Acta Oceanol. Sin., 2024, Vol. 43, No. 12, P. 113–122
https://doi.org/10.1007/s13131-024-2387-7
http://www.aosocean.com
E-mail: ocean2@hyxb.org.cn

Impacts of data sources on the predictive performance of species distribution models: a case study for *Scomber japonicus* in the offshore waters southern Zhejiang, China

Wen Ma[1, 2], Ling Ding[3], Xinghua Wu[4], Chunxia Gao[1, 5], Jin Ma[1]*, Jing Zhao[1]

[1] College of Marine Living Resource Sciences and Management, Shanghai 201306, China

[2] College of the Environment and Ecology, Xiamen University, Xiamen 361102, China

[3] Shanghai Investigation, Design & Research Institute Co., Ltd, Shanghai 200335, China

[4] China Three Gorges Corporation, Wuhan 430010, China

[5] The Key Laboratory of Sustainable Exploitation of Oceanic Fisheries Resources, Shanghai Ocean University, Shanghai 201306, China

Received 28 April 2024; accepted 12 September 2024

Abstract

As our understanding of ecology deepens and modeling techniques advance, species distribution models have grown increasingly sophisticated, enhancing both their fitting and predictive capabilities. However, the dependability of predictive accuracy remains a critical issue, as the precision of these predictions largely hinges on the quality of the base data. We developed models using both field survey and remote sensing data from 2016 to 2020 to evaluate the impact of different data sources on the accuracy of predictions for *Scomber japonicus* distributions. Our research findings indicate that the variability of water temperature and salinity data from field suvery is significantly greater than that from remote sensing data. Within the same season, we found that the relationship between the abundance of *S. japonicus* and environmental factors varied significantly depending on the data source. Models using field survey data were able to more accurately reflect the complex relationships between resource distribution and environmental factors. Additionally, in terms of model predictive performance, models based on field survey data demonstrated greater accuracy in predicting the abundance of *S. japonicus* compared to those based on remote sensing data, allowing for more accurate mastery of their spatial distribution characteristics. This study highlights the significant impact of data sources on the accuracy of species distribution models and offers valuable insights for fisheries resources management.

Key words: species distribution model, remote sensing data, field survey data, predictive performance, offshore waters southern Zhejiang, *Scomber japonicus*

Citation: Ma Wen, Ding Ling, Wu Xinghua, Gao Chunxia, Ma Jin, Zhao Jing. 2024. Impacts of data sources on the predictive performance of species distribution models: a case study for *Scomber japonicus* in the offshore waters southern Zhejiang, China. Acta Oceanologica Sinica, 43(12): 113–122, doi: 10.1007/s13131-024-2387-7

1 Introduction

With the increasing understanding of ecology and substantial improvements in the levels of modeling, more complex species distribution models have been developed, which have improved model fitting and prediction to a greater extent while providing an effective research method for understanding the complex, nonlinear relationships and interactions between organisms and their environments (Ma et al., 2020; Yu et al., 2020; Zhao et al., 2014). However, prediction performance is considered a critical issue in model applications (Guisan and Zimmermann, 2000; Luan et al., 2021). Extensive studies have been conducted on model selection (Liu et al., 2021; Ma et al., 2022b), multicollinearity (Liu et al., 2019; Zhang et al., 2020; Zhao et al., 2014), and environmental interpolation methods (Pan et al., 2021) to improve model fitting and prediction. However, only a few studies have focused on the impact of modeling data on the prediction of marine living resources. Related studies have shown that when different methods are used to access marine environmental data, there may be some differences in the data itself (La Marca et al., 2019; Lei et al., 2022), which, in turn, have implications for predicting the spatial distribution of fishery resources and exploring suitable habitats. Addressing this issue has become more urgent as conservation planning and other measures increasingly rely on spatial predictions of target fish species, as well as biodiversity based on habitat (Johnston et al., 2017; Liu et al., 2019; Zhao et al., 2014).

Marine environmental data are an important part of model building (Zhao et al., 2014), and data such as water temperature, salinity, dissolved oxygen, and chlorophyll *a* can be obtained through surveys or remote sensing (Li et al., 2014; Luan et al., 2018; Zhao et al., 2014). However, both data sources have certain advantages and disadvantages in the acquisition process. Field surveys, while able to obtain more accurate hydrological environment data, face challenges in monitoring over large areas. This

Foundation item: The Research Project of China Yangtze River Three Gorges Group Limited under contract No. 201903173; the Zhejiang Mariculture Research Institute of China under contract No. 325000.
*Corresponding author, E-mail: jma@shou.edu.cn

difficulty arises due to constraints such as the cost and time associated with surveys, often resulting in a relatively small amount of data being collected (Xue et al., 2018; Zhao et al., 2014). This limitation necessitates the use of interpolation to estimate hydrological environment data for areas not covered by surveys, which inevitably reduces data accuracy. The decrease in accuracy directly impacts the reliability of predictions regarding the spatial distribution of fishery resources (La Marca et al., 2019; Pan et al., 2021). With continuous improvements in satellite technology, hydrological environment information can be acquired on a large scale, generating continuous datasets with multiple spatial and temporal resolutions (Wu et al., 2022). These datasets can be used to assess changes in the ecological environment, and they can also be incorporated into models to assess trends in ecological services and quality (Wang et al., 2022). However, when surveying the hydrological environment via satellite, data accuracy is reduced because of factors such as spatial resolution, cloud cover occlusion, and inversion algorithms (Scales et al., 2017; Stock and Subramaniam, 2020; Welch et al., 2020), which in turn impacts target fish resource predictions. Although the applications of field surveys and remote sensing data for environmental modelling have been extensively studied, the majority of research continues to focus on land and stream environments (La Marca et al., 2019; Johnston et al., 2017). However, the specific impact of environmental information provided by different data sources on the accuracy of predictions of the spatial distribution of marine fishery resources remains an area that needs to be explored in greater depth. Given this research gap, our study aims to comprehensively assess the impact of different data sources on the predictive accuracy of marine ecological models, thereby providing a more accurate basis for assessing trends in ecological services and their quality.

As an important component of marine ecosystems, pelagic fish such as like *Scomber japonicus* are sensitive to changes in the marine environment (Pennino et al., 2020; Queiros et al., 2019). *Scomber japonicus*, a pelagic migratory fish, is widely distributed in the Pacific Ocean and Atlantic Ocean as well as in the East China Sea and Yellow Sea (Li et al., 2014). In recent years, due to factors such as overfishing, changing hydrological environment, and the deterioration of the prey environment, *S. japonicus* has gradually decreased in abundance, and it has displayed a tendency to move offshore (Cai et al., 2022). This trend underscores the critical need for accurate environmental impact predictions on fishery resources. Given the high sensitivity of such species to environmental changes, variations in the distribution of environmental resources can significantly impact their stability. Therefore, assessing the accuracy of predictions about these fishery resource distributions from different data sources is crucial, aiming to provide more effective protection measures and management strategies for these sensitive species. Consequently, we developed a generalized additive model (GAM) of *S. japonicus* distribution along the coast of southern Zhejiang, China, to determine the influence of the hydrological environment on fishery resources. Separate models were developed based on either remote sensing or field survey data from May (spring) and August (summer) of 2016–2020. The objective of this study is to elucidate the distinctive impact of data source selection on the predictive accuracy of GAMs for *S. japonicus* distribution, offering novel insights into conservation management and the sustainable utilization of fishery resources, thereby addressing a gap in current ecological modeling practices.

