ICES Journal of Marine Science, 2025, Vol. 82, Issue 1, fsaf003
https://doi.org/10.1093/icesjms/fsaf003
Received: 16 September 2024; **revised:** 16 December 2024; **accepted:** 6 January 2025
Original Article

An integrative machine learning approach to understanding South Pacific Ocean albacore tuna habitat features

Liwen Liu[1], Rong Wan[1,2,3], Feng Wu[1,2,3], Yucheng Wang[1], Yonghan Zhu[4], Cheng Zhou[1,2,3,*]

[1]College of Marine Living Resource Sciences and Management, Shanghai Ocean University, Shanghai 201306, P. R. China
[2]National Engineering Research Center for Oceanic Fisheries, Shanghai 201306, P. R. China
[3]Key Laboratory of Exploitation of Ocean Fisheries Resources, Ministry of Agriculture and Rural Affairs, Shanghai Ocean University, Shanghai 201306, P. R. China
[4]Zhongxing Digital Marine Technologies Co., Ltd, Shenzhen 518000, P. R. China

*Corresponding author. College of Marine Living Resource Sciences and Management, Shanghai Ocean University, 999 Huchenghuan Road, Lingang New District, Shanghai 201306, P. R. China. E-mail: c-zhou@shou.edu.cn

Abstract

This study employs a random forest model combined with interpretable machine learning techniques to analyze the habitat preferences of South Pacific albacore tuna, incorporating a broad range of marine environmental variables. Among these, several factors derived from mesoscale eddy structures, including eddy polarity, eddy radius, and eddy kinetic energy, are integrated to further enhance the characterization of mesoscale eddy features. Interpretable methods were applied to provide intuitive visualizations of albacore tuna habitat preferences, with a focus on the most influential factors, including seawater temperature, dissolved oxygen concentration, and normalized mesoscale eddy radius. Seawater temperature and oxygen concentration are directly linked to the physiological needs of albacore tuna, while mesoscale eddy characteristics influence foraging and behavior by altering water column properties. This study provides a comprehensive perspective on the characteristics of albacore tuna habitat and the mechanisms driving its oceanographic variables, providing valuable insights for developing location-based, practical science-based management strategies for fishery resources.

Keywords: South Pacific Ocean; albacore tuna; mesoscale eddy; interpretable machine learning; SHapley additive explanations

Introduction

Albacore tuna (*Thunnus alalunga*) is an economically important migratory species and a key target for longline fisheries in the South Pacific (Nikolic et al. 2017). However, increasing fishing pressure and environmental changes pose challenges to its sustainable management (Pauly et al. 2002). The 2024 South Pacific albacore tuna stock assessment suggests that the stock is not currently overfished but is experiencing increasing fishing pressure (Teears et al. 2024). In response to these pressures, future management strategies are likely to include quota-based systems to ensure the sustainable exploitation of albacore tuna. Therefore, gaining a comprehensive understanding of the spatial distribution and environmental drivers behind the formation of albacore tuna fishing grounds is crucial for the development of more effective fishing strategies and the advancement of fisheries science (Sund et al. 1981).

Albacore tuna is distributed across a wide range of geographical regions and ocean depths, spanning both mid-to-high latitudes in the Northern and Southern Hemispheres. Their vertical range extends from the ocean surface to depths of up to 300 meters, and they can even reach greater depths under certain conditions (Bertrand et al. 2002, Zainuddin et al. 2008, Williams et al. 2015). The distribution of albacore tuna is influenced by a complex interplay of environmental factors, including sea surface temperature (SST), salinity, dissolved oxygen concentration, and chlorophyll-a levels. These factors interact dynamically to shape their habitat preferences and distribution patterns (Mohri et al. 1999, Chen et al. 2005, Zainuddin et al. 2006, Zainuddin et al. 2008). The distribution area of South Pacific albacore tuna coincides with regions characterized by stable, active mesoscale eddy structures. Indeed, mesoscale eddies play a unique and important role in shaping the habitat of albacore tuna (Wang et al. 2003, Hu et al. 2014, Xu et al. 2017, Zhou et al. 2020). These mesoscale phenomena, commonly occurring in mid-latitude regions, significantly impact pelagic ecosystems across multiple trophic levels. Eddies have the potential to enhance productivity in nutrient-poor areas, aggregate prey, and alter habitat conditions within the water column (Arostegui et al. 2022). Such physical and biological interactions create favorable conditions for large pelagic predators. For instance, the aggregation of small swimming organisms around eddies often forms prey hotspots, attracting predators (Godø et al. 2012). Despite these insights, it remains unclear whether these interactions affect different pelagic predator species in the same way or if predators' responses to eddies are consistent across various productive regions. Understanding how these complex and dynamic marine environmental factors interact to influence albacore tuna habitats is therefore critical (Pitcher et al. 2001). Habitat models are essential tools for analyzing spa-

