li F, Lv Z, Xu B, Yuan X, Wang X. [Study on the community structure of crabs in the coastal waters along Shandong Peninsula]. Acta Oceanologica Sinica. 2017; 39(8): 48–61. Chinese.
11. Warner GF. Biology of crabs. London: Paul Elek Scientific Book Ltd; 1977.
12. Hijuelos AC, Sable SE, O'Connell AM, Geaghan JP, Lindquist DC, White ED. Application of species distribution models to identify estuarine hot spots for juvenile nekton. Estuaries and coasts. 2017; 40(4): 1183–1194.
13. Sugilar H, Park YC, Lee NH, Han DW, Han KN. Population dynamics of the swimming crab Portunus trituberculatus (Miers, 1876)(Brachyura, Portunidae) from the West Sea of Korea. International Journal of Oceanography and Marine Ecological System. 2012; 1(2): 36–49.
14. Xu XH, Yan BL, Xu JT. [Tolerance of Charybdis japonica to several Environmental Factors]. Proceedings of Conference on Environmental Pollution and Public Health (CEPPH 2012); 2012 Aug 484–487; Shanghai, China: 2012. Chinese.
15. Frusher SD, Giddins RL, Smith TJ. Distribution and abundance of grapsid crabs (Grapsidae) in a mangrove estuary: effects of sediment characteristics, salinity tolerances, and osmoregulatory ability. Estuaries. 1994; 17(3): 647–654.
16. Jones MB. Limiting factors in the distribution of intertidal crabs (Crustacea: Decapoda) in the avonheathcote estuary, christchurch. New Zealand Journal of Marine and Freshwater Research. 1976; 10 (4): 577–587.
17. Narita T, Ganmanee M, Sekiguchi H. Population dynamics of portunid crab *Charybdis bimaculata* in Ise Bay, central Japan. Fisheries Science. 2008; 74(1): 28–40.
18. Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecological Modelling. 2000; 135(2–3): 147–186.
19. Compton TJ, Leathwick JR, Inglis GJ. Thermogeography predicts the potential global range of the invasive European green crab (Carcinus maenas). Diversity and Distributions, 2010; 16(2): 243–255.
20. Herborg LM, Rudnick DA, Siliang Y, Lodge DM, MacIAAC HJ. Predicting the range of Chinese mitten crabs in Europe. Conservation Biology. 2007; 21(5): 1316–1323. https://doi.org/10.1111/j.1523-1739.2007.00778.x PMID: 17883496
21. Jensen OP, Seppelt R, Miller TJ, Bauer LJ. Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. Marine Ecology Progress Series. 2005; 299(1): 239–255.
22. Hardy SM, Lindgren M, Konakanchi H, Huettmann F. Predicting the distribution and ecological niche of unexploited snow crab (Chionoecetes opilio) populations in Alaskan Waters: a first open-access ensemble model. Integrative & Comparative Biology. 2011; 51(4): 608–622.
23. Swain DP, Wade EJ. Spatial distribution of catch and effort in a fishery for snow crab (Chionoecetes opilio): tests of predictions of the ideal free distribution. Canadian Journal of Fisheries and Aquatic Sciences. 2003; 60(8): 897–909.
24. Olden JD, Jackson DA. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology. 2002; 47(10): 1976–1995.

25. Maravelias CD, Haralabous J, Papaconstantinou C. Predicting demersal fish species distributions in the Mediterranean sea using artificial neural networks. Marine Ecology Progress Series. 2003; 255: 249–259.

26. Guisan A, Thuiller W. Predicting species distribution: offering more than simple habitat models. Ecology Letters. 2005; 8(9): 993–1009.

27. Olden JD, Jackson DA. Fish–habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. Transactions of the American Fisheries Society. 2001; 130(5): 878–897.

28. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A. Effects of sample size on the performance of species distribution models. Diversity and distributions. 2008; 14(5): 763–773.