2 Materials and methods

2.1 *Data sources*

2.1.1 *Fisheries data*

Data were obtained from fishery-independent surveys of fishery resources conducted in February (winter), May (spring), August (summer), and November (autumn) from 2016 to 2020 in the East China Sea along the coast of southern Zhejiang. Since no *S. japonicus* specimens were caught in autumn or winter, only spring and summer *S. japonicus* survey data were used in this study. The survey area mainly comprised the Yushan and Wentai fishing grounds (Fig. 1). The sites planned are divided using an average grid system, with both latitude and longitude intervals set at 0.25°. Each survey randomly designates flight paths and starting positions (i.e., the positions are fixed, but the order of sampling is random) to complete the sampling task of the entire area in the shortest time possible, ensuring the comparability of data on temporal and spatial scales. Each survey consists of 27 sites. The survey equipment included a bottom trawl approximately 95 m long, 40 m wide, 7.5 m high, with a bottom and floating substrate extending 80 m, and a mesh size of 2 cm. This trawl was towed at a speed ranging from 2 kn (1 kn = 0.514 444 4 m/s) to 4 kn. The operating time at each survey site was approximately 1 h. At each survey site, environmental data were collected simultaneously for water temperature and salinity using a WTW-Multi 3630 water quality analyzer. Water quality samples were collected, measured, and analyzed according to the specification for marine survey (GB/T 12763) (GB/T 12763.6-2007) and the specification for marine monitoring (GB 17378) (GB 17378.3-2007) (General Administration of Quality Supervision, 2007a, b).

2.1.2 *Sources and selection of environmental factors*

The hydrological environment data utilized in this study were gathered concurrently with the field survey of fishery resources. Remote sensing environmental data were obtained from the Copernicus Marine Data Store (https://marine.copernicus.eu), which provides monthly average data with a spatial resolution of 0.25° × 0.25°, consistent with the survey scale. In the remote sensing data, we exclusively utilized water temperature and salinity. The statistical values of the environmental data are listed in Fig. 2.

2.2 *Modeling process*

2.2.1 *Data analysis*

The abundance index (AI) was chosen as a relative indicator

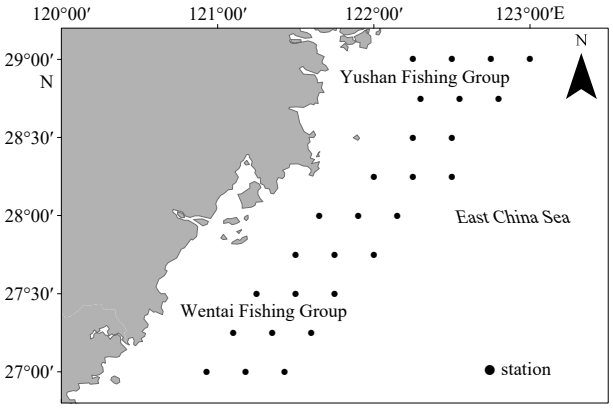

Fig. 1. Distribution of sampling stations in offshore waters of southern Zhejiang.

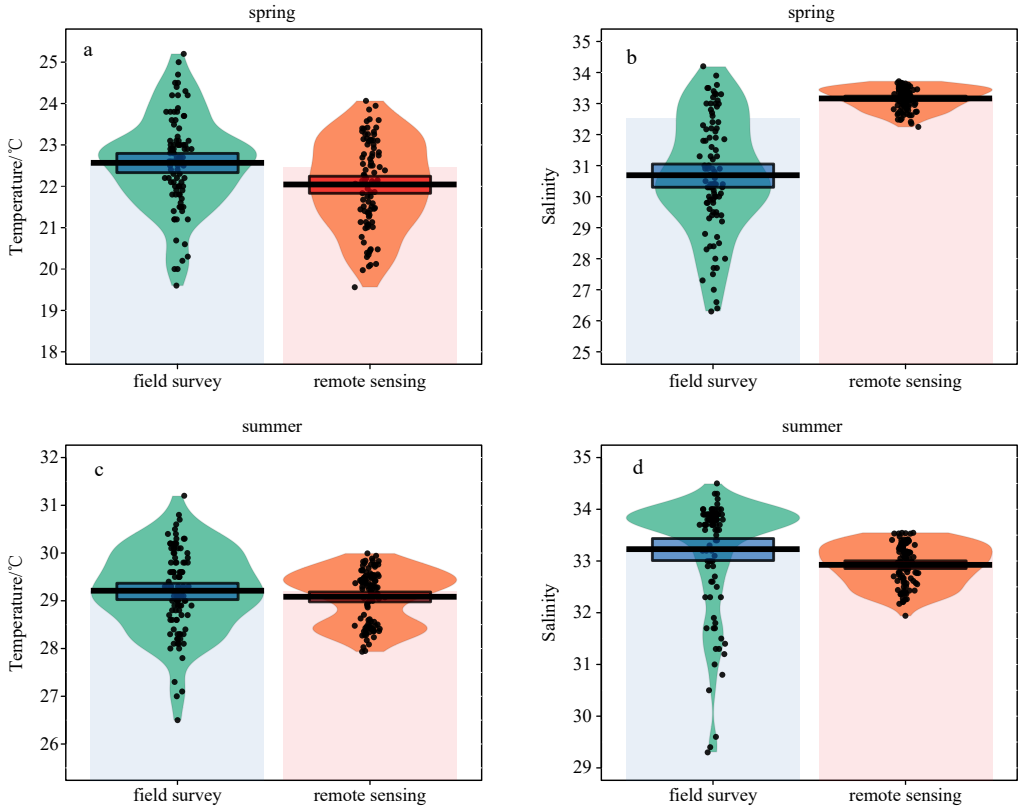

Fig. 2. Hydrological environment characterization with different data sources.

of fishery resource abundance in this study (Liu et al., 2021). The catch data were standardized by trawl time (1 h) and speed (3 kn) according to the total catch and proportion of *S. japonicus* at each station. The specific expression for AI is as follows:

$$\mathrm{AI}_k = \frac{W_k \times 3}{T_k \times S_k}, \tag{1}$$

where AI_k is the *S. japonicus* abundance index at the k-th station (unit in g/h), W_k is the total *S. japonicus* catch at the k-th station (unit in g), T_k is the operating time at the k-th station (unit in h), which is standardized to 1 h, and S_k is the trawl speed, which is standardized to 3 kn.

2.2.2 *Variable selection*

Before building the model, the mutual independence of explanatory variables was tested using a variance inflation factor (VIF). The VIF value is determined using the vif function in the car package of R (V4.0.2). When VIF > 3, there is a serious covariance problem between the environmental factors, and they need to be removed (Sagarese et al., 2014).