tial distribution patterns of tuna by examining the relationship between species occurrence and marine environmental factors, which help identify habitat preferences (Handegard et al. 2013, Arrizabalaga et al. 2015). Tuna distribution is influenced by environmental factors such as dissolved oxygen concentration, salinity, chlorophyll-a concentration, and SST (Maury et al. 2010, Lopez et al. 2017). Forecasting models for fishing grounds rely on fish population dynamics and employ methods such as statistical regression, artificial neural networks, time series analysis, and machine learning. Machine learning models, particularly supervised learning methods, have become prominent in habitat models. In habitat models, these models build predictive frameworks by analyzing historical data in relation to environmental factors (Elith et al. 2006). Generalized linear models (GLMs) extend traditional linear regression to analyze the relationship between response variables and explanatory factors. Gavaris et al. (1980) pioneered the use of GLMs to standardize catch per unit effort (CPUE) data, addressing variations due to fishing gear and correcting nominal CPUE data to remove effects not related to the abundance of the species, hence improving its use as an index of species abundance. GLM have also been used in combination with additional spatial and temporal factors, such as SST and Southern Oscillation Index (SOI) to analyze the interannual variation in tropical tuna distribution and abundance (Okamoto et al. 2001, Zainuddin et al. 2017). However, CPUE often demonstrates complex, nonlinear relationships with influencing factors, which GLMs may not capture effectively. Generalized Additive Models (GAMs) address this limitation by incorporating nonlinear functions, providing a more accurate representation of the relationship between tuna abundance and environmental factors, and assessing their relative importance. Studies by Zagaglia et al. (2004) and Setiawati et al. (2015) utilized GAMs to analyze tropical tuna habitat distribution, revealing constraints imposed by the intertropical convergence zone and influences of environmental factors such as SST and chlorophyll concentration. As the dimensionality of data has increased, traditional statistical models have become insufficient for capturing the complexity of these datasets. The advent of machine learning has significantly addressed this challenge. Random forest (RF) has been employed to analyze the vertical distribution of albacore tuna (Mondal et al. 2022), while K-Nearest Neighbors (KNN) and Support Vector Machines (SVMs) have been used to identify potential fishing zones and patterns in tuna distribution (Fitrianah et al. 2015, Yang et al. 2020). These models are evaluated through rigorous cross-validation and comparative analysis of predictive performance using metrics such as accuracy, precision, recall, aiming to enhance habitat prediction accuracy.

Many studies that use machine learning to predict albacore tuna fishing grounds have been limited by narrow temporal and spatial data coverage. Typically, these studies do not account for the vertical distribution of species such as albacore tuna, which exhibit significant vertical movement. To address these gaps, our study incorporates a comprehensive dataset spanning 6 years, from 2017 to 2022, and includes environmental factors from the ocean surface down to depths of 300 m. This extensive dataset strengthens our understanding of albacore tuna distribution patterns. Furthermore, we introduce a predictor by incorporating the distance from the fishing vessel to the mesoscale eddy center, normalized by the eddy radius. This approach addresses the limitations of models, which often rely solely on eddy kinetic energy (EKE) to represent the influence of mesoscale eddies on albacore tuna distribution, without accounting for other important factors such as eddy radius. Machine learning models in this field have often focused primarily on improving prediction accuracy without addressing model interpretability. These models, while effective, often operate as "black box" and make it difficult to understand why certain environmental factors are influential. To overcome this limitation, our study employs advanced interpretability techniques, including SHAP (SHapley Additive exPlanations), Partial Dependence Plots (PDPs), and Individual Conditional Expectation (ICE) plots. This study aims to investigate the role of various marine environmental factors in predicting albacore tuna fishing grounds using the RF algorithm. We aim to understand how temperature, dissolved oxygen, and mesoscale eddies influence the distribution of albacore tuna. We propose that the edges of mesoscale eddies, characterized by features such as eddy radius and EKE, may be particularly attractive to albacore tuna. By integrating RF with SHAP, PDP, and ICE plots, this research seeks to explore the relative importance of these environmental factors, offering insights that could support ecological studies and management strategies for albacore tuna conservation.

Materials and methods

Study area overview

The study area spans from 160°E to 80°W longitude, covering the region from the equator (0°) down to 40°S latitude. This vast oceanic expanse is well-known for its rich fisheries and diverse ecosystems, which are vital for sustaining biodiversity and ecological equilibrium. The presence of numerous islands within this region adds to its ecological complexity and richness. Moreover, this area is a significant fishing ground for various tuna species, such as bigeye tuna, skipjack tuna, albacore tuna, and yellowfin tuna.

Data sources

Fishery data

Due to restrictions on data access, this study is confined to the fishing logbooks records of Chinese industrial distant-water longline vessels. The data collected between 2017 and 2022 were provided by the Oceanic Data Center at Shanghai Ocean University. We extracted records in which the number of albacore tuna individuals accounted for >80% of the total catch by individual count. The dataset includes vessel identifiers, longitude and latitude coordinates, operation dates, catch quantities (in number), hook counts, which form a robust foundation for our analysis.

Marine environmental data

A total of 36 environmental factors were utilized in this study, encompassing a broad range of variables. These include 27 environmental parameters sourced from the Copernicus Marine Environment Monitoring Service (http://marine.copernicus.eu), which are as follows: Chlorophyll-a concentration (CHL), Sea Water Salinity (SO), Seawater Temperature (THETAO), and Dissolved Oxygen (O2) at six different depths (0, 50, 100, 150, 200, and 300 m), along with Mixed Layer Depth (MLD), Sea Surface Height (SSH), and

