· Article · Advances in Polar Science
doi:10.12429/j.advps.2025.0028 December 2025 Vol. 36 No. 4: 373-391

The distribution modeling and analysis of Antarctic krill: impacts of algorithm and spatial resolution

LI Wenxiong[1], YING Yiping[1*], ZHANG Jichang[1,2], ZHAO Yunxia[1,2], ZHU Jiancheng[1], FAN Gangzhou[1], MU Xiuxia[1] & WANG Xinliang[1,2]

[1] Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences. Key Lab of Sustainable Development of Polar Fisheries, Ministry of Agriculture and Rural Affairs. Qingdao 266071, China;
[2] Qingdao Marine Science and Technology Center, Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao 266071, China

Received 18 September 2025; accepted 10 December 2025; published online 30 December 2025

Abstract Antarctic krill (*Euphausia superba*), widely distributes around Antarctica, is a key species supporting the biodiversity of the Southern Ocean ecosystem. The Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) has thus managed the krill fishery according to a precautionary way. Currently, CCAMLR is making effort to develop a refined krill fishery management approach based on more solid science, which requires accurate predictions of krill distribution. To address this need, this study investigated the effects of algorithm and spatial resolution on the performance of Antarctic krill distribution modelling. We integrated acoustic data from 4 surveys conducted in the waters adjacent to the Antarctic Peninsula with 11 environmental variables characterizing krill prey conditions, water mass properties, and seafloor topography. These data were processed at 4 spatial resolutions (5, 10, 15, and 20 km) to fit distribution models using 4 algorithms: Random Forests (RF), Generalized Additive Models (GAM), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). Model performance was assessed and compared in terms of goodness-of-fit and predictive accuracy. The results showed that RF achieved the highest predictive performance at most resolutions, whereas GAM performed best at the coarsest resolution (20 km). XGBoost closely following RF in accuracy and demonstrated robustness as evidenced by the highly consistent partial dependence curves across resolutions. In contrast, ANN exhibited limitations with smaller sample sizes, resulting in comparatively poorer predictive performance. The analysis revealed a trade-off whereby reducing spatial resolution improved model fit and mitigated zero-inflation at the expense of fine-scale information and overall predictive accuracy. Ensemble models, integrating RF, GAM, and XGBoost, are proposed as potential balanced solutions to improve predictive stability, offering a more robust scientific basis for the refinement of krill management.

Keywords **Antarctic krill, species distribution model, algorithm selection, spatial resolution, machine learning**

Citation: Li W X, Ying Y P, Zhang J C, et al. The distribution modeling and analysis of Antarctic krill: impacts of algorithm and spatial resolution. Adv Polar Sci, 2025, 36(4): 373-391, doi: 10.12429/j.advps.2025.0028

1 Introduction

The Antarctic krill (*Euphausia superba* Dana, 1850), hereafter referred to as krill, is a keystone species in the Southern Ocean ecosystem. Serving as the primary food source for numerous predators such as whales, seals, penguins, and seabirds, krill not only plays a critical role in maintaining the biodiversity of the Antarctic marine ecosystem but also contributes significantly to the Southern Ocean's biogeochemical cycles and the global carbon cycle through vertical migration and the effect of its life history on the biological pump (Cavan et al., 2019; Murphy et al.,

* Corresponding author. E-mail: yingyp@ysfri.ac.cn

2012; Schmidt et al., 2016). Antarctic krill has an immense biomass, making it one of the most important fishery species in the Southern Ocean (Cavanagh et al., 2021).

Given the dual importance of krill for both the Antarctic marine ecosystem and fisheries, the management of the krill fishery received widespread attention. The Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) has long been guided by an ecosystem-based, precautionary, and science-based approach to manage the krill and other fisheries in its Convention Area (CCAMLR, 2023). In order to better translate scientific assessments into management and closer alignment with CCAMLR objectives, since 2019, the Scientific Committee for the Conservation of Antarctic Marine Living Resources (SC-CAMLR) has actively advanced the development of a revised krill fishery management approach (KFMA), including the science-based revision of precautionary catch limits. The revised KFMA comprises 3 parallel components of scientific work: (1) acoustic survey-based krill biomass estimation, (2) precautionary harvest rate assessment based on population dynamics, and (3) a spatiotemporal catch limit allocation approach based on krill-predator spatial overlap analysis (SOA) (CCAMLR, 2022). The SOA, in particular, requires high-quality quantitative data on the spatial distribution of krill density.

A common approach for constructing krill spatial distributions involves coupling relationships between krill data (presence, abundance or biomass) and environmental variables to develop Species Distribution Models (SDMs), which can be used to predict krill distribution based on the environmental factors (Chen and Zhu, 2022; Chen et al., 2025; Freer et al., 2025; Lin et al., 2022; Warwick-Evans et al, 2022). The main environmental factors influencing Antarctic krill distribution, considered in these studies, can be categorized into abiotic factors related to physical and chemical oceanographic characteristics (e.g., bathymetry, sea temperature, sea ice, currents and salinity) and biological factors such as phytoplankton abundance (Siegel, 2005; Warwick-Evans et al., 2022; Zhao, 2025a).

Antarctic krill exhibits temperature sensitivity, tolerating moderate short-term fluctuations but suffering metabolic and physiological constraints under larger thermal variations, ultimately altering its distribution and behavior (Brown et al., 2010; Ross et al., 1988). Sea ice provides critical refuge and food during key life stages and is therefore also generally considered to impact the krill distribution (Ryabov et al., 2017). As krill mostly feeding on micro and small organisms, especially phytoplankton, the phytoplankton concentration and distribution were often viewed as key factors influencing krill abundance and distribution (Dong et al., 2024; Marrari et al., 2008). The ocean current system of the Southern Ocean plays a crucial role in transport and flux of krill and is thus regarded as an important influence on krill distribution (Fach et al., 2006; Murphy et al., 2004). The complex water mass structure, closely related to bathymetry, and the temperature-salinity characteristics of krill habitats are also considered potential influencing factors on krill distribution (Huneke et al., 2016; Li et al., 2019).

In recent years, Species Distribution Models (SDMs) have been widely applied to marine species, employing a diverse range of algorithms (Guillaumot et al., 2021; Hunt et al., 2020; Robinson et al., 2017; Sinclair et al., 2010). Among these, the Generalized Additive Model (GAM), valued for its non-linearity and strong interpretability, is one of the most commonly used models in fisheries science (Wood, 2017). This model has been introduced into Antarctic krill research, enhancing the ability to predict krill resource distribution (Freer et al., 2025; Warwick-Evans et al., 2022). Machine learning models such as Random Forests (RF), Artificial Neural Networks (ANN), and Extreme Gradient Boosting (XGBoost) often demonstrate superior performance in handling complex non-linear relationships (Breiman, 2001; Chen and Guestrin, 2016; Maravelias et al., 2003), but their application in krill distribution studies remains a relatively emerging area of research.

Developing of krill SDMs generally relies on data from net and acoustic surveys (Atkinson et al., 2017). However, the harsh polar meteorological conditions make obtaining such data extremely costly. Consequently, survey data for krill distribution with both broad spatial coverage and high resolution are scarce. More frequently, krill distribution data are obtained through survey designs that balance these two competing factors. For SDMs, coarse spatial sampling intervals may overlook fine-scale patterns in species distributions, whereas insufficient spatial coverage can hinder the capture of broad ecological trends. Given the significant impact of spatial scale and resolution on ecological modeling (Liu and Zhu, 2022; Wiens, 1989), appropriate resolutions must be selected for developing krill SDMs.

The waters surrounding the South Shetland Islands constitute important Antarctic krill habitat and a hotspot for krill fishing activity (Zhao et al., 2025b). This area falls within the core candidate management units of the first phase (Subarea 48.1) of CCAMLR's revised KFMA (CCAMLR, 2024). Water masses from adjacent areas, like the Bellingshausen Sea and Weddell Sea, converge here, creating a complex physicochemical environment (Huneke et al., 2016), making it a suitable study area for analyzing the relationship between environmental factors and krill distribution.

The present study utilized acoustic krill assessment data from Chinese fishing vessel-based surveys (2013–2018), complemented by ocean remote sensing and reanalysis data from the Copernicus Marine Service, to develop krill SDMs with multiple algorithms and spatial-resolution. This study systematically compared the model-fit and predictive performance of 4 algorithms (GAM, RF, ANN, and XGBoost) across 4 spatial resolutions for the krill distribution models, and evaluated the effects of these options on the model robustness. The findings provide methodological insight toward improving

the establishment of krill distribution layer for the krill-predator SOA, thereby supporting the further development of the ecosystem-based management for the krill fishery.

2 Materials and methods

The workflow of the modeling methodology for Antarctic krill distribution of this study is illustrated in Figure 1, with detailed description provided in the subsequent sections.

2.1 Data sources

The acoustic krill biomass assessment data used in this study were derived from a series of surveys conducted aboard the Chinese krill fishing vessel *Fu Rong Hai* between 2013 and 2018, as fully described in Wang et al. (2025). These surveys took place in December 2013, March 2015, January 2016, and February 2018 in the waters surrounding the South Shetland Islands (Figure 2). The survey design followed 13 planned acoustic transects applied in the US AMLR survey in the western and southern areas of the South Shetland Islands (Reiss et al., 2008).

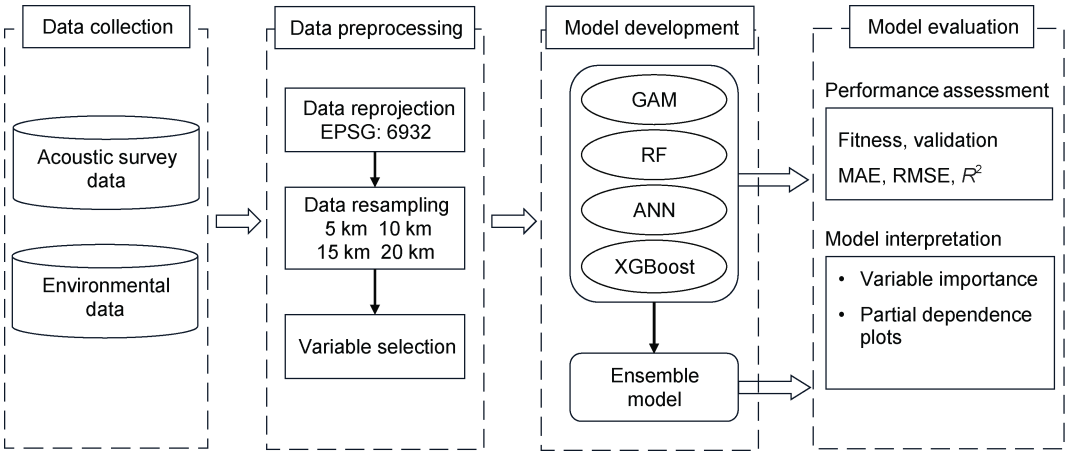

Figure 1 Flowchart of the Antarctic krill distribution modeling methodology.