2.2.3 *Building the model*

The GAM was built in two main stages. In the first stage of the GAM (GAM1), the probability of occurrence of *S. japonicus* was estimated using a binomial distribution. The second-stage GAM (GAM2) estimates the log-transformed abundance of *S. japonicus* with a Gaussian error structure and an identity link function (Liu et al., 2019; Ma et al., 2022a). Given that *S. japonicus* is a migratory fish, longitude and latitude are important explanatory variables, and their interaction was considered as a fixed variable in the modeling. The two-stage GAM full factorial expressions are as follows:

$$\text{GAM1: Logit}\,(P) = s\,(\text{lat, lon}) + s\,(T) + s\,(S) + \varepsilon, \tag{2}$$

$$\text{GAM2}: g(\ln\,(\text{AI})) = s\,(\text{lat, lon}) + s\,(T) + s\,(S) + \varepsilon, \tag{3}$$

where lat denotes latitude, lon denotes longitude, T denotes sea surface temperature, S denotes salinity, P denotes the probability of occurrence of *S. japonicus*, and ε is the random error, s denotes smooth function.

The results of GAM1 and GAM2 were combined to estimate the final log-transformed *S. japonicus* abundance (Liu et al., 2019) as follows:

$$\ln\,(y) = \ln\,(P) + \ln\,(\text{AI})\,. \tag{4}$$

2.2.4 *Model selection*

In this study, the relationships between the probability of occurrence of *S. japonicus*, abundance, and environmental factors under different data sources were constructed separately using two approaches. The goodness of fit of the model was measured using the Akaike information criterion (AIC), the smaller the AIC value, the better the fit of the model (Akaike, 1998). The two approaches for building the model are as follows.

Establishment of the best models for different seasons and data sources. In this study, the best models were selected by arranging and combining the environmental factors after the covariance test. For convenience of representation, the best models for spring and summer based on field survey and remote sensing data are referred to as $\mathrm{M_{sp,f}}$, $\mathrm{M_{sp,r}}$, $\mathrm{M_{su,f}}$, and $\mathrm{M_{su,r}}$. For the differ-

ent stages of the two-stage GAM, they are referred to as $M_{sp,f,I}$, $M_{sp,f,II}$, $M_{sp,r,I}$, $M_{sp,r,II}$, etc.

2.2.5 *Cross validation*

The accuracy and robustness of the model were evaluated using cross-validation. The data were randomly divided, with 80% allocated as the training set and 20% as the validation set, and the process was repeated 1 000 times. Throughout the cross-validation process, we established a linear regression model between the predicted abundance and observed abundance, calculating the root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (r^2). When the values of RMSE and MAE are closer to 0, and r^2 is closer to 1, it indicates that model has better predictive performance (Liu et al., 2021; Stow et al., 2009; Willmott and Matsuura, 2005).

The linear relationship between the predicted and observed values can be expressed as follows:

$$\ln Y = a + b \times \ln\widehat{y}, \tag{5}$$

where $\widehat{y}$ denotes the predicted value of the model, Y denotes the actual observed value of the model, and a and b denote deviations between the predicted and actual observed values, respectively. When a = 0 and b = 1, the predicted and actual observed densities (i.e., test data) have similar spatial patterns, and the model has good predictive performance (Li et al., 2015).

The equation for calculating the RMSE is as follows (Stow et al., 2009):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{y}_i - Y_i\right)^2}{n}}, \tag{6}$$

where n is the number of observations, Y_i is the i-th observation, and $\widehat{y}_i$ is the i-th predicted value.

The equation for the MAE is Willmott and Matsuura (2005):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y}_i - Y_i\right|. \tag{7}$$

2.2.6 *Predicting spatial distribution of S. japonicus under varying data sources*

To evaluate the accuracy of predicting the spatial distribution of resources under different data sources, this study predicted the resource distribution of *S. japonicus* in spring and summer using environmental data from different data sources and using the optimal model. In this study, the survey area was gridded according to a grid size of 0.1° × 0.1°, and the coordinates of each grid center point were obtained by calculation. The hydrological environment data for each grid center point were extracted using the inverse distance weighting method. The averages of the observed and predicted values of *S. japonicus* in different seasons from 2016 to 2020 were also overlaid, and the living habits of *S. japonicus* were combined to evaluate the prediction effects under different data sources.

In this study, all statistical analyses were performed in R (V3.6.0), the GAM was implemented by the mgcv package, and the spatial distribution of resources and station maps were plotted in Arcmap 10.8.

3 Results

3.1 *Developing predictive models for S. japonicus*

Among the multiple environmental factors selected for this study, the VIF values were all lower than 3. Therefore, the GAM between each environmental factor and abundance under different data sources was established using water temperature and salinity.

By comparing the best models under different data sources, we found differences in the composition and bias explained by the best models for the different seasons (Table 1). In Table 1, $M_{sp,f,I}$ represents the model for Phase Ⅰ using spring field data, and $M_{sp,f,II}$ represents the model for Phase Ⅱ using spring field data. $M_{sp,r,I}$ represents the model for Phase Ⅰ using spring remote sensing data, while $M_{sp,r,II}$ represents the model for Phase Ⅱ using spring remote sensing data. Similarly, $M_{su,f,I}$ represents the model for Phase Ⅰ using summer field data, and $M_{su,f,II}$ represents the model for Phase Ⅱ using summer field data. $M_{su,r,I}$ represents the model for Phase Ⅰ using summer remote sensing data, and $M_{su,r,II}$ represents the model for Phase Ⅱ using summer remote sensing data. Specifically, under different data sources, most models based on field survey data show a higher deviance explained. At the same time, in models based on field survey data, the total contributions of water temperature and salinity are both greater than those in models based on field survey data. Regarding the different stages of the model, the number of predictors in the first stage is usually higher, but compared to the second stage, the deviance explained in the first stage is lower. The range of deviance explained in the first stage is from 18.6% to 25.1%, while in the second stage, it ranges from 29.9% to 64.3% (Table 1).

3.2 *Relationship between abundance of S. japonicus and environmental factors*

Figures 3 and 4 show the relationships between the distribution of *S. japonicus* and temperature and salinity across different

Table 1. Optimal GAM compositions under different data sources

Model	Stage	Optimal model	Deviance explained of each factor/%	Deviance explained/%	AIC
$M_{sp,f}$	$M_{sp,f,I}$	lat, lon	8.8	22.9	114.88
		T	5.1		
		S	9.0		
$M_{sp,f}$	$M_{sp,f,II}$	lat, lon	18.8	43.9	216.10
		S	25.1		
$M_{sp,r}$	$M_{sp,r,I}$	lat, lon	8.8	25.1	116.33
		T	6.3		
		S	10.0		
$M_{sp,r}$	$M_{sp,r,II}$	lat, lon	18.8	29.9	227.33
		T	11.1		
$M_{su,f}$	$M_{su,f,I}$	lat, lon	12.5	25.1	160.65
		T	6.4		
		S	6.2		
$M_{su,f}$	$M_{su,f,II}$	lat, lon	36.1	64.3	193.03
		T	24.4		
		S	3.8		
$M_{su,r}$	$M_{su,r,I}$	lat, lon	12.5	18.6	167.94
		S	6.1		
$M_{su,r}$	$M_{su,r,II}$	lat, lon	36.1	48.2	206.49
		T	0.7		
		S	11.9		

seasons and data sources. In spring (Figs 3a and b), $M_{sp,f,I}$ shows a distinct negative linear relationship between the probability of occurrence of *S. japonicus* and water temperature within the range of 20.5℃ to 25.2℃. Conversely, within the range of 19.5℃ to 24℃, $M_{sp,r,I}$ exhibits that the probability of occurrence of *S. japonicus* initially rises with increasing temperature, then displays a declining trend. In summer, the relationship between probability of occurrence and water temperature forms a dome shape, with the probability of occurrence peaking at 29℃ in the range of 26.5℃ to 33.5℃ (Fig. 3d). Furthermore, in $M_{su,f,II}$, the highest abundance of *S. japonicus* was observed around a water temperature of 29℃ (Fig. 3e). However, in the model based on remote sensing data ($M_{su,r,II}$), a trend of gradually decreasing *S. japonicus* abundance with increasing water temperature was observed, which differs significantly from the results of $M_{su,f,II}$ (Figs 3e and f).