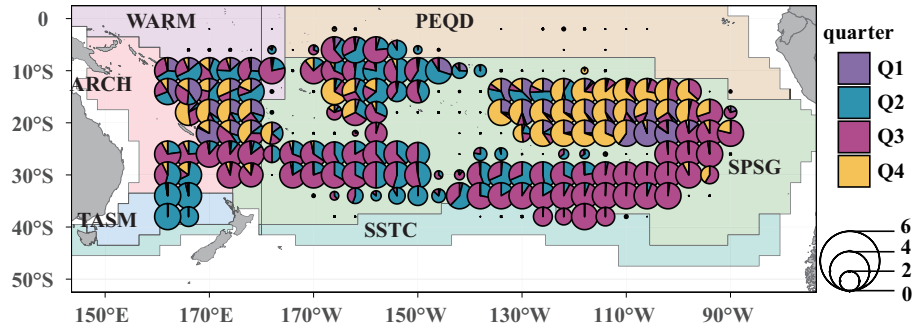

Figure 1. Distribution of albacore tuna CPUE across four quarters, overlaid on six distinct ecological zones. These zones are labeled as ARCH (Archipelagic Deep Basins Province), PEQD (Pacific Equatorial Divergence Province), SPSG (South Pacific Subtropical Gyre Province), SSTC (Southern Subtropical Convergence Province), TASM (Tasman Sea Province), and WARM (Western Pacific Warm Pool Province).

Table 1. Abbreviations and descriptions of oceanic environmental variables.

Abbreviation	Description
MLD	Mixed layer depth, indicating the vertical mixing zone of the ocean
SSH	Sea surface height, representing the sea level height anomalies
SO	Sea water salinity at six different depths (0, 50, 100, 150, 200, 300 m)
THETAO	Seawater temperature at six different depths (0, 50, 100, 150, 200, 300 m)
CHL	Chlorophyll-a concentration at six different depths (0, 50, 100, 150, 200, 300 m)
O2	Dissolved oxygen at six different depths (0, 50, 100, 150, 200, 300 m)
EKE	Eddy kinetic energy
INSIDE	Whether the latitude and longitude of the operation are within 2.5 times the radius of the cyclone and anticyclone, represented by 0 for "no" and 1 for "yes"
NECR	(Normalized distance to the cyclone eddy core radius) is the distance from the cyclone to the operating vessel divided by the cyclone radius
NEAR	(Normalized distance to the anticyclone eddy core radius) is the distance from the anticyclone to the operating vessel divided by the anticyclone radius
ARCH	Trades—Archipelagic Deep Basins Province
PEQD	Trades—Pacific Equatorial Divergence Province
SPSG	Westerlies—S. Pacific Subtropical Gyre Province
SSTC	Westerlies—S. Subtropical Convergence Province
TASM	Westerlies—Tasman Sea Province
WARM	Trades—W. Pacific Warm Pool Province

EKE. These parameters are provided as daily observations with spatial resolutions ranging from a finer resolution of 0.083° × 0.083° (~8.3 km) to a coarser resolution of 0.25° × 0.25° (~25 km). And, the study includes two factors related to mesoscale vortices: cyclones and anticyclones, with data generated based on the methodology and open-source code from Dong et al. (2022). Additionally, we incorporated ecological zone identifiers based on the latitude and longitude from the fishing logbooks records to determine the ecological zone in which each fishing operation took place. As shown in Fig. 1, six distinct zones were identified: Trades—Archipelagic Deep Basins Province (ARCH), Trades—Pacific Equatorial Divergence Province (PEQD), Westerlies—S. Pacific Subtropical Gyre Province (SPSG), Westerlies—S. Subtropical Convergence Province (SSTC), Westerlies—Tasman Sea Province (TASM), and Trades—W. Pacific Warm Pool Province (WARM). These zones were defined according to the prevailing role of physical forcing in regulating phytoplankton distribution (Longhurst et al. 1995), with the data sourced from Marine Regions (https://www.marineregions.org). Given that these are factor-type variables, we applied one-hot encoding to integrate them into the dataset. The abbreviations and brief descriptions of the oceanic environmental variables used in this study are summarized in Table 1.

Data processing

Data extraction

Given that the catch log dataset spans 6 years, it inevitably contains missing values, outliers, and some records that do not meet the criteria for accurate modeling. Therefore, a thorough data extraction and cleaning process was conducted to address these issues and ensure the dataset's reliability and suitability for analysis. Following this preprocessing, a refined dataset was extracted, which was suitable for integration with fishery environmental factors. This cleaned dataset was chosen as the core data for the modeling process. It includes detailed records of key parameters such as geographic coordinates (longitude and latitude), timestamp, weight of albacore tuna, number of albacore tuna caught, and number of hooks used. By utilizing date, longitude, and latitude data, we can relate these with marine environmental factors. Traditional approaches to EKE focus primarily on quantifying the intensity of mesoscale variability, yet they overlook the spatial extent of mesoscale eddies. To address this limitation, our study incorporated the quantification of mesoscale eddy radii. This quantification formed the basis of an enhanced framework, enabling us to develop a more comprehensive method for assessing the impact of mesoscale eddies on fishing activities, as illustrated in Fig. 2. We identified the cyclone and anticyclone