29. Segurado P, Araújo MB. An evaluation of methods for modelling species distributions. Journal of Biogeography. 2004; 31(10): 1555–1568.

30. Hallstan S, Johnson RK, Sandin L. Effects of dispersal-related factors on species distribution model accuracy for boreal lake ecosystems. Diversity. 2013; 5(2): 393–408.

31. França S, Cabral HN. Predicting fish species distribution in estuaries: Influence of species' ecology in model accuracy. Estuarine, Coastal and Shelf Science. 2016; 180: 11–20.

32. Santika T, Hutchinson MF. The effect of species response form on species distribution model prediction and inference. Ecological Modelling. 2009; 220(19): 2365–2379.

33. Li M, Zhang C, Xu B, Xue Y, Ren Y. Evaluating the approaches of habitat suitability modelling for white-spotted conger (Conger myriaster). Fisheries Research. 2017; 195: 230–237.

34. Guisan A, Jr TCE, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling. 2002; 157(2–3): 89–100.

35. Li Z, Ye Z, Wan R, Zhang C. Model selection between traditional and popular methods for standardizing catch rates of target species: a case study of Japanese Spanish mackerel in the gillnet fishery. Fisheries Research. 2015; 161: 312–319.

36. Zhang SY, Zhang HJ, Jiao JP, Li YS, Zhu KW. [Change of ecological environment of artificial reef waters in Haizhou Bay]. Journal of Fisheries of China. 2006; 30(4): 475–480. Chinese.

37. Sun X; Zhang S, Zhao Y, Zhang H. [Community structure of fish and macroinvertebrates in the artificial reef sea area of Haizhou Bay]. Journal of Shanghai Ocean University. 2010; 19(4): 505–513. Chinese.

38. Jongman RH, Ter Braak CJ, Van Tongeren OF. 1st ed. Data analysis in community and landscape ecology. Cambridge: Cambridge university press; 1995.

39. Brosse S, Guegan JF, Tourenq JN, Lek S. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological modelling. 1999; 120 (2–3): 299–311.

40. Brosse S, Lek S. Relationships between environmental characteristics and the density of age-0 Eurasian perch Perca fluviatilis in the littoral zone of a lake: a nonlinear approach. Transactions of the American Fisheries Society. 2002; 131(6): 1033–1043.

41. Shepard FP. Nomenclature based on sand-silt-clay ratios. Journal of Sedimentary Research. 1954; 24 (3): 151–158.

42. Parra HE, Pham CK, Menezes GM, Rosa A, Tempera F, Morato T. Predictive modeling of deep-sea fish distribution in the Azores. Deep-Sea Research Part II: Topical Studies in Oceanography. 2017; 145: 49–60.

43. Olaya-Marín EJ, Martínez-Capel F, Vezza P. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge & Management of Aquatic Ecosystems. 2013; 139(409): 07.

44. Hastie T, Tibshirani R. Generalized Additive Models. 1st ed. London: Chapman and Hall; 1990.

45. Leathwick JR, Elith J, Hastie T. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling. 2006; 199(2): 188–196.

46. Breiman L. Random forests. Machine learning. 2001; 45(1): 5–32.

47. Bradter U, Kunin WE, Altringham JD, Thom TJ, Benton TG. Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods in Ecology & Evolution. 2013; 4(2): 167–174.

48. Liaw A, Wiener M. Classification and regression by random Forest. R News. 2002; 2 (3): 18–22. Available from: https://www.researchgate.net/publication/228451484.

49. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323(6088): 533–536.

50. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Computation. 1992; Jan. 4(1): 1–58.

51. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). New York: Springer; 1998. pp. 199–213.

52. Garson GD. Interpreting neural-network connection weights. Artificial Intelligence Expert. 1991; 6(4): 47–51.

53. Goh ATC. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 1995; 9(3): 143–151.

54. Olden JD. Predictive models for freshwater fish community composition. Ph.D. Thesis, The University of Toronto. 2000. Available from: http://hdl.handle.net/1807/13839.