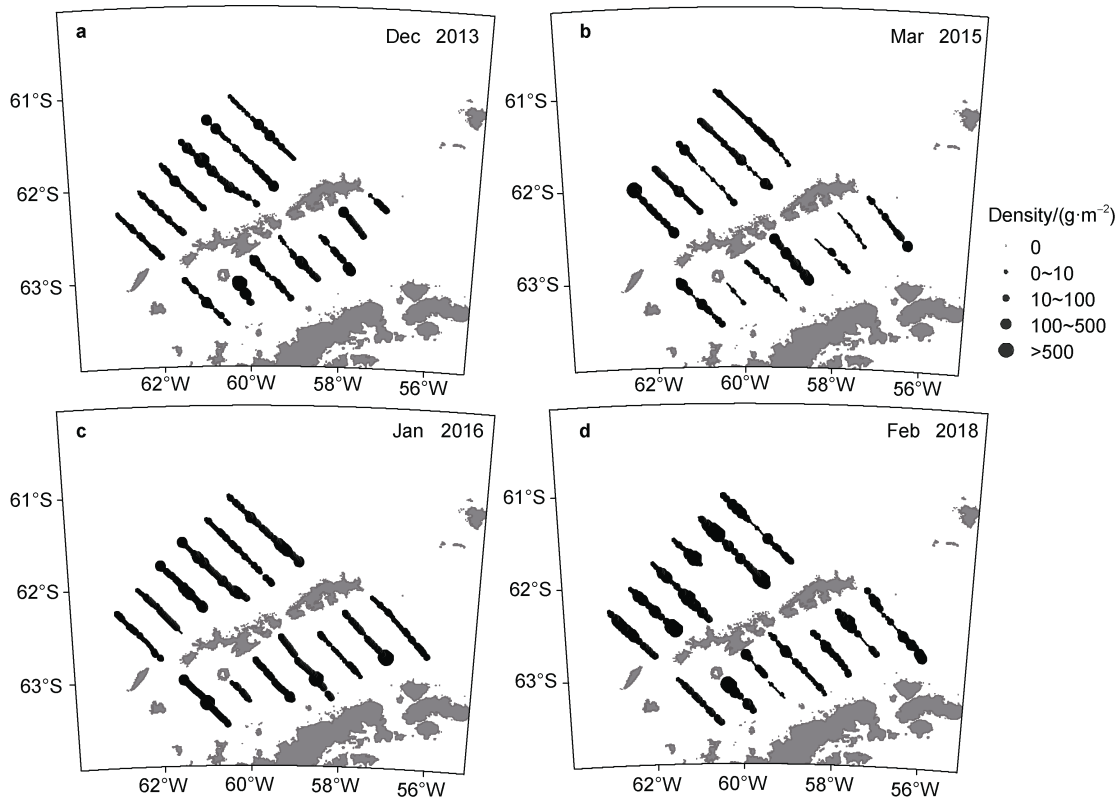

Figure 2 Distribution of Antarctic krill biomass in the waters surrounding the South Shetland Islands. **a**, December 2013; **b**, March 2015; **c**, January 2016; **d**, February 2018.

Acoustic data were collected using a Simrad EK60 scientific echosounder (operating at 38, 70, and 120 kHz). Data were processed and analyzed using Echoview fisheries acoustic post-processing software. The acoustic assessment of krill resource density followed the methodology recommended by the Working Group on Acoustic Survey and Analysis Methods of the SC-CAMLR (Krafft et al., 2021). Krill echo identification was performed using a swarm-based method: the echo integration layer was set from 15 m to 250 m below the sea surface, with manual adjustments applied as necessary. The nautical area scattering coefficient for krill was calculated and exported at a resolution of 1 n mile using 120 kHz acoustic data.

Based on their established ecological relevance to Antarctic krill, 11 environmental covariates were selected to represent key habitat dimensions, including the physical environment (e.g., sea surface temperature and salinity), food availability (e.g., chlorophyll-*a* concentration), and seafloor topography (e.g., bathymetry) (Table 1). These covariates include 3 static variables, that is, unchanging with survey time: bathymetry, slope, and distance from the shelf break (the 1000 m isobath was used to represent the position of the shelf break). The 8 remaining variables were dynamic variables: distance from the sea ice edge (defined as 15% sea ice concentration), sea surface temperature, bottom temperature, sea surface salinity, sea surface height, mixed layer thickness, current velocity, and chlorophyll-*a* concentration. Finally, year and month were included as factorial predictors in an exploratory capacity to assess the potential magnitude of temporal effects on krill distribution.

2.2 Data resampling and variable selection

Owing to the substantial disparity in longitudinal and latitudinal distances within the Antarctic Peninsula region, all data in this study were reprojected into the equal-area WGS_1984_EASE-Grid_2.0_South (EPSG: 6932) projection system. The study area was subsequently discretized into grid cells based on true ground distance (in km), creating regular grids at 4 spatial resolutions (5, 10, 15, and 20 km) to account for their potential effects on the model. For acoustic survey samples, the arithmetic mean of krill density within each corresponding grid cell was computed. Environmental raster data were resampled using bilinear interpolation to ensure consistency across resolutions. Furthermore, to mitigate the effects of multicollinearity among environmental variables, both Variance Inflation Factor (VIF) tests and Pearson correlation analysis were applied. When 2 variables were highly correlated ($|R| > 0.7$), the variable demonstrating a stronger correlation with krill density was retained for subsequent modeling.

2.3 Modeling algorithm

Given the zero-inflated and nonlinear characteristics of krill density data, we selected 4 representative algorithms, GAM (statistical model with interpretable smooths), RF and XGBoost (tree-based ensemble learners with high robustness), and ANN (deep-learning architecture with nonlinear flexibility), to compare their ecological applicability. The implementations for establishing the relationship between krill density and environmental factors by using these four algorithms were described as follows. All analyses were implemented in R version 4.3.0 (R Core Team, 2023).

Table 1 Summary of environmental covariates evaluated for model input

Covariate	Abbreviation	Spatial resolution	Units	Data source
Bathymetry	bathy	1 km×1 km	m	GEBCO Compilation Group, 2023
Bathymetric slope	slope	1 km×1 km	°	Calculated in ArcGIS 10.6 from the GEBCO 2023 bathymetry data
Distance to shelf break	distbreak	1 km×1 km	km	
Sea surface salinity	sss	0.083°×0.083°	PSU	Copernicus Marine Service, 2023
Sea surface temperature	sst	0.083°×0.083°	℃	
Bottom temperature	bottomT	0.083°×0.083°	℃	
Sea surface height above geoid	zos	0.083°×0.083°	m	
Mixed layer thickness	mlotst	0.083°×0.083°	m	
Geostrophic current velocity	velocity	0.083°×0.083°	$m \cdot s^{-1}$	
Distance from ice edge (15% ice concentration)	distSIE	0.083°×0.083°	km	Calculated in ArcGIS 10.6 from sea ice concentration, data obtained from the Copernicus Marine Service (2023)
Chlorophyll-*a* concentration	chla	4 km×4 km	$mg \cdot m^{-3}$	Copernicus Marine Service, 2025

2.3.1 Generalized Additive Model (GAM)

GAM represent a nonparametric extension of generalized linear model, characterized by additive effects of explanatory variables and the ability to capture nonlinear relationships between the response variable and predictors (Wood, 2017). A key advantage of GAM lies in their capacity to model complex nonlinear patterns while maintaining interpretability, as the contribution of each variable can be visually represented. The general form of a GAM is expressed as:

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

where Y is the response variable, $g(\ )$ is the link function, $f_i(\ )$ is a smooth function of explanatory variables spline, x_i are explanatory variables, α is the intercept term, n is the number of explanatory variables, and ε is the residual error term. In our study, the GAM model was implemented using the "mgcv" package, with a Tweedie error distribution specified for the model.

2.3.2 Random Forests (RF)

RF is an ensemble learning algorithm based on decision trees that enhances predictive accuracy and robustness by constructing multiple trees and aggregating their predictions (Breiman, 2001). This method effectively handles nonlinear relationships, high correlations among explanatory variables, and missing data, while demonstrating strong resistance to overfitting. The algorithm operates through 4 key steps: bootstrap sampling, where multiple subsets of the original data are randomly sampled with replacement; feature subspace selection, where a random subset of features is selected at each node split; independent tree construction, where each tree is grown to maximum depth without pruning; and prediction aggregation, where the final output is determined by averaging all tree predictions. This dual randomization process in both data and feature selection ensures model diversity and generalization capability, making RF particularly suitable for modeling complex ecological relationships. In our study, the RF model was implemented using the "randomForest" package with the parameter configuration set to ntree = 1000 (number of decision trees) and mtry = 3 (number of randomly selected features for node splitting).

2.3.3 Artificial Neural Networks (ANN)

ANN is a machine learning model designed to mimic biological neural networks. It captures complex relationships in data through interconnected neurons and activation functions (Maravelias et al., 2003). Theoretical studies have shown that a feedforward network with a single hidden layer can approximate any continuous function to arbitrary precision, provided a sufficient number of hidden neurons are available, thereby enabling the modeling of highly nonlinear patterns. There are many types of ANN, a hidden-layer is used in our study as a feedforward network. The network consists of 3 layers: an input layer (explanatory variables), a hidden layer (with multiple neurons), and an output layer (response variables). Due to the sensitivity of neural networks to the resolution of input features, all variables were standardized before being fed into the model using the following formula:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \quad (2)$$

where x is the original feature value, μ is the mean value, σ is the standard deviation of the feature, $x_{\text{standardized}}$ is the standardized value. This process ensures that all features have a mean of 0 and a standard deviation of 1, thereby eliminating the influence of resolution differences on model training. In our study, the ANN model was implemented using the "nnet" package with the parameter configuration set to size = 10 (number of hidden neurons), decay = 0.01 (weight decay coefficient), and maxit = 500 (maximum number of iterations).

2.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble learning algorithm based on gradient boosting. It employs Classification and Regression Trees as base learners and operates by iteratively adding new trees that correct the residuals of the previous ensemble, thereby progressively reducing model bias (Chen and Guestrin, 2016). The algorithm incorporates regularization terms in its objective function and utilizes second-order Taylor expansion to approximate the loss function, which significantly enhances computational efficiency, improves generalization, and mitigates overfitting. Compared to other machine learning methods, XGBoost offers superior computational speed, stronger predictive performance, and enhanced robustness in handling missing values and sparse matrix data, characteristics particularly advantageous in modeling marine species distribution. Owing to these capabilities, XGBoost has been successfully applied in ecological modeling studies (Muñoz-Mas et al., 2019), and is well-suited for predicting the distribution of Antarctic krill resources. In our study, the XGBoost model was implemented using the "mlr3" package with the parameter configuration set as follows: nrounds = 500 (number of boosting iterations), eta = 0.01 (learning rate), max_depth = 5 (maximum tree depth), and objective = "reg:tweedie" (loss function for Tweedie regression).

2.3.5 Ensemble modeling

To leverage the complementary strengths of individual algorithms and enhance predictive robustness, we developed an ensemble model that integrates predictions from GAM, RF, and XGBoost. Given the comparatively poorer and unstable predictive performance of the ANN across spatial resolutions (as detailed in Section 3.3), it was excluded from the ensemble. The ensemble model was constructed using a performance-based weighted averaging approach, where the weights were determined by combining the predictive accuracy and explanatory power of each base model.

The ensemble prediction y_{den} was computed as:

$$y_{\text{den}} = W_{\text{GAM}} y_{\text{GAM}} + W_{\text{RF}} y_{\text{RF}} + W_{\text{XGBoost}} y_{\text{XGBoost}} \quad (3)$$

where y_{GAM}, y_{RF} and y_{XGBoost} denote the predictions from GAM, RF, and XGBoost, respectively; w_{GAM}, w_{RF}, and w_{XGBoost} are their corresponding normalized weights.

The weighting scheme was developed through a 3-step process.

Root Mean Squared Error (RMSE)-based weights: weights were assigned inversely proportional to each model's RMSE.

$$W_{\mathrm{RMSE},\,m}=\frac{1/E_{\mathrm{RMSE},\,m}}{\sum_{k\in M}1/E_{\mathrm{RMSE},\,k}} \tag{4}$$

where $E_{\mathrm{RMSE,m}}$ represents the RMSE of model m, M={GAM, RF, XGBoost}.

Coefficient of Determination (R^2)-based weights: weights were assigned proportional to each model's R^2 value.

$$W_{R^2,m}=\frac{R_m^2}{\sum_{k\in M}R_k^2} \tag{5}$$

where R_m^2 represents the R_2 of model m M={GAM, RF, XGBoost}.

Combined weights: the final weights were computed as the normalized average of the RMSE-based and R^2-based weights:

$$W_m=\frac{W_{\mathrm{RMSE},\,m}+W_{R^2,m}}{\sum_{k\in M}\left(W_{\mathrm{RMSE},\,k}+W_{R^2,k}\right)} \tag{6}$$

This dual-metric weighting strategy ensures that models demonstrating both high predictive accuracy (low RMSE) and strong explanatory power (high R^2) receive greater influence in the ensemble prediction, the ensemble model was implemented in R using a custom EnsembleModel class.