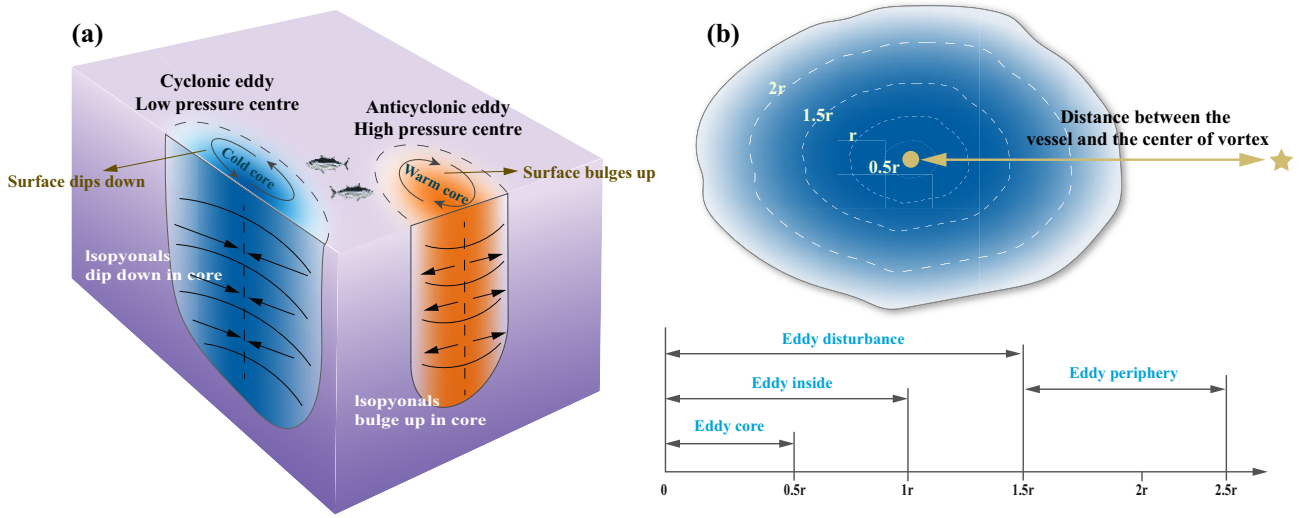

Figure 2. (a) Cyclonic vortices are typically characterized by the upward displacement of isopycnals and isothermal surfaces, leading to lower water temperatures in the central region of the vortex compared to the surrounding sea areas. Consequently, cyclonic vortices are often referred to as cold vortices. In contrast, anticyclonic vortices display the opposite characteristics, with elevated temperatures in their central regions. (b) This figure illustrates the regions defined by normalized radii: We divide the eddy from inside to outside into eddy core (0–0.5r), eddy inside (0–1r), eddy disturbance (0–1.5r), and eddy periphery (1.5–2.5r). The positions of the fishing vessel and the vortex core are indicated by small stars and circles, respectively.

closest to the fishing operation, based on both temporal and spatial proximity to the fishing vessel's coordinates and operational time. After this identification, we extracted the radius and center point of each eddy and calculated the distance from the fishing vessel to the center of both the cyclone and the anticyclone. To assess the proximity of the fishing vessel to each eddy, we computed the distance ratio by dividing the distance from the vessel to the eddy center by the radius of the respective eddy. These distance ratios, which reflect the intensity of the mesoscale eddy's influence on the fishing vessel, were incorporated as marine environmental factors in our analysis. The formula used to calculate this distance ratio is as follows:

$$NECR(NEAR) = \frac{\arccos(\sin(VESSEL_{lat}) \times \sin(EDDY_{lat}) + \cos(VESSEL_{lat}) \times \cos(EDDY_{lat}) \times \cos(EDDY_{lon} - VESSEL_{lon})) \times R}{EDDY_r}. \quad (1)$$

Note: The latitude of $VESSEL_{lat}$ represents the latitude of the fishing vessel during operations, while $EDDY_{lat}$ denotes the latitude of the centroid of mesoscale eddies. $VESSEL_{lon}$ corresponds to the longitude of the fishing vessel, and $EDDY_{lon}$ indicates the longitude of the centroid of mesoscale eddies. $EDDY_r$ signifies the radius of the mesoscale eddy, and R denotes the Earth's radius. *NECR* (Normalized distance to the cyclone eddy core radius) is the distance from the cyclone to the operating vessel divided by the cyclone radius. *NEAR* (Normalized distance to the anticyclone eddy core radius) is the distance from the anticyclone to the operating vessel divided by the anticyclone radius.

Calculation of CPUE

In this study, CPUE was computed as the ratio of catch to effort for each fishing operation based on the exact latitude and longitude of the fishing event without aggregating them into a grid. This approach allows for a high level of spatial resolution and avoids any loss of detail that might occur from grid aggregation. The target variable for classification was CPUE. This study classified fisheries using five different thresholds (30%, 40%, 50%, 60%, and 70%). For each threshold, binary classification was conducted using various models for analysis. Areas with CPUE values greater than the threshold were classified as high-yield fishing grounds, while areas with CPUE values below the threshold were categorized as low-yield fishing grounds. The formula used to calculate CPUE as follows:

$$CPUE_{ymdijv} = \frac{C_{ymdijv}}{H_{ymdijv}} \times 1000 \quad (2)$$

Note: Where $CPUE_{ymdijv}$, C_{ymdijv}, H_{ymdijv} represent the CPUE (calculated as the number of albacore tuna caught per 1 000 hooks), total individual catch of albacore tuna, and the total number of hooks deployed, respectively, for y year, m month, d day, longitude i, latitude j, and *vessel* v.