Figure 4 shows the relationship between the distribution of *S. japonicus* and salinity. In spring, $M_{su,f,I}$ displays a trend where the occurrence probability of *S. japonicus* first decreases and then increases with rising salinity (Fig. 4a). However, $M_{sp,r,I}$ shows that within the salinity range of 32.2 to 33.8, the occurrence probability of *S. japonicus* first decreases and then stabilizes with increasing salinity (Fig. 4c). In summer, although the relationship between the occurrence probability of *S. japonicus* and salinity varies across different data sources, a trend of gradually decreasing occurrence probability is observed in high salinity areas (Figs 4d and f). Furthermore, there are also notable differences in the second stage models. $M_{su,f,II}$ shows that within the salinity range of 28 to 34.5, the abundance of *S. japonicus* exhibits first rises and then falls (Fig. 4e). Meanwhile, $M_{su,r,II}$ indicates that within the salinity range of 32.1 to 34.5, the abundance of *S. japonicus* shows a gradual declining trend (Fig. 4g).

3.3 *Model evaluation*

The results of 1 000 cross validations showed that the RMSE and MAE of the field survey data model were smaller than those of the remote sensing data model in the same season. Meanwhile, the r^2 values of the field survey data model were larger than those of the remote sensing data model in both seasons. Therefore, the prediction performance of the field survey data-based model was more accurate than that of the remote sensing data-based model (Table 2).

3.4 *Distribution of S. japonicus under different data sources*

Mapping the resource distribution of *S. japonicus* in spring and summer from 2016 to 2020, by calculating the average values of abundance, revealed large differences in spatial distribution under different data sources in the same season (Fig. 5). In spring, the field survey data model showed that the predicted high values of *S. japonicus* were mainly concentrated in the

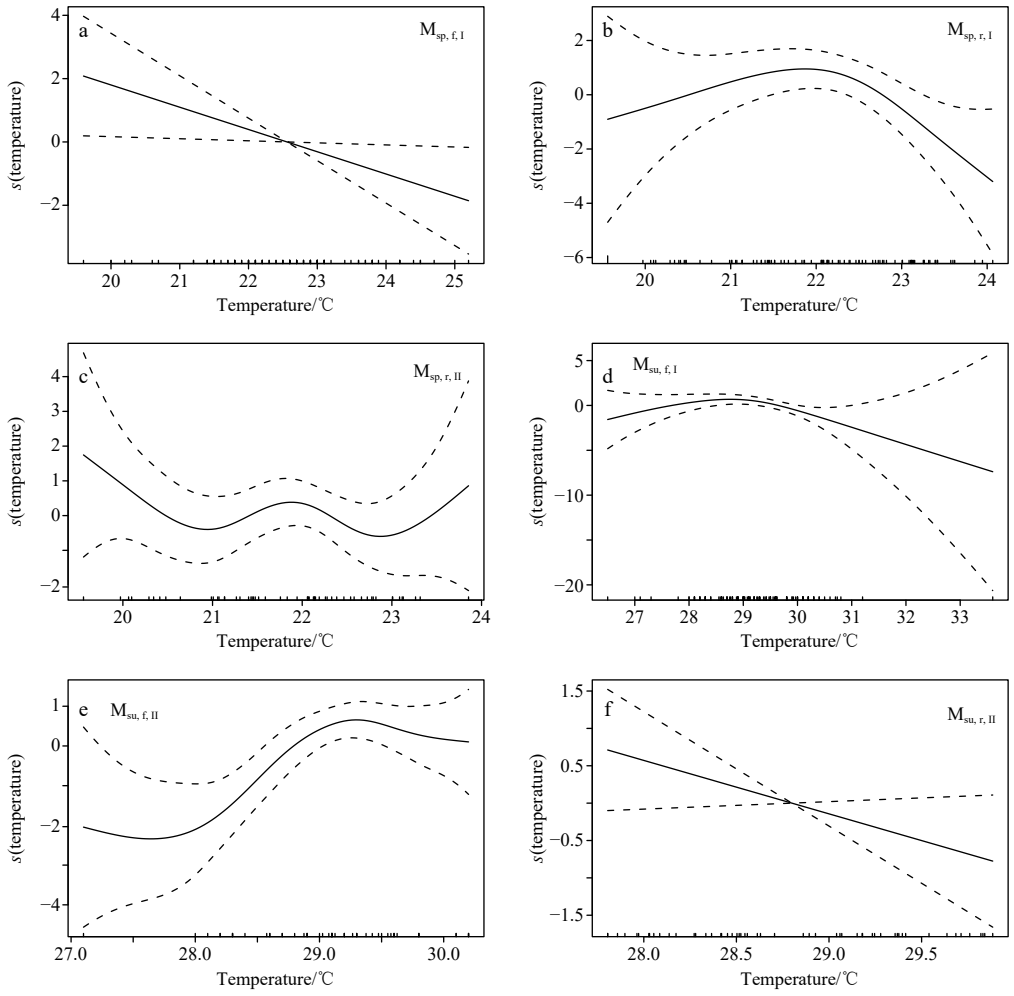

Fig. 3. Relationship between distribution of *Scomber japonicus* and temperature across different seasons and various data sources. $s(\cdot)$ denotes smooth function. The dashed line represents the 95% confidence interval.