2.4 Modeling evaluation

To fully utilize the available data, the entire dataset was employed to assess the model's fitting performance. A 10-fold cross-validation procedure was implemented to evaluate the model's predictive performance: the complete dataset was partitioned into 10 subsets, with 9 subsets used as the training set and the remaining subset as the test set in each iteration; this process was repeated 10 times. Model performance was quantified using 3 metrics: the Mean Absolute Error (MAE), RMSE, and R^2.

MAE represents the average of the absolute differences between predicted values and observed values. It is calculated using the following formula:

$$E_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}\left|O_i-P_i\right| \tag{7}$$

Mean Squared Error (MSE) quantifies the average squared difference between predicted and observed values, while the RMSE is defined as the square root of MSE, representing the actual scale of prediction errors. The formulas are as follows:

$$E_{\mathrm{RMSE}}=\sqrt{E_{\mathrm{MSE}}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(O_i-P_i\right)^2} \tag{8}$$

R^2 is used to quantify the proportion of variance in the dependent variable that is explained by the model. The formula is as follows:

$$R^2=1-\frac{\sum_{i=1}^{n}\left(O_i-P_i\right)^2}{\sum_{i=1}^{n}\left(O_i-\overline{O}\right)^2},\quad \overline{O}=\frac{1}{n}\sum_{i=1}^{n}O_i \tag{9}$$

where n is the number of observations, O_i denotes the observed values, P_i represents the predicted values, $\overline{O}$ is the mean of the observed values.

3 Results

3.1 Distribution of krill density data across spatial resolutions

The distribution of krill density data across the 4 spatial resolutions (5, 10, 15, and 20 km) is summarized in Table 2. The overall density range of 0–1047.20 g·m^{-2}, was observed at the 5 km, 10 km, and 15 km resolutions, while 0–697.85 g·m^{-2} was recorded at the 20 km resolution. The data exhibited a right-skewed distribution at all resolutions, as evidenced by mean values (33.32 to 37.05 g·m^{-2}) consistently exceeding the medians (8.90 to 14.89 g·m^{-2}). This skewness attenuated with increasing spatial resolution: the number of data points (n) decreased from 895 to 274, zero-values declined markedly (from 128 to 12), and both the median and kurtosis increased, indicating reduced data dispersion and a stronger concentration of values around the mean (Figure 3).

Table 2 Descriptive statistics of krill density (g·m^{-2}) at different spatial resolutions

Spatial resolution	*n*	Zeros	Mean	Median	Max
5 km×5 km	895	128	33.32	8.90	1,047.20
10 km×10 km	499	46	36.61	11.50	1,047.20
15 km×15 km	356	21	37.05	14.33	1,047.20
20 km×20 km	274	12	36.79	14.89	697.85

3.2 Environmental variable selection and model specification

At 4 spatial resolutions (5, 10, 15, and 20 km grids), environmental variables exhibiting collinearity included: zos-bathy (R: −0.80 to −0.82), zos-distbreak (R: 0.79 to 0.82), and bathy-distbreak (R: −0.93) (Figure 4). Considering both the collinearity among predictors and their ecological relevance to krill density, zos and bathy were excluded, while the remaining nine environmental variables were retained for subsequent modeling. The general structure of the species distribution model is expressed as:

$$\begin{aligned}\ln(\text{density}+1)\sim{}&\text{sst}+\text{bottomT}+\text{sss}+\text{mlotst}+\text{velocity}+\\&\text{distSIE}+\text{chla}+\text{slope}+\text{distbreak}+\\&\text{year}+\text{month}\end{aligned} \tag{10}$$

where the left side is a response variable that represents the natural logarithm of (krill density + 1), while the additive combination of all environmental and spatiotemporal predictor variables on the right side is used to describe the response variable.

3.3 Comparative performance of different models across spatial resolutions

As shown in Table 3, the fitting performance of GAM, RF, ANN, and XGBoost models was evaluated across 4

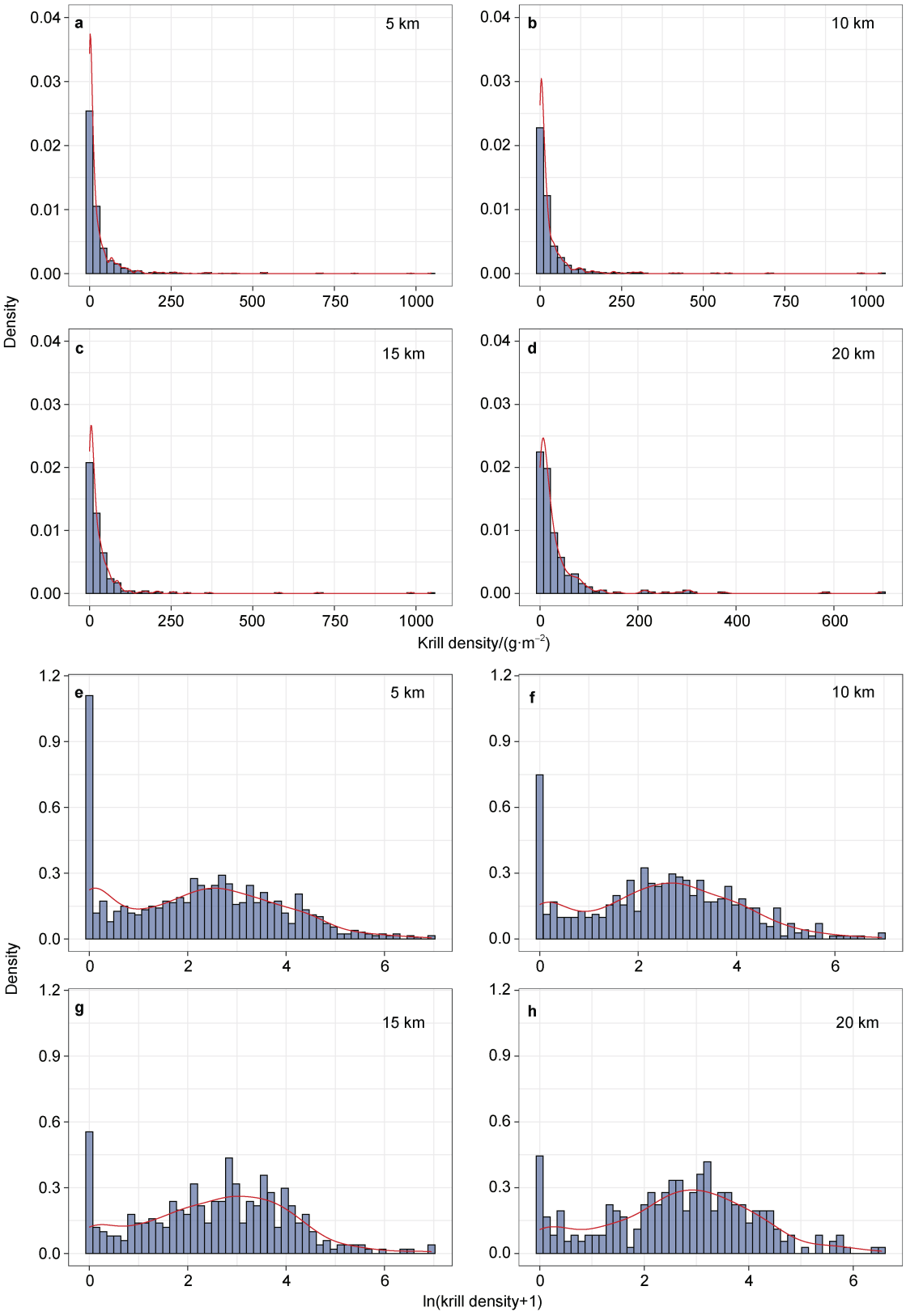

Figure 3 Histograms of krill density distribution at different spatial resolutions. **a**–**d**, histograms of the original (untransformed) krill density values (g·m^{-2}) at 5, 10, 15, and 20 km resolutions, respectively. The *y*-axis (density) represents the relative frequency of observations. **e**–**h**, corresponding distributions of the log-transformed values ln(krill density+1). The transformation was applied to normalize the strongly right-skewed data and reduce the influence of extreme values for modeling.

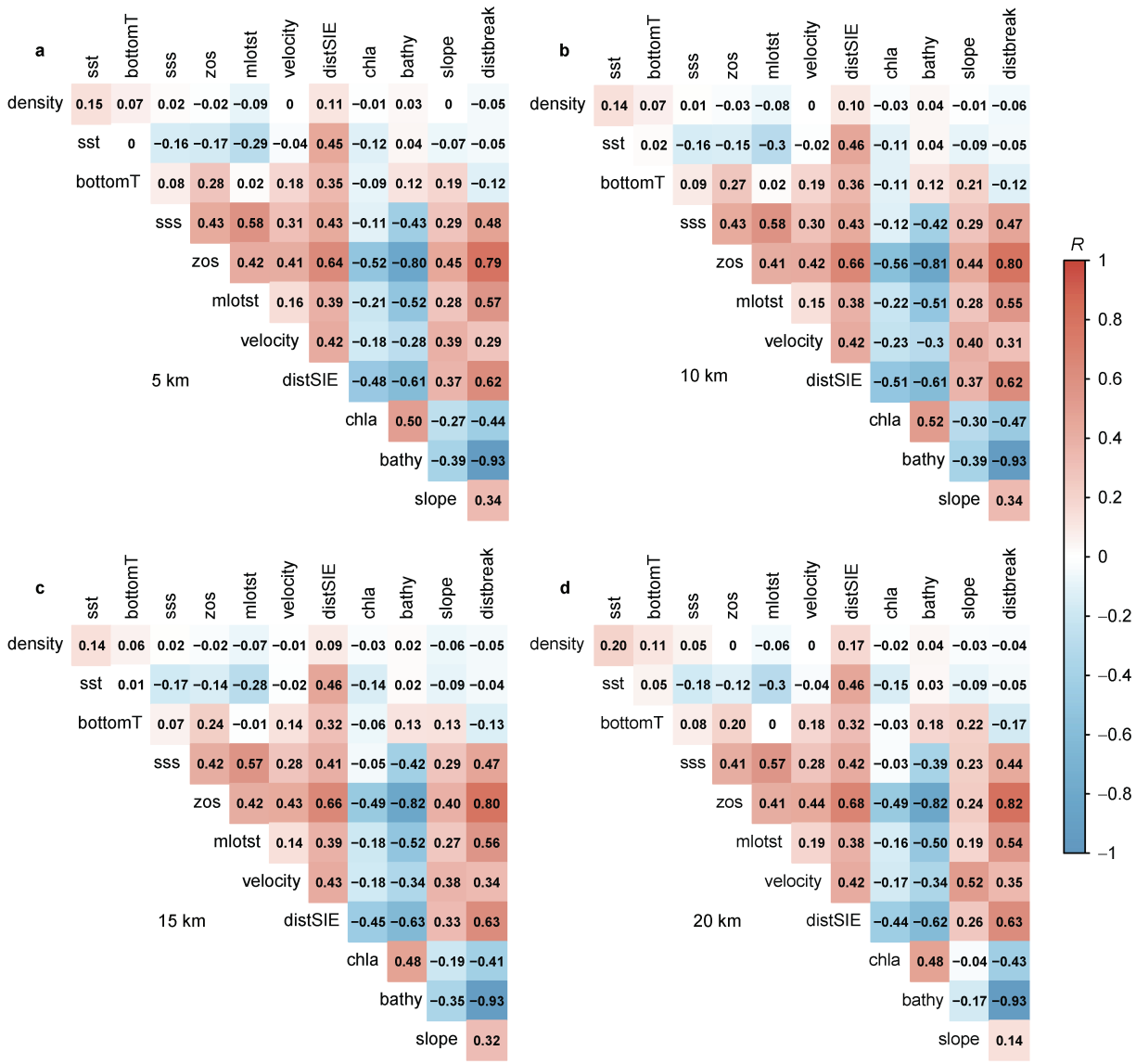

Figure 4 Heatmap of correlation coefficients among environmental factors at different spatial resolutions. **a**, 5 km spatial resolution; **b**, 10 km spatial resolution; **c**, 15 km spatial resolution; **d**, 20 km spatial resolution.