Model development process

Methodological framework

This study followed a systematic workflow to develop a robust predictive model by integrating fisheries logbook data and environmental factors through rigorous data cleaning and matching. After defining CPUE thresholds, we explored several machines learning algorithms, including RF, KNN, SVM, Neural Networks (NNET), and Recursive Partitioning and Regression Trees (RPART), with RF being the primary model and the others serving as control models for comparison. To ensure prediction reliability, we implemented five-fold cross-validation, where the dataset was split into training and test subsets in various configurations. This allowed for a comprehensive comparison of model performance across different

thresholds and parameters. The optimal model was selected based on performance metrics, followed by detailed analysis using SHAP-based methods to interpret the contribution of environmental factors (e.g. dissolved oxygen, chlorophyll, and temperature) to the model's predictions, providing a clear and interpretable understanding of the results. As shown in Fig. 3.

Five-fold cross-validation

To systematically assess model performance and identify the best-performing model, we utilized five-fold cross-validation. This method strikes a good balance between model evaluation accuracy, computational efficiency, and data utilization. It provides reliable performance estimates while avoiding excessive computational resource consumption. This method involves dividing the training dataset into five stratified folds. Each model is trained on four of these folds and validated on the remaining fold iteratively. This process is repeated five times to ensure a robust evaluation across different data subsets. Performance metrics such as accuracy, balanced accuracy, precision, recall, F1 score, and ROC curve are calculated for each validation iteration, providing a comprehensive assessment of each model's performance.

Hyperparameter tuning

For each model, hyperparameters are optimized to enhance performance. Techniques such as grid search and random search are utilized to examine various hyperparameter combinations. Subsequently, cross-validation is employed to assess the performance of each combination and identify the optimal set of hyperparameters. This approach ensures that each model achieves optimal efficiency and generalization. Systematic hyperparameter tuning and grid search were conducted to determine the best hyperparameters for each model, as presented in Table 2.

Hyperparameters are critical settings that dictate the training process of machine learning models. Unlike parameters, which are learned during the training phase, hyperparameters are predefined before training begins and can significantly impact model performance. For the models analyzed in this study, the following hyperparameters were employed: RPART : The complexity parameter ($cp = 0.02264251$) controls the size of the decision tree by pruning branches that do not substantially improve model performance, thereby reducing overfitting; NNET: The size parameter, set to 5, specifies the number of neurons in the hidden layer, affecting the network's ability to learn complex patterns. The decay parameter, set to 0.1, acts as a regularization term to penalize large weights and mitigate overfitting; RF: The mtry hyperparameter, set to 40, determines the number of variables sampled at each split, influencing the randomness and diversity of the trees within the forest; SVM: The sigma parameter, set to 0.02002539, defines the kernel width for the radial basis function, affecting the influence of individual training examples. The C parameter, set to 1, serves as a regularization term that balances the trade-off between margin maximization and classification error minimization; KNN: The k parameter, set to 9, specifies the number of nearest neighbors considered during prediction, with a higher k value leading to greater smoothing of predictions.

Additionally, a common hyperparameter across all models is k-fold $= 5$, which refers to 5-fold cross-validation. In this process, the dataset is divided into five subsets. The model is trained on four of these subsets and validated on the remaining subset. This procedure is repeated five times, allowing for a more robust evaluation of performance metrics across all subsets. The choice of 5-fold cross-validation strikes a balance between computing time, dataset size, and model performance, offering a good compromise between model evaluation accuracy and computational efficiency.

Comprehensive analysis of model performance indicators

Machine learning evaluation metrics typically encompass five key indicators: accuracy, precision (P), recall (R), F1 score (F1), and balanced accuracy (BA). The formulas for calculating these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \tag{3}$$

$$\text{BalanceAccuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right), \tag{4}$$

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{5}$$

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{6}$$

$$\text{F1} = \frac{2\text{PR}}{\text{P} + \text{R}} \times 100\%. \tag{7}$$

Note: True Positive (TP): The number of samples where both the actual value and the predicted value are positive; True Negative (TN): The number of samples where both the actual value and the predicted value are negative; False Positive (FP): The number of samples where the actual value is negative but the predicted value is positive, indicating an incorrect prediction of negative instances as positive; False Negative (FN): The number of samples where the actual value is positive but the predicted value is negative, indicating an incorrect prediction of positive instances as negative.

SHAP, PDP, and ICE

SHAP values explain the contribution of each feature to individual predictions by assigning importance scores based on cooperative game theory. This approach offers both global and local interpretations of the model. Global SHAP provides an overview of the significance of each feature across the entire dataset, while local SHAP reveals how each feature impacts specific predictions. The SHAP summary plot illustrates the overall importance of each feature in the dataset. The x-axis represents the SHAP values, where values farther from 0 indicate a greater contribution of the feature to the model's output. Each feature can have both positive and negative impacts on the predictions, and these features are listed along the y-axis. Each point in the plot represents a SHAP value for a specific prediction and feature. Areas with denser points indicate a higher sample density for that feature, forming thicker regions. The color of the points reflects the feature values: red indicates higher feature values, while blue indicates lower feature values. By analyzing the distribution of red and blue points, we can infer the directional impact of each feature on the predictions. If red points are primarily concentrated on the right side of the x-axis (positive SHAP values), it indicates that higher feature values have a positive impact on the predictions, pushing the predicted value higher. Conversely, if blue points are concentrated on the right side, it suggests that lower feature values positively contribute to the predictions. If red points