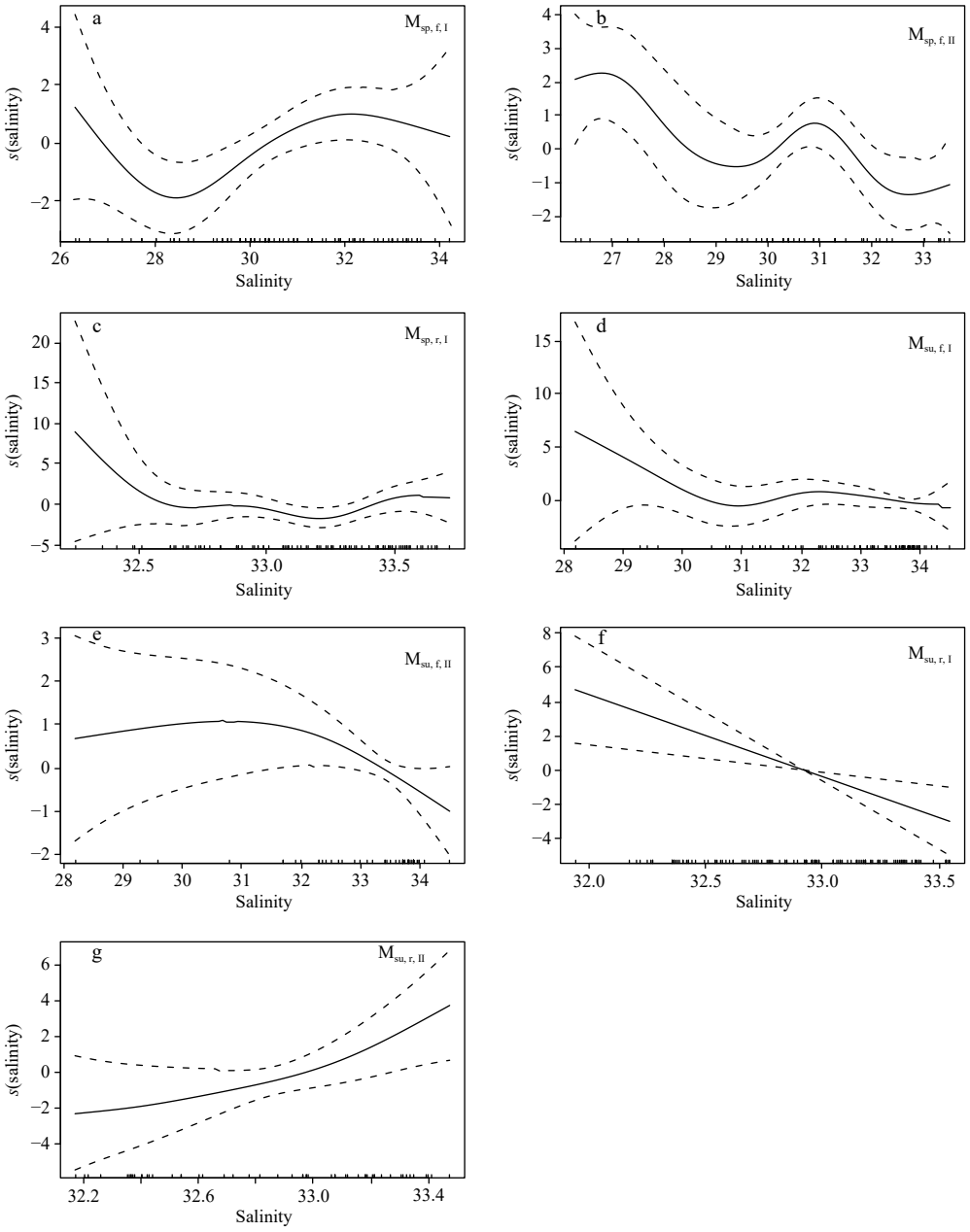

Fig. 4. Relationship between distribution of *Scomber japonicus* and salinity across different seasons and various data sources. $s(\cdot)$ denotes smooth function. The dashed line represents the 95% confidence interval.

Table 2. Results of 1 000 cross-validations of models

Optimal model	r^2	RMSE/(g·h^{-1})	MAE/(g·h^{-1})
$M_{sp,f}$	0.18	3 215.9	1 395.9
$M_{sp,r}$	0.11	3 281.8	1 398.8
$M_{su,f}$	0.17	1 006.7	468.4
$M_{su,r}$	0.07	1 017.1	484.8

northern part of the study area and in nearshore waters. The model based on remote sensing data exhibits a similar pattern.

In summer, the field survey data model showed that *S. japonicus* was mainly concentrated on the offshore side, and the overall characteristics showed that the abundance was greater in the seaward waters than in the inshore waters. Compared to models based on field survey data, the spatial distribution predictions based on remote sensing data differ, with a smaller range of high-abundance water areas (Fig. 5).

In addition, by comparing the opatimal model under two different data sources with the actual distribution of the resource characteristics, the accuracy of the field survey data-based model was higher (Fig. 5).

4 Discussion

4.1 *Comparison of model prediction performance*

As understanding of the uncertainty of species distribution models and their widespread use in predictions increases, there is a need to enhance the understanding and application of model evaluations to ensure the accuracy and stability of model predictions (Luan et al., 2021). The predictive performance of models highly depends on the quantity and quality of modeling to ensure accuracy and reliability (Guillera-Arroita et al., 2015; La Marca et al., 2019). Therefore, it is necessary to compare the ac-

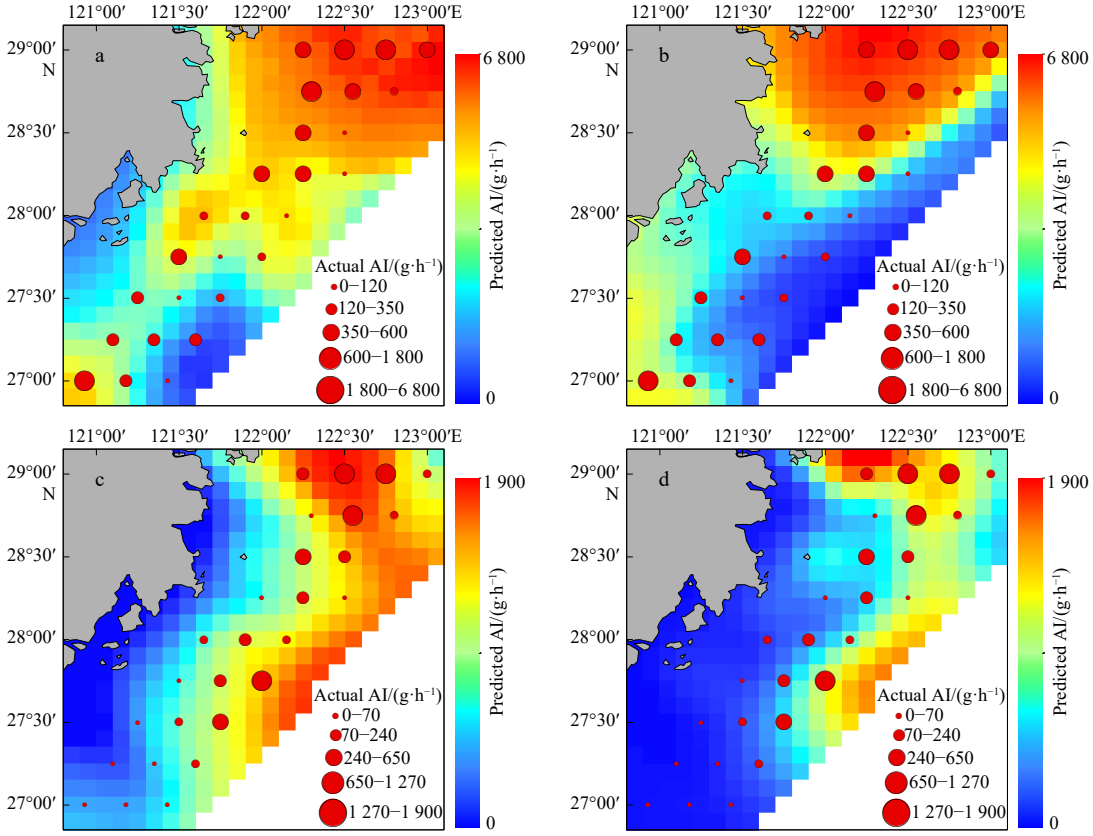

Fig. 5. Overlay of actual and predicted abundance of *Scomber japonicus* under the optimal models. The specific categorizations are as follows. a. Spring field data models, b. spring remote sensing data models, c. summer field data models, and d. summer remote sensing data models.

curacy and predictive performance of models under different data sources to fully understand the impact of different source data on spatial predictions of resources. In this study, two different data sources (field survey data and remote sensing data) are applied to predict the distribution of *S. japonicus* off the southern coast of Zhejiang. The fitting effects and predictiveness of the two different data sources are evaluated using deviance explained, cross-validation methods (RMSE, MAE, and r^2). Results indicate that the model based on field survey data is more accurate than the one based on remote sensing data. These results can be used for the conservation and spatial management of the region and its species. For example, the southern offshore areas of Zhejiang are breeding grounds for many important fishery species, but due to human activities and other factors, there is a significant decline in resources. The local government has formulated policies to manage them, aiming to restore the condition of these waters (Han et al., 2024). The findings and techniques of this study can help determine the current state of areas to be managed, thereby enabling the formulation of more effective conservation measures.