spatial resolutions (5, 10, 15, and 20 km). RF consistently demonstrated superior performance at most spatial resolutions (MAE 0.48–0.50; RMSE 0.61–0.63; R^2 0.82–0.85). The minimal fluctuation of these metrics across resolutions indicates strong robustness to spatial resolution variations. XGBoost ranked as the second-best performing model, showing improved fitting performance with increasing spatial resolution. At the 20 km spatial resolution, XGBoost achieved optimal performance (MAE 0.39; RMSE 0.50; R^2 0.88), exceeding that of RF. The ensemble model demonstrated competitive performance across all resolutions, consistently ranking between the best and second-best individual models. At 5 km resolution, the ensemble achieved MAE=0.69, RMSE=0.85, and R^2=0.72, outperforming both GAM and ANN but slightly behind RF and XGBoost. This pattern persisted across intermediate resolutions, with the ensemble effectively balancing the

Table 3 Model fitting performance across spatial resolutions and algorithms

Spatial resolution	Method	RMSE	MAE	R^2
5 km×5 km	GAM	1.49	1.24	0.14
	RF	**0.62**	**0.49**	**0.85**
	ANN	1.18	0.93	0.46
	XGBoost	0.98	0.78	0.62
	Ensemble	0.85	0.69	0.72
10 km×10 km	GAM	1.41	1.15	0.17
	RF	**0.63**	**0.49**	**0.83**
	ANN	0.99	0.75	0.59
	XGBoost	0.78	0.60	0.74
	Ensemble	0.78	0.62	0.74

continued

Spatial resolution	Method	RMSE	MAE	R^2
15 km×15 km	GAM	1.37	1.12	0.13
	RF	**0.63**	**0.50**	**0.82**
	ANN	0.95	0.73	0.58
	XGBoost	0.65	0.52	0.80
	Ensemble	0.71	0.58	0.76
20 km×20 km	GAM	1.26	1.00	0.23
	RF	0.61	0.48	0.82
	ANN	0.65	0.46	0.80
	XGBoost	**0.50**	**0.39**	**0.88**
	Ensemble	0.64	0.50	0.81

Notes: The best performance values for each metric at each spatial resolution are highlighted in bold. RMSE, Root Mean Square Error; MAE, Mean Absolute Error; R^2, Coefficient of Determination.

strengths of its constituent models. In comparison, both ANN and GAM exhibited relatively weaker overall performance.

The boxplots constructed from 10-fold cross-validation results systematically reveal the performance variations of the four models across different spatial resolutions (Figure 5). At 5, 10, and 15 km resolutions, RF demonstrated better predictive performance than the other 3 algorithms, achieving its strongest predictive capability at the 5 km resolution (MAE=1.13±0.10, RMSE=1.42±0.11, R^2=0.23±0.08). GAM exhibited unique resolution-responsive characteristics, showing reduced errors and increased explanatory power as spatial resolution became coarser. At the 20 km resolution, its performance surpassed that of RF, demonstrating optimal results. XGBoost achieved second-best performance at 5, 10, and 15 km resolutions, but exhibited obvious performance degradation at the 20 km resolution. ANN consistently showed the poorest performance

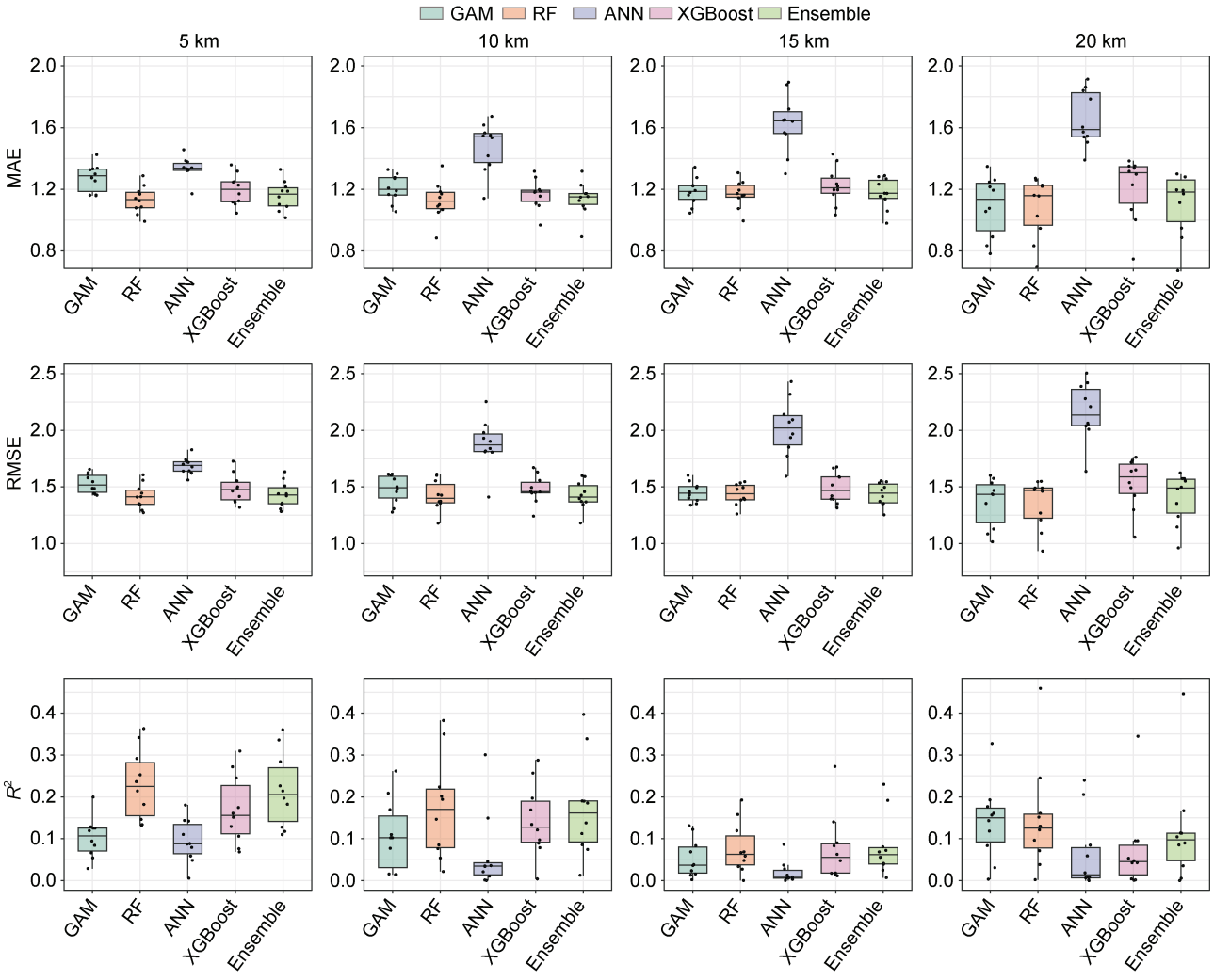

Figure 5 Evaluation of model predictive performance using 10-fold cross-validation.

across all resolutions, with predictive capability continuously declining as resolution coarsened, revealing its inherent limitations in modeling small-sample ecological data. Notably, the increased box heights for error metrics across all models at the coarse resolution (20 km) indicate that information loss during spatial aggregation amplifies model uncertainty. The ensemble model demonstrated competitive, albeit not the best, predictive performance across all spatial resolutions. As detailed in Table 3, its accuracy metrics (MAE, RMSE and R^2) consistently ranked between the best and second-best performing individual models. This pattern of delivering consistently intermediate performance was observed across most resolution scenarios. More importantly, as shown in Figure 5, the ensemble model exhibited reduced variance in its predictions compared to the individual models, indicating enhanced robustness and lower uncertainty.

3.4 Environmental predictor importance and partial dependence responses across models and spatial resolutions

Despite identifying a consistent set of key environmental drivers, the variable importance analysis revealed that the perceived influence of these predictors was highly algorithm-dependent, with tree-based methods (RF and XGBoost) showing greater stability across spatial resolutions than ANN and GAM (Figure 6). The GAM model identified distSIE, distbreak, sst, and chla as the most correlative environmental variables, whose relative importance varied with spatial resolution, while other variables remained relatively stable. In RF, distSIE, sst, mlotst, and bottomT were the relatively important variables, whereas XGBoost highlighted distSIE, mlotst, bottomT, and distbreak. Both RF and XGBoost maintained largely consistent variable importance rankings across resolutions. However, as spatial resolution coarsened, the error induced by single variables decreased in RF models, whereas it increased in XGBoost models. Conversely, the ANN model exhibited instability in relative variable importance, showing inconsistent rankings across spatial resolutions, indicating its failure to clearly identify dominant environmental factors. For the ensemble model, the variable importance reflected a composite of its constituent models'patterns. distSIE emerged as the most consistently important variable across all resolutions, followed by mlotst and sst. This pattern aligns closely with the results from RF and XGBoost, confirming the dominance of these environmental drivers in shaping krill distribution.

Analysis of partial dependence plots revealed significant differences in the response patterns of environmental factors to krill density among the 4 models across spatial resolutions. The GAM model produced the smoothest and most interpretable response curves, showing negative correlations between krill density and both sst and velocity, and positive correlations with sss and distSIE. The response patterns of other variables exhibited minor variations but remained within ecologically reasonable ranges (Figure 7). The RF model maintained consistent response patterns across resolutions from 5 km to 20 km, though predicted values increased with coarser spatial resolutions. The RF response curves displayed complex nonlinear patterns that may challenge ecological interpretation (Figure 8). The ANN model generated unstable and unpredictable partial dependence patterns across resolutions, with response curves varying dramatically between different resolutions. Frequent sign reversals and non-monotonic patterns were observed, lacking ecological plausibility (Figure 9). The XGBoost model produced the most stable and ecologically meaningful partial dependence curves across resolutions. The response curves exhibited smooth nonlinear patterns that remained consistent in both form and predicted values from 5 km to 20 km resolutions. For key predictors including distSIE, mlotst, and bottomT, relatively higher krill density was observed under conditions where distance to sea ice edge exceeded 150 km, mixed layer depth was less than 20 m, and bottom temperature was higher than −0.25 ℃ (Figure 10). The ensemble model generated

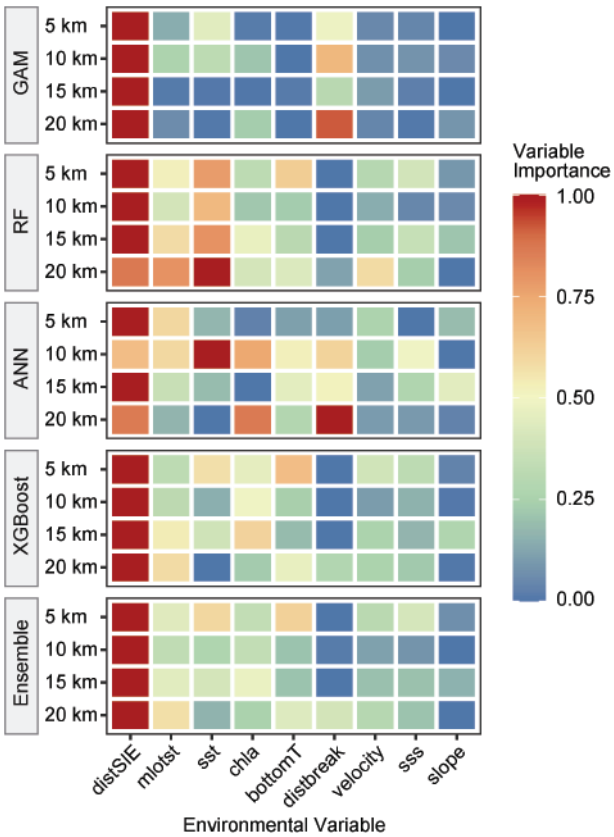

Figure 6 Comparison of standardized variable importance across different models and spatial resolutions. Variable importance was calculated using Permutation Importance (based on RMSE) and was standardized to a [0,1] scale for each model-resolution combination to enable cross-comparison.