55. Franklin J. Mapping Species Distributions: Spatial Inference and Prediction. London: Cambridge University Press; 2010.

56. Smith P. A. Autocorrelation in logistic regression modelling of species' distribution. Global Ecology and Biogeography Letters. 1994; 4: 47–61.

57. Tanaka K, Chen Y. Spatiotemporal variability of suitable habitat for American lobster (Homarus americanus) in Long Island Sound. Journal of Shellfish Research. 2015; 34 (2): 531–543.

58. Zhang H, Zimba PV. Analyzing the effects of estuarine freshwater fluxes on fish abundance using artificial neural network ensembles. Ecological Modelling. 2017; 359: 103–116.

59. Chen C, Beardsley RC, Cowles GW. An unstructured grid, finite-volume coastal ocean model (FVCOM) system. Oceanography. 2006; 19(1): 78–89.

60. Gritti ES, Gaucherel C, Crespo-Perez M-V, Chuine I. How Can Model Comparison Help Improving Species Distribution Models? PLOS ONE. 2013; 8(7): e68823. https://doi.org/10.1371/journal.pone.0068823 PMID: 23874779

61. Guisan A, Zimmermann NE, Elith J, Graham CH, Phillips S, Peterson AT. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecological monographs. 2007; 77(4): 615–630.

62. Chahouki MAZ, Ahvazi LK, Azarnivand H. Comparison of three modeling approaches for predicting plant species distribution in mountainous scrub vegetation (Semnan Rangelands, Iran). Polish Journal of Ecology. 2012; 60(2): 277–289.

63. Froeschke BF, Tissot P, Stunz GW, Froeschke JT. Spatiotemporal predictive models for juvenile southern flounder in Texas estuaries. North American Journal of Fisheries Management. 2013; 33(4): 817–828.

64. Nitze I, Schulthess U, Asche H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. Proc. of the 4th GEOBIA; 2012 May 35–40; Rio de Janeiro, Brazil: 2012.

65. Chen XZ, Fan W, Cui XS, Zhou W, Tang F. [Fishing ground forecasting of Thunnus alalung in Indian Ocean based on random forest]. Acta Oceanologica Sinica. 2013; 35(1): 158–164. Chinese.

66. Zeng Y, Yeo DCJ. Assessing the aggregated risk of invasive crayfish and climate change to freshwater crabs: A Southeast Asian case study. Biological Conservation. 2018; 223: 58–67.

67. Thuiller W. BIOMOD–optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology. 2003; 9(10): 1353–1362.

68. Mcpherson JM, Jetz W. Effects of species' ecology on the accuracy of distribution models. Ecography. 2007; 30(1): 135–151.

69. Kwak SN, Park JM, Huh SH. Seasonal Variations in Species Composition and Abundance of Fish and Decapods in an Eelgrass (Zostera marina) Bed of Jindong Bay. Journal of the Korean Society of Marine Environment & Safety. 2014; 20(3): 259–269.

70. Lv J, Liu P, Gao B, Wang Y, Wang Z, Chen P, et al. Transcriptome analysis of the *Portunus trituberculatus*: de novo assembly, growth-related gene identification and marker discovery. PLOS one. 2014; 9 (4): e94055. https://doi.org/10.1371/journal.pone.0094055 PMID: 24722690

71. Rushton SP, Ormerod SJ, Kerby G. New paradigms for modelling species distributions? Journal of Applied Ecology. 2004; 41(2): 193–200.

72. Zurell D, Zimmermann NE, Sattler T, Nobis MP, Schröder B. (2016). Effects of functional traits on the prediction accuracy of species richness models. Diversity and Distributions. 2016 Aug; 22(8): 905–917.

73. Godsoe W, Franklin J, Blanchet FG. Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. Ecology and evolution. 2017; 7(2): 654–664. https://doi.org/10.1002/ece3.2657 PMID: 28116060