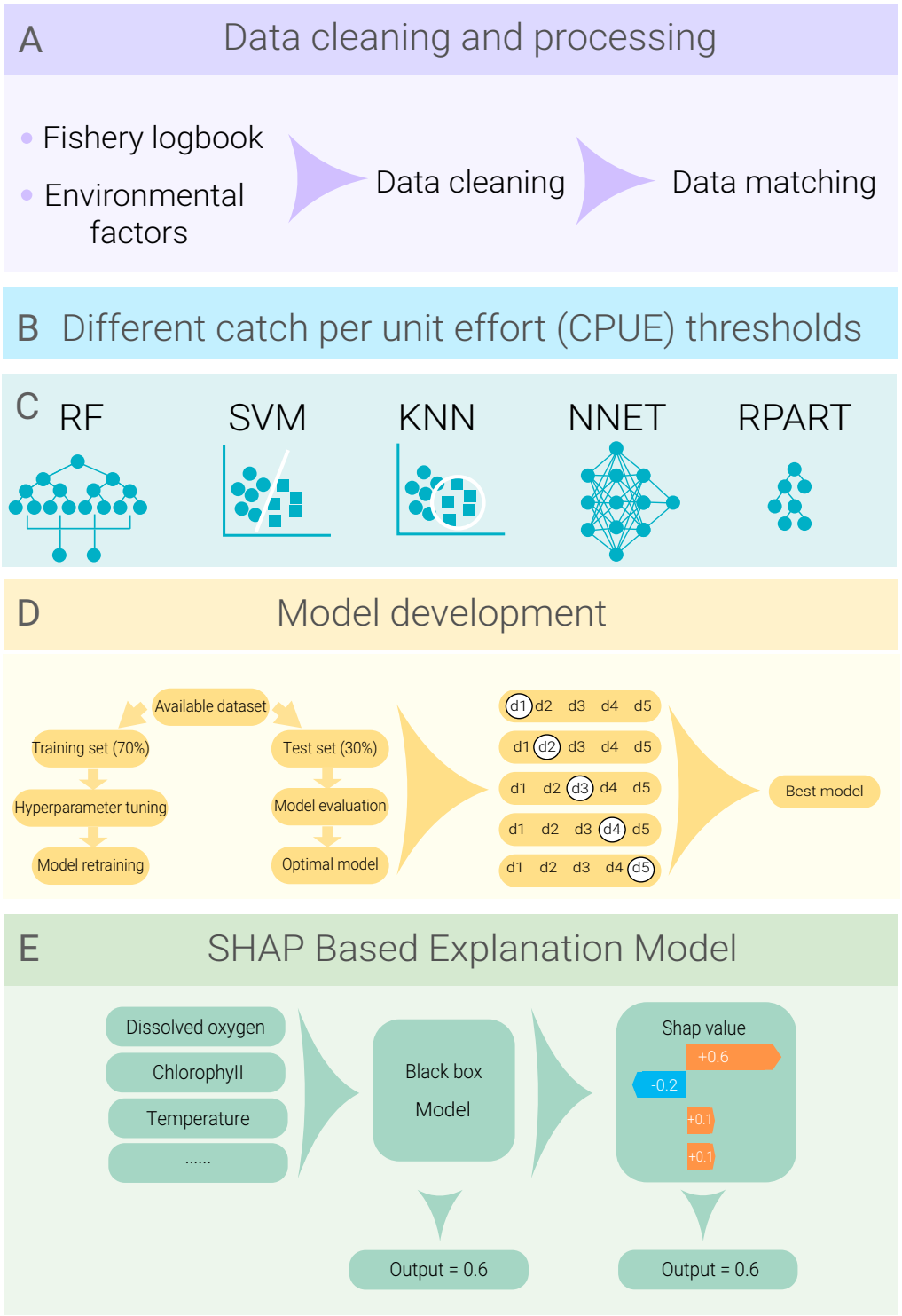

Figure 3. The training framework for the model involves partitioning the original dataset into a training set and a test set, with 70% allocated for training and 30% for testing. The training set, denoted as D, is further divided into five equally sized subsets: d1, d2, d3, d4, and d5. Each model—K-Nearest Neighbors (KNN), Random Forest (RF), Recursive Partitioning and Regression Trees (RPART), Support Vector Machine (SVM), and Neural Network (NNET)—is trained iteratively using these subsets. In each iteration, four subsets are utilized for training while the remaining subset is used as the test set. Hyperparameters and grid search parameters are continuously optimized to determine the best-performing model.

are concentrated on the left side of the *x*-axis (negative SHAP values), it means that higher feature values negatively impact the predictions, decreasing the predicted value. Similarly, blue points concentrated on the left side suggest that lower feature values have a negative impact on the predictions.

ICE plots display a separate line for each instance, showing how the instance's prediction changes when a specific feature is varied. PDPs, which illustrate the average effect of a feature, are a global method because they focus on the overall average, rather than on specific instances. The counterpart

Table 2. Best tune for each model.

Model	Hyperparameters
RPART	*cp* = 0.02264251; *k*-fold = 5
NNET	size = 5; decay = 0.1; *k*-fold = 5
RF	mtry = 24; *k*-fold = 5
SVM	sigma = 0.02002539; C = 1; *k*-fold = 5
KNN	*k* = 9; *k*-fold = 5

to a PDP for individual data instances is the ICE plot (Goldstein et al. 2017). The functions used to generate the ICE curves, PDP curves, and SHAP summary plots in this study are all from the h2o package in R (for details, please refer to https://github.com/h2oai). An ICE plot visualizes how the prediction depends on a feature for each individual instance, resulting in one line per instance, in contrast to the single line representing the overall average in a PDP. A PDP is essentially the average of the lines in an ICE plot. To generate the values for each line (and instance), all other features are held constant, and the feature of interest is replaced with values from a grid, followed by making predictions for these newly created instances using the black-box model.