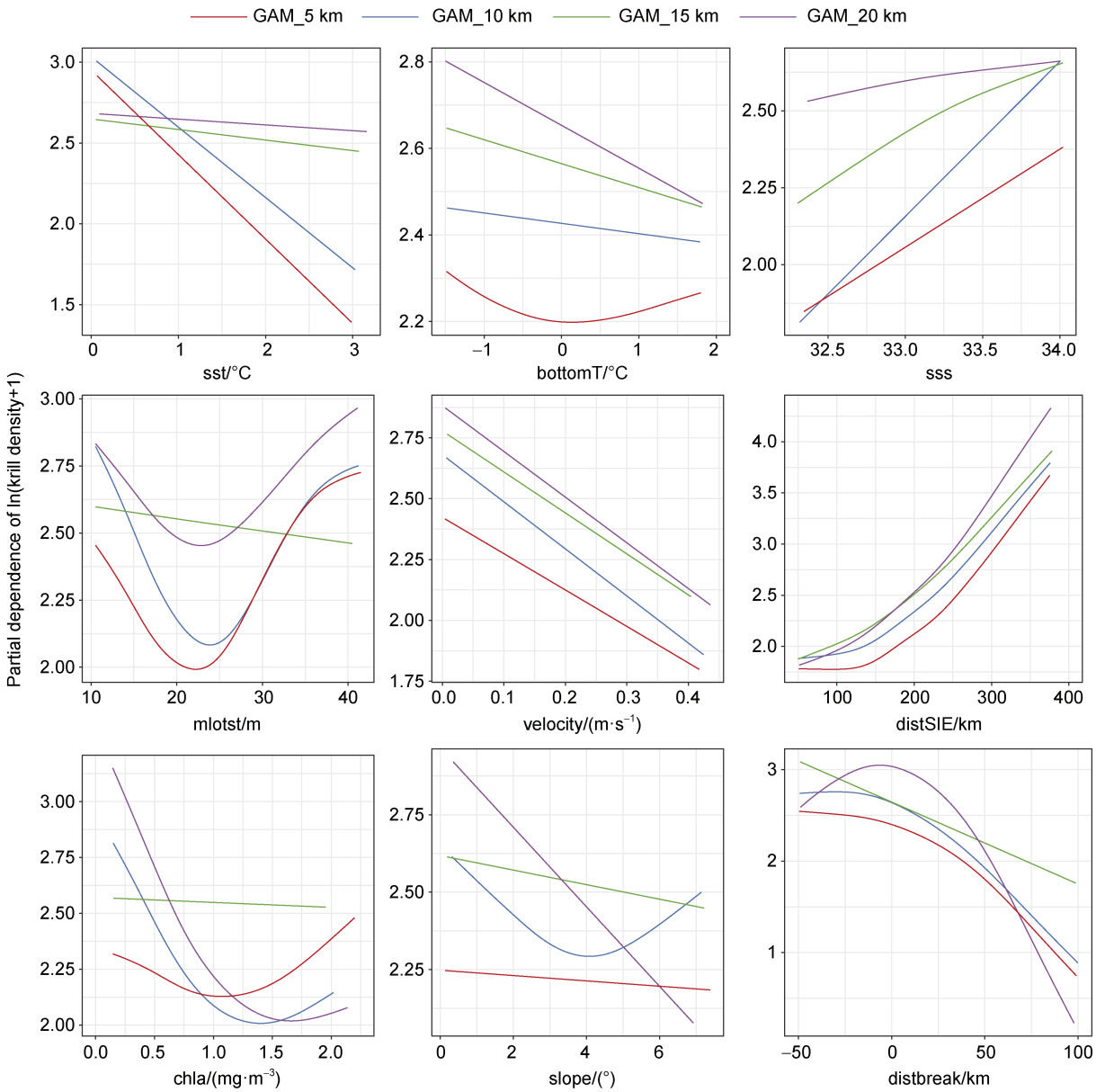

Figure 7 Partial dependence plots for environmental factors from the GAM at different spatial resolutions.

partial dependence curves that represented a balanced compromise between the smooth, interpretable patterns of GAM and the complex, data-driven responses of RF and XGBoost (Figure 11). The curves exhibited smoother transitions than individual tree-based models while capturing more nuanced responses than GAM, thereby combining the statistical robustness of averaging with ecological realism.

4 Discussion

Four modeling approaches, including GAM, RF, ANN, and XGBoost, were evaluated in their performance on predicting Antarctic krill distribution, across different spatial resolutions (5, 10, 15 and 20 km). The interaction between modeling algorithms and spatial resolution revealed complex and nuanced differences when handling ecological datasets with small sample sizes and zero-inflation. Overall, RF demonstrated the strongest comprehensive performance and robustness to changes in spatial resolution, while the differential behavior of the models at fine versus coarse resolutions reflects their varying sensitivities to multi-scale drivers of ecological processes.

4.1 Model algorithm suitability for the characteristics of krill distribution data

Antarctic krill exhibits a highly aggregated and heterogeneous spatial distribution (Burns et al, 2022), resulting in a zero-inflated, right skewed structure of

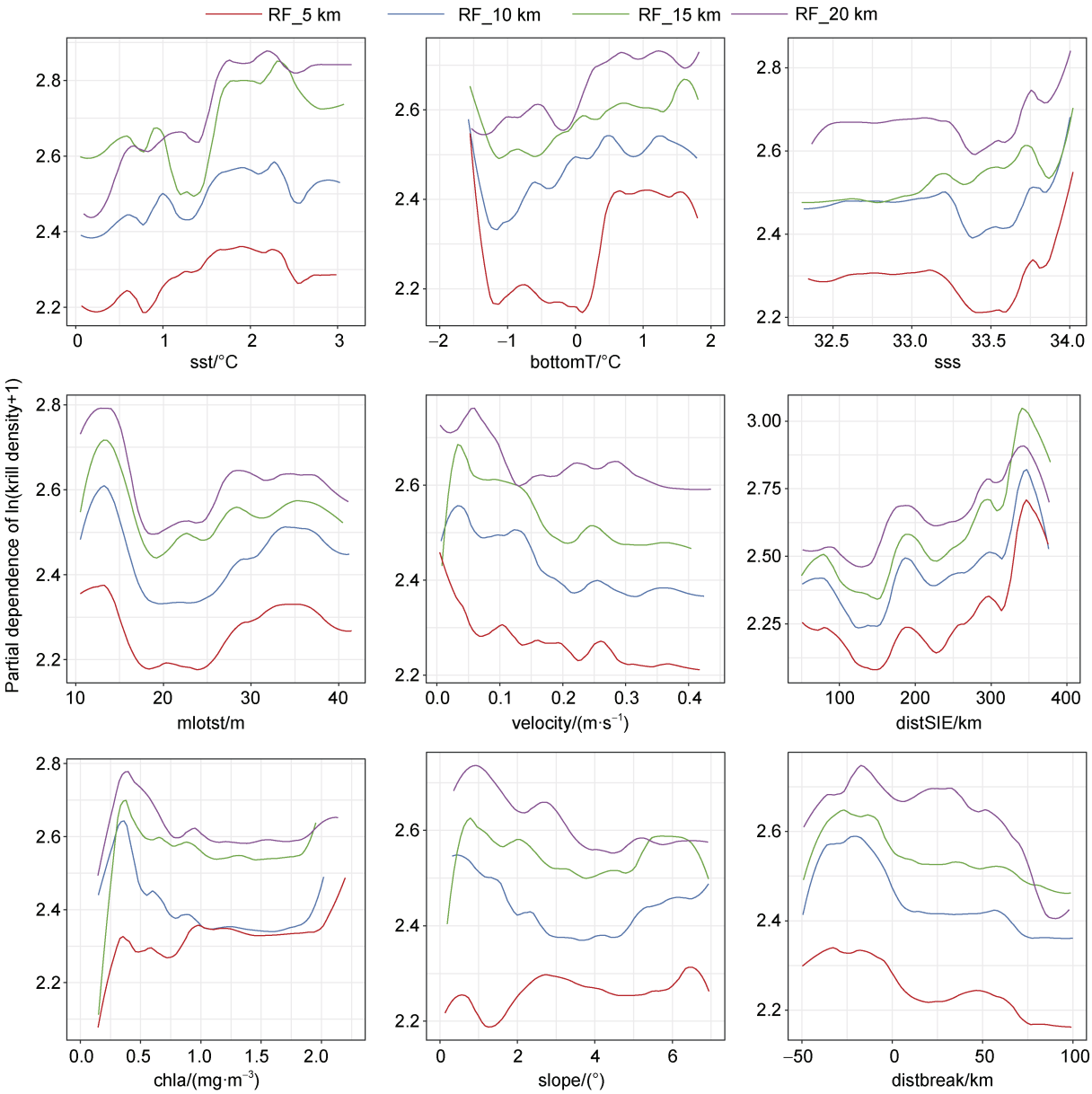

Figure 8 Partial dependence plots for environmental factors from the RF at different spatial resolutions.

density data. This structure comprises a large number of zero observations and a continuous, positively skewed distribution of positive values, approximately log-normal.

For GAM and XGBoost, which can incorporate specific distributional assumptions, the Tweedie distribution is often employed. The Tweedie distribution, an exponential dispersion model, is particularly effective in simultaneously modeling data with a high frequency of zeros and a continuous, skewed positive component (Mustafa, 2025). This makes it widely used in ecological modeling of species distributions with zero-inflation data (Shono, 2008).

However, the Tweedie distribution requires the parameterization of a specific variance-mean relationship (Kokonendji, 2004). This fixed structure, while advantageous for model interpretation and stability, may be less flexible than entirely non-parametric methods (like RF) for capturing exceptionally complex or unforeseen structures in the data.

The RF algorithm, by contrast, leverages binary recursive partitioning through tree-based splitting rules to effectively separate regions of zero aggregation from areas with non-zero responses. It does not require prior assumptions about data distribution and is naturally robust to zero-inflated structures (Liaw and Wiener, 2002).