This difference may stem from discrepancies in accuracy between different data sources. Field survey data are typically collected during specific time frames, providing accurate environmental conditions for specific moments and locations, which align more closely with resource survey data. In contrast, remote sensing data span a wider time range, and this temporal mismatch might reduce its effectiveness (La Marca et al., 2019). Furthermore, although significant advancements in remote sensing technology have made it possible to provide a more comprehensive and extended series of environmental variables, the potential for reduced data quality due to cloud cover, atmospheric conditions, as well as technological limits and the spatial resolution of the data, might impact the effectiveness of its application (Lei et al., 2022).

A few studies have compared the impact of field survey data and remote sensing data on model predictive performance, and their conclusions are consistent with our findings. For instance, Johnston et al. (2017), in predicting the biodiversity of freshwater fish and benthic invertebrate communities in streams, found that the model based on field survey data had significantly better predictive performance than the model based on remote sensing data. Similarly, La Marca et al. (2019), in examining the effects of data source and species distribution model (SDM) methodologies on the predictive performance of small mammal distribution and their implication in determining spatial conservation priorities, also found that models based on field surveys had better predictiveness.

In summary, our study findings suggest that although remote sensing data play a crucial role in environmental monitoring, in some cases, field survey data, due to its higher accuracy and specificity in time, can more precisely reflect the variability of hydrological environments, thereby showing better fitting results in model predictions.

4.2 *Data source effects on predicting S. japonicus habitat preferences*

This study compared spatial distribution prediction models of *S. japonicus* constructed based on different data sources, reveal-

ing significant differences in prediction outcomes (Fig. 5). According to previous research, from March to May each year, sexually mature *S. japonicus* migrate to the nutrition-rich and topographically complex coastal waters of southern Zhejiang to spawn, influenced by the strengthening of warm currents and rising water temperatures, while individuals with immature gonads continue to migrate northward until they mature and return to the spawning area (Deng et al., 1991; Chen et al., 2017). Based on this, we hypothesized that the abundance of *S. japonicus* in the coastal waters of southern Zhejiang and the northern regions would be significantly higher in spring than in other waters. However, comparing the prediction results of different data sources, although the models from both data sources are relatively consistent in the overall distribution pattern and align with the habitual distribution of *S. japonicus*, significant differences exist at a finer scale. Particularly in spring, the high-abundance areas predicted by the model based on remote sensing data are noticeably smaller in range than those predicted by the model based on field survey data, and are more inclined towards coastal waters, which diverges significantly from actual observational characteristics (Figs 5a and b). Accurately understanding these discrepancies in predictions is crucial for formulating differentiated management strategies and adaptive measures, as fisheries managers rely on accurate habitat distribution information to identify key habitats for species (Zhao et al., 2014; Zhang et al., 2020).

In summer, the model based on field survey data predicts that *S. japonicus* mainly concentrates in offshore waters far from the coast (Fig. 5c), which is consistent with field survey data and related research findings. Previous studies have indicated that as the size of *S. japonicus* increases and their foraging ability improves, they gradually migrate to waters farther from the shore for feeding (Zhang et al., 2017). However, although the model constructed with remote sensing data can also generally display this characteristic, its predictive capability is noticeably insufficient (Fig. 5d). Therefore, choosing remote sensing data or field survey data as the data source for the model could have significantly different impacts on the formulation of resource conservation and fishery management strategies.

4.3 *Impact of different data sources on the relationship between S. japonicus distribution and environmental factors*

The distribution and habits of fish are closely related to human activities and environmental changes. Accurately exploring the relationship between the distribution of fishery resources and environmental factors, as well as predicting their spatial distribution, not only helps to understand the response mechanisms of fishery resource populations to environmental factors but also has significant importance for the assessment and formulation of fishery resources management strategies. Although many environmental factors affect the population of *S. japonicus* (Li and Chen, 2009; Yu et al., 2018), the impact of these factors on population dynamics varies. Among the many environmental parameters, temperature and salinity are considered very important variables for predicting *S. japonicus* fishing grounds (Li et al., 2014; Yu et al., 2018). This study indicates that, except for the spring phase one models ($M_{sp,f,I}$ and $M_{sp,r,I}$), the models using remote sensing data generally have lower total contributions from water temperature and salinity compared to those based on field survey data (Table 1). This indicates that in the remote sensing data models, water temperature and salinity, which are crucial environmental factors for the distribution of *S. japonicus* resources, were not adequately reflected, potentially affecting the accuracy of model predictions to some extent. This difference might be explained by discrepancies between field survey data and remote sensing data. Specifically, the range of water temperature and salinity shown in field survey data was significantly larger than that in remote sensing data (Fig. 2), which more accurately reflects the environmental variability between different survey stations. Therefore, models constructed based on field survey data are more likely to accurately capture the relationship between *S. japonicus* distribution and environmental factors, thoroughly analyze the specific impact of each environmental factor on the distribution of *S. japonicus*, and thus make more accurate predictions.

Water temperature is one of the most frequently measured parameters in the marine environment, influencing not only the metabolic rate and physiological state of fish but also affecting the structure of the entire food chain by regulating primary productivity (Selleslagh and Amara, 2008). In this study, we found significant differences in the relationship between water temperature and environmental factors even within the same season and stage when analyzing with different data sources. Specifically, in the data analysis of $M_{sp,f,I}$, there was a negative correlation between the occurrence probability of *S. japonicus* and water temperature. Whereas in the data analysis of $M_{sp,r,I}$, a dome-shaped relationship was shown with the highest occurrence probability at 22℃. By comparing with other studies on *S. japonicus* in the East China Sea, we found some study results are consistent with the conclusions of $M_{sp,f,I}$, but significantly different from the findings of $M_{sp,r,I}$. For example, studies by Cui and Chen (2007) showed a significant negative correlation between the relative resource amount of *S. japonicus* in the East China Sea and sea surface water temperature during April to May. Hiyama et al. (2002) also found that the spawning success rate of *S. japonicus* negatively correlated with water temperature in the East China Sea. These findings suggest that the relationship between the occurrence probability of *S. japonicus* and water temperature analyzed based on field survey data might be more accurate in spring. Similarly, in summer, our study also found that most conclusions are consistent with the results of the model based on field survey data ($M_{su,f,II}$) compared to other studies in the East China Sea (Chen et al., 2009; Li and Chen, 2009; Li et al., 2014), and significantly different from the results of $M_{su,r,II}$. For instance, Chen et al. (2009), when exploring suitable habitats for *S. japonicus* in the East China Sea during summer using the Habitat Suitability Index model, found that over 90% of *S. japonicus* catch came from areas with water temperatures of 28.0–29.4℃. Therefore, we believe that compared to remote sensing data, field survey data can provide more accurate results when exploring the relationship between *S. japonicus* distribution and water temperature.