Results

Performance comparison across CPUE thresholds

Comparison of Receiver Operating Characteristic across five models

The Receiver Operating Characteristic (ROC) curve is a critical tool for evaluating the performance of classification models, demonstrating the trade-offs between sensitivity and specificity across various threshold levels. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1–specificity) at different thresholds, providing a comprehensive assessment of a model's ability to discriminate between positive and negative classes.

Among the models analyzed, the RF model demonstrated the highest performance. Its ROC curves, shown in Figs 4a, b, c, d, and e, are positioned near the upper left corner, indicating a high true positive rate and a relatively low false positive rate across various thresholds. Additionally, the AUC values further validate the RF model's exceptional performance, reflecting its superior discriminative ability and robustness. In comparison, the KNN and SVM models also demonstrated strong performance, with their ROC curves partially overlapping. A closer examination of their AUC values reveals that KNN slightly outperforms SVM. Both models achieved high AUC scores, indicating their ability to maintain a high true positive rate while keeping the false positive rate low. These results confirm that both KNN and SVM are reliable classifiers, with robust discriminative capabilities. The NNET model and the recursive partitioning model (RPART) demonstrated the weakest performance among the models evaluated. Both models struggled to effectively balance sensitivity and specificity, which hindered their classification effectiveness. Their results reflect the poorest discriminative power, which limits their usefulness when compared to other models.

In summary, the comparative analysis shows that the RF model is the most robust classifier, exhibiting the highest sensitivity, the lowest false positive rate, and the highest AUC values across all CPUE thresholds.

Comparison of key metrics across CPUE thresholds

In the comparative analysis of five machine learning models under varying CPUE thresholds (30th percentile, 40th percentile, 50th percentile, 60th percentile, and 70th percentile), significant differences were observed across key metrics, including accuracy, recall, balanced accuracy, F1 score, and precision. The sunburst charts effectively visualize these variations through color gradients, where dark red indicates superior performance, while dark green signifies weaker performance.

As shown in Fig. 5, the RF model consistently exhibited the most robust and stable performance across all CPUE thresholds, with its metrics predominantly shown in dark red. It demonstrated exceptional results at every threshold, achieving high accuracy, recall, balanced accuracy, and F1 scores with minimal variability between thresholds. The KNN model showed moderate performance, at the 30th, 40th, and 70th percentiles, where it was represented in lighter red to orange, indicating acceptable but less consistent performance compared to RF. At the 50th and 60th percentiles, the model's performance declined, with more green tones emerging, reflecting increased variability and reduced reliability. The NNET and RPART models exhibited the weakest and least stable performance. Their metrics were largely represented in green or dark green across all CPUE thresholds, indicating significant declines in accuracy, recall, and precision. These models showed low reliability and substantial variability across thresholds.

Overall, the RF model emerged as the most stable and high-performing option across all CPUE thresholds, consistently represented in dark red across all metrics. In contrast, KNN and SVM demonstrated moderate, threshold-dependent performance, while NNET and RPART exhibited notable weaknesses, particularly at higher fishing intensities. This analysis confirms the RF model as the optimal choice for CPUE classification under varying conditions.

Variable importance analysis of the best model

In the variable importance analysis presented in Fig. 6, thetao_50 (seawater temperature at 50 meters depth) emerged as the most influential predictor with an absolute importance score of 100. This highlights the primary role of temperature in the prediction process, particularly in mid-shallow water layers. o2_0 (surface dissolved oxygen concentration) ranked second with an importance score of 80.35, further emphasizing the critical role of dissolved oxygen in the model. thetao_0 (surface seawater temperature) ranked third with a score of 61.14, demonstrating the significant influence of surface temperature on the target variable. o2_50 (dissolved oxygen at 50 meters) followed with a score of 59.22, reinforcing the importance of seawater temperature and dissolved oxygen as key environmental factors for distribution predictions, particularly in surface and mid-shallow waters. Additionally, NECR and NEAR showed notable importance, scoring 53.56 and 52.84, respectively. Other secondary but noteworthy variables include o2_100 (dissolved oxygen at 100 m, 47.39), MLD (46.48), thetao_100 (temperature at 100 m, 46.38), and so_100 (salinity at 100 m, 45.35). While their influence is comparatively lower, they still enhance the model's explanatory power. Overall, temperature, dissolved oxygen, NECR, and NEAR emerge as the most critical environmental factors in this analysis.