The poor performance of the ANN can be largely attributed to its high demand for training data. ANN requires substantial sample sizes to reliably estimate its numerous connection weights (Goodwin et al., 2022). Consideration with fewer than 1,000 samples available, the model may be prone to memorizing noise in the training

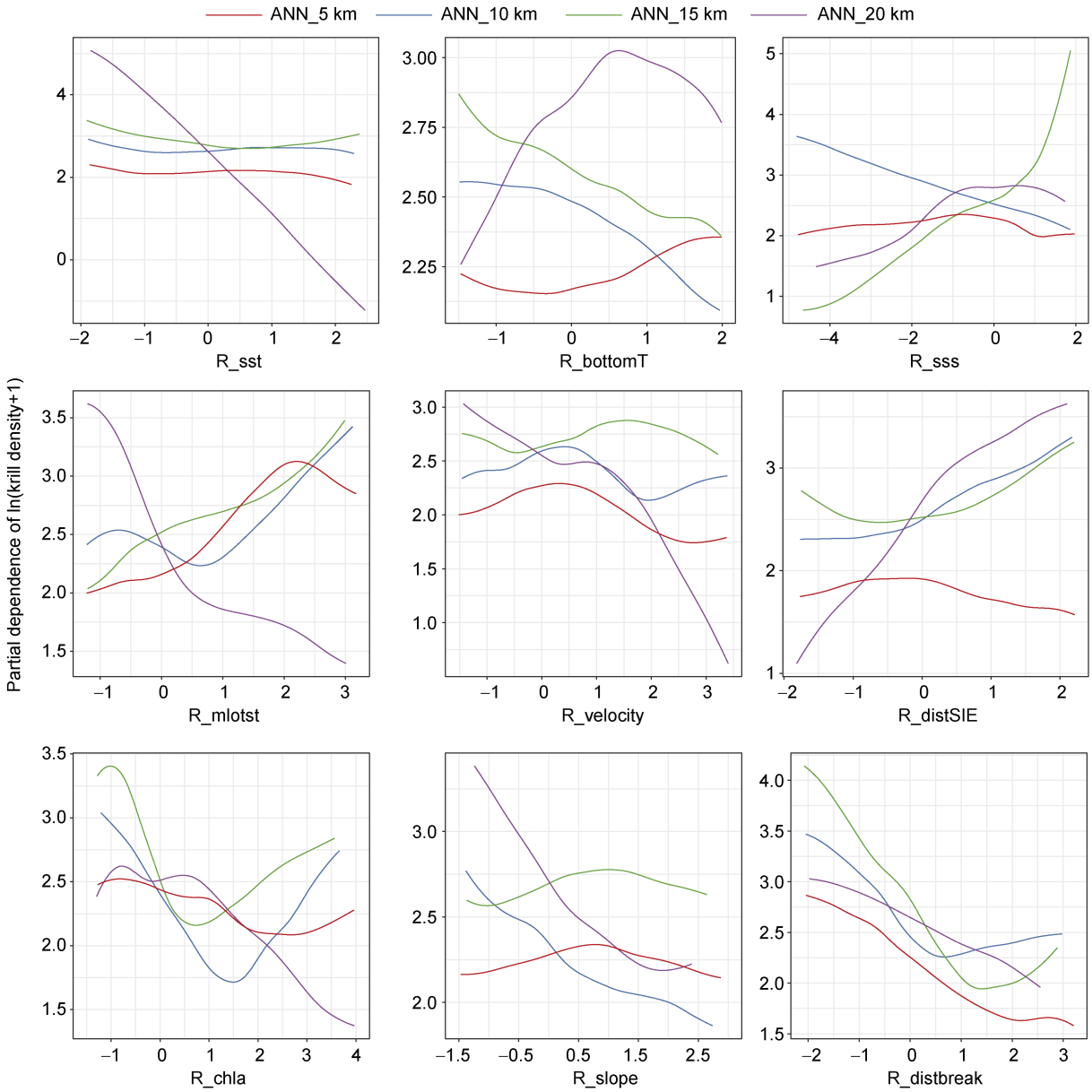

Figure 9 Partial dependence plots for environmental factors from the ANN at different spatial resolutions.

data rather than learning generalizable patterns, and leading to a collapse in its predictive generalization ability. In this study, the krill distribution dataset contained 895 samples at the finest resolution (5 km), with even smaller sample sizes at coarser resolutions. Furthermore, the zero-inflated, right-skewed distribution of krill density data may pose additional challenges for standard ANN architectures. ANN remains a powerful tool for species distribution modeling, but its application is better suited for studies with massive datasets, such as those derived from satellite telemetry, unmanned aerial vehicles, or environmental DNA (Swart et al., 2020; Suarez-Bregua et al., 2022). In contrast, the better performance of RF can be attributed to its use of bootstrap aggregation (bagging) and random feature subspace selection, which effectively reduce the variance of individual decision trees and enhance generalization under limited sample conditions.

Although XGBoost improves prediction accuracy through a gradient boosting framework that iteratively corrects errors in previous models, it is more sensitive to hyperparameters than RF (Wang et al., 2021). Under small-sample conditions, achieving optimal performance necessitates extensive and computationally expensive hyperparameter tuning, which may limit its practical applicability in some cases.

GAM did not perform well in goodness of fit, however, it performed better in predictions under coarse resolution scenarios than other models in this study. GAM's primary strength lies in its interpretability, making it particularly valuable in ecological studies with limited sample sizes

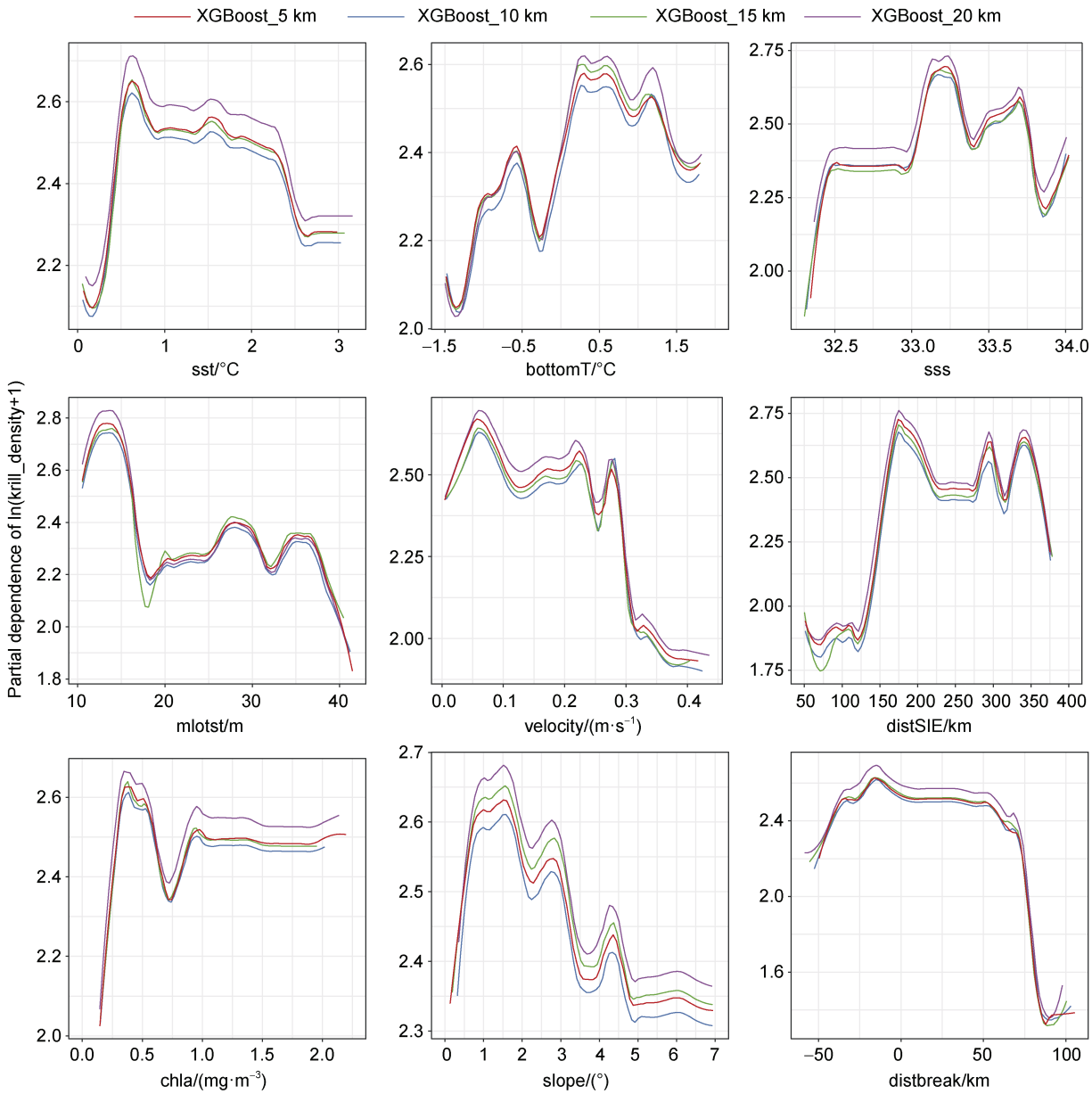

Figure 10 Partial dependence plots for environmental factors from the XGBoost at different spatial resolutions.

(Wood, 2017). Through the use of regularized additive smooth functions, GAM can mitigate overfitting and clearly reveal and quantify the nonlinear relationships between environmental covariates and species distributions (Wood, 2017).

Given the complementary strengths and limitations of these individual algorithms, RF's robust predictive accuracy, XGBoost's stable ecological inference, and GAM's interpretability, a logical progression is to combine them. The ensemble model developed in this study, which integrates GAM, RF, and XGBoost through a performance-based weighting scheme, was designed for this exact purpose. Its primary strength lies not in achieving peak predictive accuracy, but in its superior robustness and consistent intermediate performance. In practical ecological applications and management scenarios, a model that reliably avoids the poorest outcomes is often more valuable than one that oscillates between the best and worst. By effectively mitigating the risk of substantial prediction errors that can occur with individual models under specific conditions, the ensemble provides a more reliable and trustworthy foundation for the spatial management planning of krill fisheries.

4.2 Impacts of spatial resolution on model performance

Changes in spatial resolution not only alter sample size but also reshape the statistical distribution of data and the ecological information they convey, thereby impacting model performance. As spatial resolution became finer,

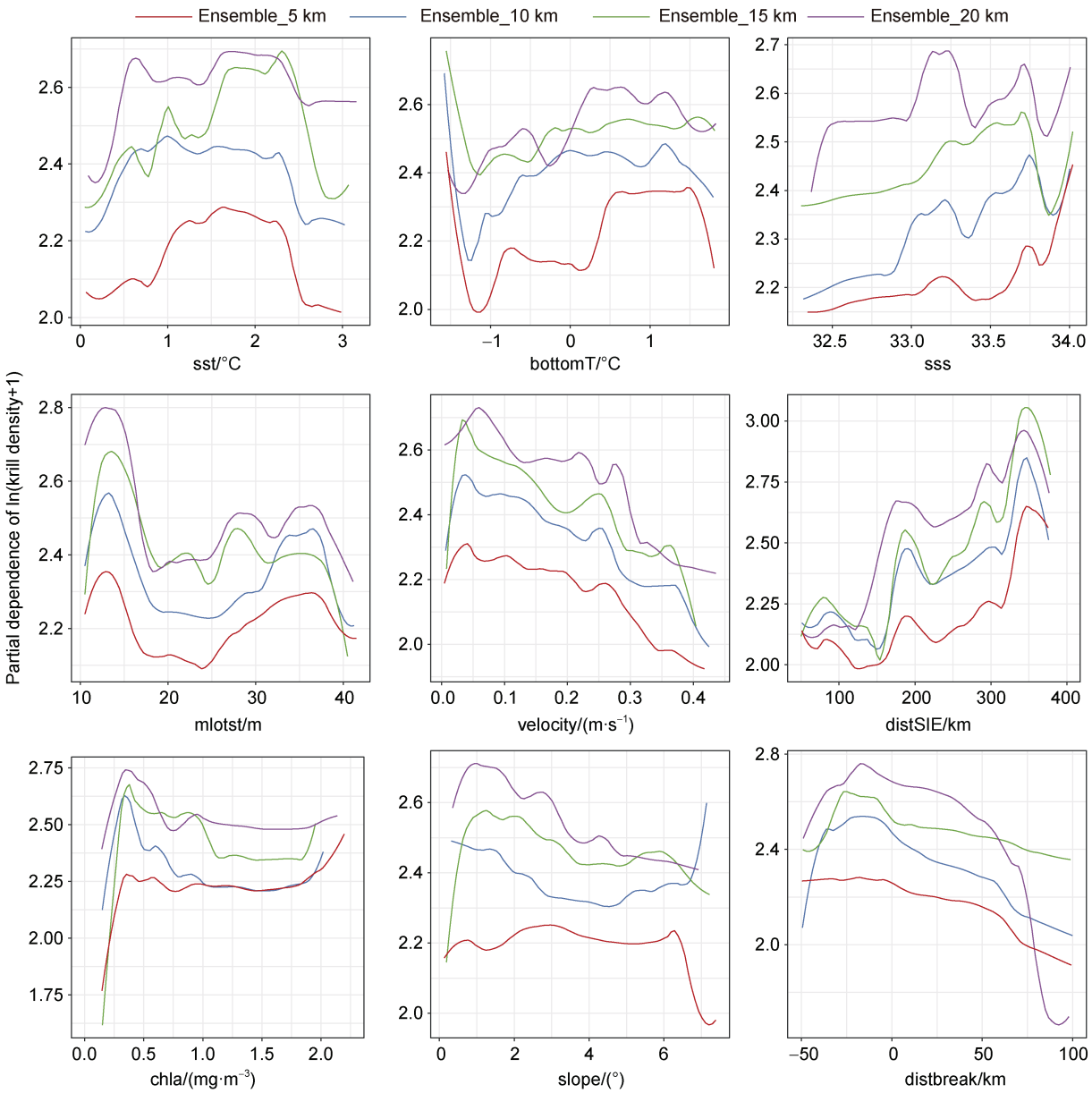

Figure 11 Partial dependence plots for environmental factors from the ensemble model at different spatial resolutions.

prediction accuracy tends to improve, however, the goodness-of-fit statistics showed differences by model. When resolution coarsened from 5 km to 20 km, there were a reduction in sample size from 895 to 274 and a significant decrease in the proportion of zero values. At coarser resolutions, spatial averaging within grid cells lead to smoothing of both environmental variables and species data, reducing data variability and consequently improving goodness-of-fit metrics for some model. However, this smoothing came at the cost of information loss. Finer resolutions preserved more detailed local habitat heterogeneity and fine-scale ecological processes.