Salinity is closely related to the timing of the onset and the growth rate of various developmental stages of fish, playing a very important role in the distribution of fishery resources (Liu et al., 2019; Zhang et al., 2021). This study revealed that within the same season, analyses based on different data sources showed that the occurrence probability and abundance of *S. japonicus* might exhibit different relationships with salinity (Fig. 4). For example, we noticed significant differences in salinity changes during the first phase of spring ($M_{sp,f,I}$ and $M_{sp,r,I}$) and the second phase of summer ($M_{su,f,II}$ and $M_{su,r,II}$), sometimes even showing opposite trends. We believe this phenomenon is mainly caused by differences in the range of salinity changes across different data sources. In models based on field survey data, the wider range of salinity variation allows for a more precise fitting of the relation-

ship between salinity and *S. japonicus* distribution. However, in models based on satellite remote sensing data, changes in salinity are not very pronounced even when resource levels vary, thereby affecting the model's fitting effectiveness. Moreover, Chen et al. (2009) found that the suitable salinity for *S. japonicus* was 33.6 to 34.2, which differs from the conclusions of this study. This discrepancy may stem from the approach of this study to build models in stages, which, compared to studying the overall relative abundance of resources, highlights the impact of salinity on the distribution of *S. japonicus* resources, especially when the data contains a large number of zero values.

Furthermore, incorporating latitude and longitude as fixed explanatory variables in the model is crucial because these spatial data not only influence environmental factors such as water temperature and salinity but also indicate specific geographic areas where *S. japonicus* may congregate (Yasuda et al., 2014). Including the interaction of latitude and longitude helps us more accurately depict the species' response to different geographical environments and make more precise predictions about its distribution. Therefore, considering latitude and longitude is essential for understanding the distribution patterns of *S. japonicus* across various data sources and how these patterns are influenced by specific geographic locations.

We acknowledge that although this study utilizes two-stage GAM and conducts a comparative analysis of model prediction effects under different data sources by cross-validation and spatial prediction distribution, its prediction performance is not very good. In practice, some biotic and abiotic factors (e.g., dissolved oxygen, climate change, predators), in addition to the environmental factors used in this study, can also drive the spatial distribution and habitat selection of fishery resources (Dai et al., 2020; Liu et al., 2019). Therefore, more biotic and abiotic factors will be incorporated in future studies, as well as more models will be tried, which may help the fitting effect and prediction performance of the prediction models to further accurately predict their distribution characteristics, thus providing more support for the conservation and sustainable use of fishery resources.

Acknowledgements

We would like to thank to the teachers and students from the Laboratory of Quantitative Fisheries Stock & Ecosystem Assessment and Management, Shanghai Ocean University.

References

Akaike H. 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, eds. Selected Papers of Hirotugu Akaike. New York, NY, USA: Springer, 199–213

Cai Kai, Kindong R, Ma Qiuyun, et al. 2022. Growth heterogeneity of Chub mackerel (*Scomber japonicus*) in the Northwest Pacific Ocean. Journal of Marine Science and Engineering, 10(2): 301, doi: 10.3390/jmse10020301

Chen Xinjun, Li Gang, Feng Bo, et al. 2009. Habitat suitability index of Chub mackerel (*Scomber japonicus*) from July to September in the East China Sea. Journal of Oceanography, 65(1): 93–102, doi: 10.1007/s10872-009-0009-9

Chen Weifeng, Peng Xin, Wang Zhenhua, et al. 2017. Community structure characteristics of fishes in the coastal area of south Zhejiang during autumn and winter. Ocean Development and Management (in Chinese), 34(11): 111–119

Cui Ke, Chen Xinjun. 2007. Study of the relationships between SST and mackerel abundances in the Yellow and East China Seas. South China Fisheries Science (in Chinese), 3(4): 20–25

Dai Libin, Hodgdon C, Tian Siquan, et al. 2020. Comparative performance of modelling approaches for predicting fish species richness in the Yangtze River Estuary. Regional Studies in Marine Science, 35: 101161, doi: 10.1016/j.rsma.2020.101161

Deng Jingyao, Zhao Chuanyan. 1991. Marine Fisheries Biology (in Chinese). Beijing: Agriculture Publishing Press

General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. 2007a. GB/T 12763.6-2007 Specifications for Oceanographic Survey-Part 6: Marine Biological Survey (in Chinese). Beijing: China Standards Press

General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. 2007b. GB 17378.3-2007 The Specification for Marine Monitoring-Part 3: Sample Collection, Storage and Transportation (in Chinese). Beijing: China Standards Press

Guillera-Arroita G, Lahoz-Monfort J J, Elith J, et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography, 24(3): 276–292, doi: 10.1111/geb.12268

Guisan A, Zimmermann N E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135(2/3): 147–186, doi: 10.1016/S0304-3800(00)00354-9

Han Dongyan, Kindong R, Wang Wen, et al. 2024. Effects of jellyfish and black seabream releasing on marine ecosystems: A mass balance approach for the coastal area of southern Zhejiang, China. Ocean & Coastal Management, 248: 106948, doi: 10.1016/j.ocecoaman.2023.106948

Hiyama Y, Yoda M, Ohshimo S. 2002. Stock size fluctuations in chub mackerel (*Scomber japonicus*) in the East China Sea and the Japan/East Sea. Fisheries Oceanography, 11(6): 347–353, doi: 10.1046/j.1365-2419.2002.00217.x

Johnston M R, Elmore A J, Mokany K, et al. 2017. Field-measured variables outperform derived alternatives in Maryland stream biodiversity models. Diversity and Distributions, 23(9): 1054–1066, doi: 10.1111/ddi.12598

La Marca W, Elith J, Firth R S C, et al. 2019. The influence of data source and species distribution modelling method on spatial conservation priorities. Diversity and Distributions, 25(7): 1060–1073, doi: 10.1111/ddi.12924

Lei Lin, Wang Jintao, Chen Xinjun. 2022. Influence of environmental data of different sources on marine species habitat modeling: A case study for *Ommastrephes bartramii* in the Northwest Pacific Ocean. Acta Oceanologica Sinica, 41(1): 76–83, doi: 10.1007/s13131-021-1896-x

Li Gang, Chen Xinjun. 2009. Study on the relationship between catch of mackerel and environmental factors in the East China Sea in summer. Journal of Marine Sciences, 27(1): 1–8.