RF, with its ability to efficiently capture complex feature interactions and model high-dimensional nonlinear relationships, can exploit these fine-scale patterns to achieve higher predictive accuracy. In face of different variance and noise associated with resolution, RF can also keep robustness on goodness-of-fit.

Coarsening disproportionately harms ANN prediction performance. Already constrained by limited sample size, the reduction in data volume at coarser resolutions prevents the model from learning meaningful patterns, leading to a sharp decline in performance.

In this study, spatial coarsening substantially reduced the proportion of zero values (from 14.3% at 5 km to 4.4% at 20 km) thereby alleviating the challenges posed by zero-inflated data structures during model fitting. This reduction is a key factor contributing to improved goodness-of-fit statistics at coarser resolutions. Moreover, the diminished severity of zero-inflation at coarser

resolutions creates more favorable conditions for the application of the Tweedie distribution in both XGBoost and GAM, as the variance-mean relationship becomes more stable. Additionally, this condition benefits GAM by enabling it to achieve predictive performance comparable to, or surpassing, that of RF at coarser resolution.

4.3 Sensitivity of algorithms to spatial resolution

XGBoost demonstrated remarkable resolution invariance, maintaining nearly identical relative importance of environmental variables and generating almost completely consistent partial dependence curves across all 4 spatial resolutions. This robustness suggests that XGBoost may successfully capture fundamental ecological relationships in the data and be insensitive to spatial resolution. Given the superior stability of XGBoost, future studies can explore the potential of other gradient boosting methods, such as LightGBM or CatBoost, which may offer additional advantages (Dorogush et al., 2018; Ke et al., 2017).

RF demonstrated nearly consistent relative importance of environmental variables and maintained the shape and trends of partial dependence curves across resolution. However, the absolute values of predicted density increased with coarser spatial resolutions. This resolution-dependent magnitude variation suggested the necessity of considering spatial resolution when interpreting quantitative predictions from RF (Georganos et al., 2021). Importantly, the model's ability to consistently identify the relative influence of variables and the shape of their relationships with krill density confirms its robust interpretability for uncovering consistent ecological patterns across scales.

GAM generated the smoothest and most interpretable response curves, consistent with its inherent modelling feature (Wood, 2017). However, certain variables exhibited substantial variability across resolutions, occasionally displaying anomalous patterns. This instability likely reflects increased uncertainty in estimating smooth functions under limited sample sizes (Wood, 2020).

For the data used by the present study, ANN demonstrated the poorest resolution stability, with both the relative importance of environmental variables and partial dependence plots demonstrating high sensitivity to spatial resolution. Trends for identical variables even showed contradictory patterns across different resolutions. This behavior indicates pronounced overfitting and extreme sensitivity to noise (Olden et al., 2008). Rather than learning generalizable ecological relationships, the ANN memorized resolution-specific spurious patterns. These findings emphasize the need for cautious implementation of complex "black-box" algorithms in data-limited ecological applications (Li et al., 2023).

4.4 Ecological insights and methodological implications

Despite differences in the ranking of variable importance across modeling algorithms and spatial resolutions, the consistent identification of several key environmental factors as highly important helps to clarify which ones exhibit stable, model-detectable associations with krill distribution.

Chlorophyll-*a* concentration and mixed layer depth were identified as important variables in most models. These parameters reflect the prey availability for krill and highlight the crucial role of "bottom-up" control mechanisms in the Southern Ocean ecosystem (Murphy et al., 2007).

Sea surface temperature, sea surface salinity, and bottom temperature essentially reflect the horizontal dimension of water mass structure. The waters surrounding the Antarctic Peninsula exhibit a complex composition of water masses, each characterized by distinct temperature-salinity properties and nutrient profiles that may shape biological distributions (Martinson et al., 2008). In this study, partial dependence plots from multiple models indicate that krill density peaks within specific ranges of temperature and salinity. This suggests a strong association between krill distribution and water masses with particular physicochemical characteristics.

Distance to sea ice edge (distSIE) and shelf break emerged as significant predictor variables, revealing the importance of large-scale geographic patterns in structuring krill distribution. Given that our data were collected during summer surveys, distSIE is unlikely to reflect direct sea ice effects, but rather serves as a broad-scale geospatial proxy (Atkinson et al., 2008). This variable may integrate multiple ecological processes: proximity to overwintering habitats (Siegel, 2005), productivity gradients following sea ice retreat (Ducklow et al., 2007), and legacy effects of ice-associated historical ecological processes (Quetin and Ross, 2003).

As previously discussed, coarsening spatial resolution in this study smoothed data distributions and alleviated zero-inflation, but at the potential cost of overlooking local ecological processes. This trade-off is closely linked to the distinctive patchy distribution feature of krill across multiple resolutions, ranging from meters to tens of kilometers (Tarling et al., 2009), which is itself strongly associated with the environmental factors of their habitat. Different ecological processes operate at distinct spatial scales: krill foraging behavior is likely influenced by phytoplankton patches at small resolutions (Hamner and Hamner, 2000), whereas population dynamics are largely controlled by ocean circulation and water mass structures at larger scales (Murphy et al., 2007).

4.5 Limitations and future directions

This study has several limitations that point toward valuable future research directions. First, the models were based solely on the summer acoustic surveys around the South Shetland Islands. Consequently, they may not fully capture seasonal distribution dynamics associated with winter sea-ice dynamics and prey availability (Siegel, 2005),

and their spatial extrapolation to wider areas is limited. Future study should extend data coverage in both seasonal and spatial dimensions to enable the development of robust and full annual distribution models. Second, the ensemble framework presented here is a first step. Future work could explore more sophisticated ensemble techniques, such as stacking (Wolpert, 1992) or Bayesian model averaging (Dormann et al., 2018), which might further enhance predictive performance and uncertainty quantification.

5 Conclusions

By systematically multiple modelling algorithms across spatial resolutions, this study demonstrates that the optimal choice for modeling Antarctic krill distribution depends critically on the research objective, whether it prioritizes predictive accuracy, ecological interpretability, or operational robustness.

Our findings reveal a fundamental trade-off: while finer spatial resolutions (5 km) generally enhanced the predictive capability of algorithms like RF, coarser resolutions (20 km) improved the goodness-of-fit for models like GAM by alleviating zero-inflation. RF consistently delivered high predictive stability, whereas XGBoost produced the most resolution-invariant ecological inferences. The GAM proved invaluable for generating interpretable response curves, despite its variable predictive performance.

Considering the limitations of individual algorithms, the ensemble model emerged as a robust solution that reconciles these trade-offs. Although it did not achieve the highest peak accuracy, it provided consistently intermediate performance and effectively hedged against the poor performance or instability of any single constituent model. This makes the ensemble approach particularly suitable for risk-averse applications like spatial fishery management, where reliable and trustworthy predictions are paramount.

Consequently, we recommend an ensemble modeling strategy: XGBoost for ecological inference, RF for high-resolution prediction, and GAM for interpretation. Future efforts should focus on refining such integrated frameworks to further enhance the accuracy and reliability of krill distribution forecasts, thereby directly supporting the science-based management of the Antarctic krill fishery and the conservation of the Southern Ocean ecosystem.

Acknowledgment We are grateful to the captains and crews of the F/V *Fu Rong Hai* and the scientific observers onboard for helping with data collection during the cruises. The research was funded by the National Key R&D Program of China (Grant no. 2022YFC2807504), the Marine S&T Fund of Shandong Province for Qingdao Marine Science and Technology Center (Grant no. 2022QNLM030002-1), and the Central Public-interest Scientific Institution Basal Research (Grant no. 2023TD02).

References

Atkinson A, Siegel V, Pakhomov E A, et al. 2008. Oceanic circumpolar habitats of Antarctic krill. Mar Ecol Prog Ser, 362: 1-23, doi:10.3354/meps07498.

Atkinson A, Hill S L, Pakhomov E A, et al. 2017. KRILLBASE: a circumpolar database of Antarctic krill and salp numerical densities, 1926–2016. Earth Syst Sci Data, 9(1): 193-210, doi:10.5194/essd-9-193-2017.

Breiman L. 2001. Random Forests. Mach Learn, 45(1): 5-32, doi:10.1023/A:1010933404324.

Brown M, Kawaguchi S, Candy S, et al. 2010. Temperature effects on the growth and maturation of Antarctic krill (*Euphausia superba*). Deep Sea Res Part II Top Stud Oceanogr, 57(7/8): 672-682, doi:10.1016/j.dsr2.2009.10.016.

Burns A L, Schaerf T M, Lizier J, et al. 2022. Self-organization and information transfer in Antarctic krill swarms. Proc R Soc B, 289(1969): 20212361, doi:10.1098/rspb.2021.2361.

Cavan E L, Belcher A, Atkinson A, et al. 2019. The importance of Antarctic krill in biogeochemical cycles. Nat Commun, 10(1): 4742, doi:10.1038/s41467-019-12668-7.

Cavanagh R D, Melbourne-Thomas J, Grant S M, et al. 2021. Future risk for Southern Ocean ecosystem services under climate change. Front Mar Sci, 7: 615214, doi:10.3389/fmars.2020.615214.

CCAMLR. 2022. Report of the 41st Meeting of the Commission for the Conservation of Antarctic Marine Living Resources. Hobart: CCAMLR.

CCAMLR. 2023. Fishery report 2023: *Euphausia superba* in Area 48. (2024-05-03) [2025-08-31]. https://fishdocs.ccamlr.org/FishRep_48_KRI_2023.html.

CCAMLR. 2024. Report of the 43rd Meeting of the Commission for the Conservation of Antarctic Marine Living Resources. Hobart: CCAMLR.

Chen T Q, Guestrin C. 2016. XGBoost: a scalable tree boosting system. San Francisco: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, doi:10.1145/2939672.2939785.

Chen X Z, Zhu G P. 2022. Habitat of Antarctic krill (*Euphausia superba*) in the Bransfield Strait based on ensembled species distribution model. J Fish China, 46(3): 390-401, doi:10.11964/jfc.20211213239 (in Chinese with English abstract).

Chen Z, Chen X Z, Zhu G P. 2025. Ecological niche and habitat use of Antarctic krill (*Euphausia superba*) and the pelagic tunicate (*Salpa thompsoni*) under future climate change: further overlap or separation? Biodivers Conserv, 34(5): 1771-1792, doi:10.1007/s10531-025-03042-9.

Copernicus Marine Service. 2023. Global ocean physics reanalysis (GLORYS12V1). (2023-11-30) [2025-08-31]. doi:10.48670/moi-00021.

Copernicus Marine Service. 2025. Global ocean colour (Copernicus-GlobColour), Bio-Geo-Chemical, L4 (monthly and interpolated) from satellite observations (1997-ongoing). [2025-08-31]. doi:10.48670/moi-00281.

Dong Y F, Liu H, Zhu G P. 2024. Effects of sea ice and Southern Annular Mode on the length-based index of the Antarctic krill (*Euphausia superba*) population in the Bransfield Strait. Antarctic, J Fish Sci China, 30(12): 1543-1555, doi:10.12264/JFSC2023-0266 (in Chinese with English abstract).

Dormann C F, Calabrese J M, Guillera-Arroita G, et al. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol Monogr, 88(4):

485-504, doi:10.1002/ecm.1309.

Dorogush A V, Ershov V, Gulin A. 2018. CatBoost: gradient boosting with categorical features support. arXiv: 1810.11363. https://arxiv.org/abs/1810.11363

Ducklow H W, Baker K, Martinson D G, et al. 2007. Marine pelagic ecosystems: the West Antarctic Peninsula. Philos Trans R Soc Lond B Biol Sci, 362(1477): 67-94, doi:10.1098/rstb.2006.1955.

Fach B A, Hofmann E E, Murphy E J. 2006. Transport of Antarctic krill (*Euphausia superba*) across the Scotia Sea. Part II: krill growth and survival. Deep Sea Res Part I Oceanogr Res Pap, 53(6): 1011-1043, doi:10.1016/j.dsr.2006.03.007.

Freer J J, Warwick-Evans V, Skaret G, et al. 2025. A new dynamic distribution model for Antarctic krill reveals interactions with their environment, predators, and the commercial fishery in the South Scotia Sea region. Limnol Oceanogr, 70(4): 833-849, doi:10.1002/lno.12809.

GEBCO Compilation Group. 2023. GEBCO 2023 Grid. (2023-04-18) [2025-08-31]. doi:10.5285/f98b053b-0cbc-6c23-e053-6c86abc0af7b.

Georganos S, Grippa T, Niang Gadiaga A, et al. 2021. Geographical Random Forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int, 36(2): 121-136, doi:10.1080/10106049.2019.1595177.

Goodwin M, Halvorsen K T, Jiao L, et al. 2022. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. ICES J Mar Sci, 79(2): 319-336, doi:10.1093/icesjms/fsab255.

Guillaumot C, Danis B, Saucède T. 2021. Species distribution modelling of the Southern Ocean benthos: a review on methods, cautions and solutions. Antarct Sci, 33(4): 349-372, doi:10.1017/s0954102021000183.

Hamner W M, Hamner P P. 2000. Behavior of Antarctic krill (*Euphausia superba*): schooling, foraging, and antipredatory behavior. Can J Fish Aquat Sci, 57(S3): 192-202, doi:10.1139/f00-195.

Huneke W G C, Huhn O, Schröeder M. 2016. Water masses in the Bransfield Strait and adjacent seas, austral summer 2013. Polar Biol, 39(5): 789-798, doi:10.1007/s00300-016-1936-8.

Hunt T N, Allen S J, Bejder L, et al. 2020. Identifying priority habitat for conservation and management of Australian humpback dolphins within a marine protected area. Sci Rep, 10(1): 14366, doi:10.1038/s41598-020-69863-6.

Ke G L, Meng Q, Finley T, et al. 2017. LightGBM: a highly efficient gradient boosting decision tree. Long Beach: 31st Conference on Neural Information Processing Systems (NeurIPS), doi:10.5555/3294996.3295074.

Kokonendji C C, Dossou-Gbété S, Demetrio C G B. 2004. Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demetrio classes. Barcelona: Sort, 201-214.

Krafft B A, MacAulay G J, Skaret G, et al. 2021. Standing stock of Antarctic krill (*Euphausia superba* Dana, 1850) (Euphausiacea) in the Southwest Atlantic sector of the Southern Ocean, 2018–19. J Crustacean Biol, 41(3): ruab046, doi:10.1093/jcbiol/ruab046.

Li H J, Wang B, Niu X, et al. 2023. Application of machine learning technology in ecology. Chin J Ecol, 42(11): 2767-2775. doi: 10.13292/j.1000-4890.202311.009 (in Chinese with English abstract).

Li Y J, Li Y, Wei Z X. 2019. Study on the properties and exchanges of water masses in the region of Antarctic Peninsula. Haiyang Xuebao, 41(9): 13-25, doi:10.3969/j.issn.0253−4193.2019.09.002 (in Chinese with English abstract).

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News, 2(3): 18-22.

Lin S Y, Zhao L, Feng J L. 2022. Predicted changes in the distribution of Antarctic krill in the Cosmonaut Sea under future climate change scenarios. Ecol Indic, 142: 109234, doi:10.1016/j.ecolind.2022.109234.

Liu H, Zhu G P. 2022. Scale effect of spatial pattern for Antarctic krill (*Euphausia superba*) distribution around Antarctic Peninsula based on changepoint analysis. J Fish China, 46(3): 359-367, doi:10.11964/jfc.20211213240 (in Chinese with English abstract).

Maravelias C D, Haralabous J, Papaconstantinou C. 2003. Predicting demersal fish species distributions in the Mediterranean Sea using artificial neural networks. Mar Ecol Prog Ser, 255: 249-258, doi:10.3354/meps255249.

Marrari M, Daly K L, Hu C M. 2008. Spatial and temporal variability of SeaWiFS chlorophyll *a* distributions west of the Antarctic Peninsula: Implications for krill production. Deep Sea Res Part II Top Stud Oceanogr, 55(3/4): 377-392, doi:10.1016/j.dsr2.2007.11.011.

Martinson D G, Stammerjohn S E, Iannuzzi R A, et al. 2008. Western Antarctic Peninsula physical oceanography and spatio–temporal variability. Deep Sea Res Part II Top Stud Oceanogr, 55(18/19): 1964-1987, doi:10.1016/j.dsr2.2008.04.038.

Muñoz-Mas R, Gil-Martínez E, Oliva-Paterna F J, et al. 2019. Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.): introducing XGBoost to eco-informatics. Ecol Inform, 53: 100974, doi:10.1016/j.ecoinf.2019.100974.

Murphy E J, Watkins J L, Meredith M P, et al. 2004. Southern Antarctic Circumpolar Current Front to the northeast of South Georgia: horizontal advection of krill and its role in the ecosystem. J Geophys Res Oceans, 109(C1): 2002JC001522, doi:10.1029/2002JC001522.

Murphy E J, Watkins J L, Trathan P N, et al. 2007. Spatial and temporal operation of the Scotia Sea ecosystem: a review of large-scale links in a krill centred food web. Philos Trans R Soc Lond B Biol Sci, 362(1477): 113-148, doi:10.1098/rstb.2006.1957.

Murphy E J, Cavanagh R D, Hofmann E E, et al. 2012. Developing integrated models of Southern Ocean food webs: including ecological complexity, accounting for uncertainty and the importance of scale. Prog Oceanogr, 102: 74-92, doi:10.1016/j.pocean.2012.03.006.

Mustafa Z. 2025. Tweedie distribution: a statistical solution for unusually dispersed data. Sciencestatistics, 3(1): 29-37, doi:10.24127/sciencestatistics.v3i1.8003.

Olden J D, Lawler J J, Poff N L. 2008. Machine learning methods without tears: a primer for ecologists. Q Rev Biol, 83(2): 171-193, doi:10.1086/587826.

Quetin L B, Ross R M. 2003. Episodic recruitment in Antarctic krill *Euphausia superba* in the Palmer LTER study region. Mar Ecol Prog Ser, 259: 185-200, doi:10.3354/meps259185.

R Core Team. 2023. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Reiss C S, Cossio A M, Loeb V, et al. 2008. Variations in the biomass of Antarctic krill (*Euphausia superba*) around the South Shetland Islands, 1996–2006. ICES J Mar Sci, 65(4): 497-508, doi:10.1093/icesjms/fsn033.

Robinson N M, Nelson W A, Costello M J, et al. 2017. A systematic review of marine-based Species Distribution Models (SDMs) with recommendations for best practice. Front Mar Sci, 4: 421,

doi:10.3389/fmars.2017.00421.

Ross R M, Quetin L B, Kirsch E. 1988. Effect of temperature on developmental times and survival of early larval stages of *Euphausia superba* Dana. J Exp Mar Biol Ecol, 121(1): 55-71, doi:10.1016/0022-0981(88)90023-8.

Ryabov A B, de Roos A M, Meyer B, et al. 2017. Competition-induced starvation drives large-scale population cycles in Antarctic krill. Nat Ecol Evol, 1(7): 0177, doi:10.1038/s41559-017-0177.

Schmidt K, Schlosser C, Atkinson A, et al. 2016. Zooplankton gut passage mobilizes lithogenic iron for ocean productivity. Curr Biol, 26(19): 2667-2673, doi:10.1016/j.cub.2016.07.058.

Shono H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fish Res, 93(1/2): 154-162, doi:10.1016/j.fishres.2008.03.006.

Siegel V. 2005. Distribution and population dynamics of *Euphausia superba*: summary of recent findings. Polar Biol, 29(1): 1-22, doi:10.1007/s00300-005-0058-5.

Sinclair S J, White M D, Newell G R. 2010. How useful are species distribution models for managing biodiversity under future climates? Ecol Soc, 15(1): 8.

Suarez-Bregua P, Álvarez-González M, Parsons K M, et al. 2022. Environmental DNA (eDNA) for monitoring marine mammals: Challenges and opportunities. Front Mar Sci, 9: 987774, doi:10.3389/fmars.2022.987774.

Swart S, du Plessis M D, Thompson A F, et al. 2020. Submesoscale fronts in the Antarctic marginal ice zone and their response to wind forcing. Geophys Res Lett, 47(6): e2019GL086649, doi:10.1029/2019GL086649.

Tarling G A, Klevjer T, Fielding S, et al. 2009. Variability and predictability of Antarctic krill swarm structure. Deep Sea Res Part I Oceanogr Res Pap, 56(11): 1994-2012, doi:10.1016/j.dsr.2009.07.004.

Wang S B, Zhuang J Q, Zheng J, et al. 2021. Application of Bayesian hyperparameter optimized random forest and XGBoost model for landslide susceptibility mapping. Front Earth Sci, 9: 712240, doi:10.3389/feart.2021.712240.

Wang X L, Zhang J C, Yu X T, et al. 2025. Inter-annual variation of Antarctic krill (*Euphausia superba*) density around the South Shetland Islands during 2013–2019. ICES J Mar Sci, 82(8): fsaf083, doi:10.1093/icesjms/fsaf083.

Warwick-Evans V, Fielding S, Reiss C S, et al. 2022. Estimating the average distribution of Antarctic krill *Euphausia superba* at the northern Antarctic Peninsula during austral summer and winter. Polar Biol, 45(5): 857-871, doi:10.1007/s00300-022-03039-y.

Wiens J A. 1989. Spatial scaling in ecology. Funct Ecol, 3(4): 385, doi:10.2307/2389612.

Wolpert D H. 1992. Stacked generalization. Neural Netw, 5(2): 241-259, doi:10.1016/S0893-6080(05)80023-1.

Wood S N. 2017. Generalized Additive Models: an introduction with R. New York: Chapman and Hall/CRC, doi:10.1201/9781315370279.

Wood S N. 2020. Inference and computation with Generalized Additive Models and their extensions. TEST, 29(2): 307-339, doi:10.1007/s11749-020-00711-5.

Zhao Y X, Wang X L, Ying Y P, et al. 2025a. Research progress on the impact of environmental factors on Antarctic krill resources and fishery dynamic. J Fish Sci China, 2025, 32(2): 258-275, doi:10.12264/JFSC2024-0285 (in Chinese with English abstract).

Zhao Y X, Wang X L, Zhang J C, et al. 2025b. Seasonal dynamics of an Antarctic krill resource hot spot in the Bransfield Strait, Antarctica. Mar Coast Fish, 17(3): vtaf012, doi:10.1093/mcfafs/vtaf012.