Li Gang, Chen Xinjun, Lei Lin, et al. 2014. Distribution of hotspots of Chub mackerel based on remote-sensing data in coastal waters of China. International Journal of Remote Sensing, 35(11/12): 4399–4421, doi: 10.1080/01431161.2014.916057

Li Bai, Cao Jie, Chang J H, et al. 2015. Evaluation of effectiveness of fixed-station sampling for monitoring American lobster settlement. North American Journal of Fisheries Management, 35(5): 942–957, doi: 10.1080/02755947.2015.1074961

Liu Xiaoxue, Gao Chunxia, Zhao Jing, et al. 2021. Modeling and comparison of count data containing zero values: a case study of *Setipinna taty* in the south inshore of Zhejiang, China. Environmental Science and Pollution Research, 28(34): 46827–46837, doi: 10.1007/s11356-021-13440-5

Liu Xiaoxiao, Wang Jing, Zhang Yunlei, et al. 2019. Comparison between two GAMs in quantifying the spatial distribution of *Hexagrammos otakii* in Haizhou Bay, China. Fisheries Research, 218: 209–217, doi: 10.1016/j.fishres.2019.05.019

Luan Jing, Zhang Chongliang, Ji Yupeng, et al. 2021. Matching data types to the objectives of species distribution modeling: An evaluation with marine fish species. Frontiers in Marine Science, 8: 771071, doi: 10.3389/fmars.2021.771071

Luan Jing, Zhang Chongliang, Xu Binduo, et al. 2018. Modelling the spatial distribution of three *Portunidae* crabs in Haizhou Bay,

China. PLoS One, 13(11): e0207457, doi: 10.1371/journal.pone.0207457

Ma Wen, Gao Chunxia, Qin Song, et al. 2022a. Do two different approaches to the season in modeling affect the predicted distribution of fish? A case study for *Decapterus maruadsi* in the offshore waters of southern Zhejiang, China. Fishes, 7(4): 153, doi: 10.3390/fishes7040153

Ma Wen, Gao Chunxia, Tang Wei, et al. 2022b. Relationship between *Engraulis japonicus* resources and environmental factors based on multi-model comparison in offshore waters of Southern Zhejiang, China. Journal of Marine Science and Engineering, 10(5): 657, doi: 10.3390/jmse10050657

Ma Jin, Li Bai, Zhao Jing, et al. 2020. Environmental influences on the spatio-temporal distribution of *Coilia nasus* in the Yangtze River estuary. Journal of Applied Ichthyology, 36(3): 315–325, doi: 10.1111/jai.14028

Pan Shaoyuan, Tian Siquan, Wang Xuefang, et al. 2021. Comparing different spatial interpolation methods to predict the distribution of fishes: A case study of *Coilia nasus* in the Changjiang River Estuary. Acta Oceanologica Sinica, 40(8): 119–132, doi: 10.1007/s13131-021-1789-z

Pennino M G, Coll M, Albo-Puigserver M, et al. 2020. Current and future influence of environmental factors on small pelagic fish distributions in the Northwestern Mediterranean Sea. Frontiers in Marine Science, 7: 622, doi: 10.3389/fmars.2020.00622

Queiros Q, Fromentin J M, Gasset E, et al. 2019. Food in the sea: size also matters for pelagic fish. Frontiers in Marine Science, 6: 385, doi: 10.3389/fmars.2019.00385

Sagarese S R, Frisk M G, Cerrato R M, et al. 2014. Application of generalized additive models to examine ontogenetic and seasonal distributions of spiny dogfish (*Squalus acanthias*) in the Northeast (US) shelf large marine ecosystem. Canadian Journal of Fisheries and Aquatic Sciences, 71(6): 847–877, doi: 10.1139/cjfas-2013-0342

Scales K L, Hazen E L, Jacox M G, et al. 2017. Scale of inference: on the sensitivity of habitat models for wide-ranging marine predators to the resolution of environmental data. Ecography, 40(1): 210–220, doi: 10.1111/ecog.02272

Selleslagh J, Amara R. 2008. Environmental factors structuring fish composition and assemblages in a small macrotidal estuary (eastern English Channel). Estuarine, Coastal and Shelf Science, 79(3): 507–517, doi: 10.1016/j.ecss.2008.05.006

Stock A, Subramaniam A. 2020. Accuracy of empirical satellite algorithms for mapping phytoplankton diagnostic pigments in the open ocean: a supervised learning perspective. Frontiers in Marine Science, 7: 599, doi: 10.3389/fmars.2020.00599

Stow C A, Jolliff J, McGillicuddy Jr D J, et al. 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, 76(1/2): 4–15, doi: 10.1016/j.jmarsys.2008.03.011

Wang Junbang, Ding Yuefan, Wang Shaoqiang, et al. 2022. Pixel-scale historical-baseline-based ecological quality: Measuring impacts from climate change and human activities from 2000 to 2018 in China. Journal of Environmental Management, 313: 114944, doi: 10.1016/j.jenvman.2022.114944

Welch H, Brodie S, Jacox M G, et al. 2020. Considerations for transferring an operational dynamic ocean management tool between ocean color products. Remote Sensing of Environment, 242: 111753, doi: 10.1016/j.rse.2020.111753

Willmott C J, Matsuura K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1): 79–82, doi: 10.3354/cr030079

Wu Xiaojing, He Honglin, Zhang Li, et al. 2022. Spatial sampling design optimization of monitoring network for terrestrial ecosystem in China. Science of the Total Environment, 847: 157397, doi: 10.1016/j.scitotenv.2022.157397

Xue Ying, Tanaka K, Yu Huaming, et al. 2018. Using a new framework of two-phase generalized additive models to incorporate prey abundance in spatial distribution models of juvenile slender lizardfish in Haizhou Bay, China. Marine Biology Research, 14(5): 508–523, doi: 10.1080/17451000.2018.1447673

Yasuda T, Yukami R, Ohshimo S. 2014. Fishing ground hotspots reveal long-term variation in chub mackerel *Scomber japonicus* habitat in the East China Sea. Marine Ecology Progress Series, 501: 239–250, doi: 10.3354/meps10679

Yu Hao, Cooper A R, Infante D M. 2020. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. Ecological Modelling, 432: 109202, doi: 10.1016/j.ecolmodel.2020.109202

Yu Wei, Guo Ai, Zhang Yang, et al. 2018. Climate-induced habitat suitability variations of chub mackerel *Scomber japonicus* in the East China Sea. Fisheries Research, 207: 63–73, doi: 10.1016/j.fishres.2018.06.007

Zhang Qiuhua, Cheng Jiahua, Xu Hanxiang, et al. 2017. Fishery Resources and Sustainable Utilization in the East China Sea (in Chinese). Shanghai: Fudan University Press, 212–219

Zhang Yunlei, Xue Ying, Xu Binduo, et al. 2021. Evaluating the effect of input variables on quantifying the spatial distribution of croaker *Johnius belangerii* in Haizhou Bay, China. Journal of Oceanology and Limnology, 39(4): 1570–1583, doi: 10.1007/s00343-020-0193-4

Zhang Yunlei, Yu Huaming, Yu Haiqing, et al. 2020. Optimization of environmental variables in habitat suitability modeling for mantis shrimp *Oratosquilla oratoria* in the Haizhou Bay and adjacent waters. Acta Oceanologica Sinica, 39(6): 36–47, doi: 10.1007/s13131-020-1546-8

Zhao Jing, Cao Jie, Tian Siquan, et al. 2014. A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices. Aquatic Ecology, 48(3): 297–312, doi: 10.1007/s10452-014-9484-1