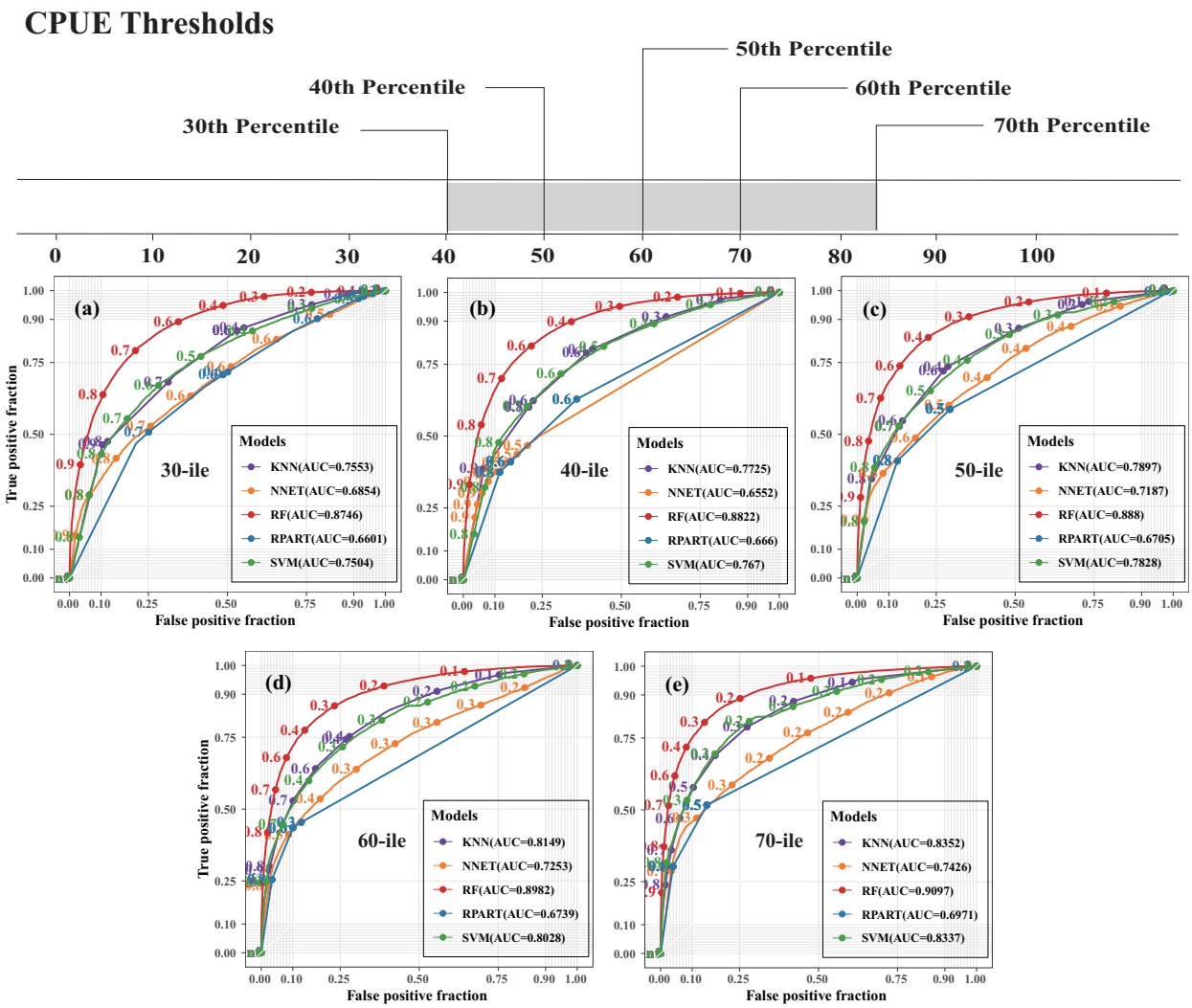

Figure 4. Comparison of ROC curves of machine learning models under different CPUE thresholds. The gray bar at the top represents the range of CPUE values corresponding to the number of individuals. 30-ile. The 30-ile represents the 30th percentile, the 40-ile represents the 40th percentile, the 50-ile represents the 50th percentile, the 60-ile represents the 60th percentile, the 70-ile represents the 70th percentile.

SHAP summary plot of key oceanographic features

As shown in Fig. 7, the points for o2_0, o2_50, thetao_50, and thetao_0 are relatively dispersed, indicating that these four features contribute significantly to the model's output. The distribution of points for o2_0 and o2_50 exhibits a consistent color trend, with blue points concentrated on the left side and red points on the right. This suggests that lower dissolved oxygen values negatively impact the predictions, while higher dissolved oxygen values positively contribute to the predictions. In contrast, the color distribution for thetao_0 and thetao_50 is the opposite of that for dissolved oxygen. Blue points are concentrated on the right, while red points are on the left, indicating that lower temperatures are associated with favorable habitat conditions in this dataset.

For the features EKE, INSIDE, NECR, and NEAR, all points—regardless of whether they are red (high values) or blue (low values)—are concentrated on the right side of the *x*-axis. This indicates that these features consistently have a positive impact on the predictions. The color gradient for INSIDE transitions from blue to red, indicating that contributions within eddy regions have a greater positive impact on the predictions compared to those outside the eddy regions. In contrast, the points for EKE, NECR, and NEAR show a gradient from red to blue, suggesting that smaller values of these features have a stronger positive influence on the predictions. These findings demonstrate that eddy regions, particularly their inner areas, have a greater positive impact on the predictions.

SHAP summary plot of key ecological zone

As shown in Fig. 8, among the six regions, only the WARM region has red points distributed on the left side of the *x*-axis, while points from the other regions are all concentrated on the right side. This indicates that being either within or outside an ecological region generally has a positive impact on the predictions. In terms of color variation, only the ARCH region exhibits a clear transition from blue to red, with red points positioned farther from the *x*-axis, suggesting that the ARCH region has a significant positive influence on the predictions.

For PEQD, the pattern reveals a red-to-blue-to-red transition, indicating that being within or outside this region typ-

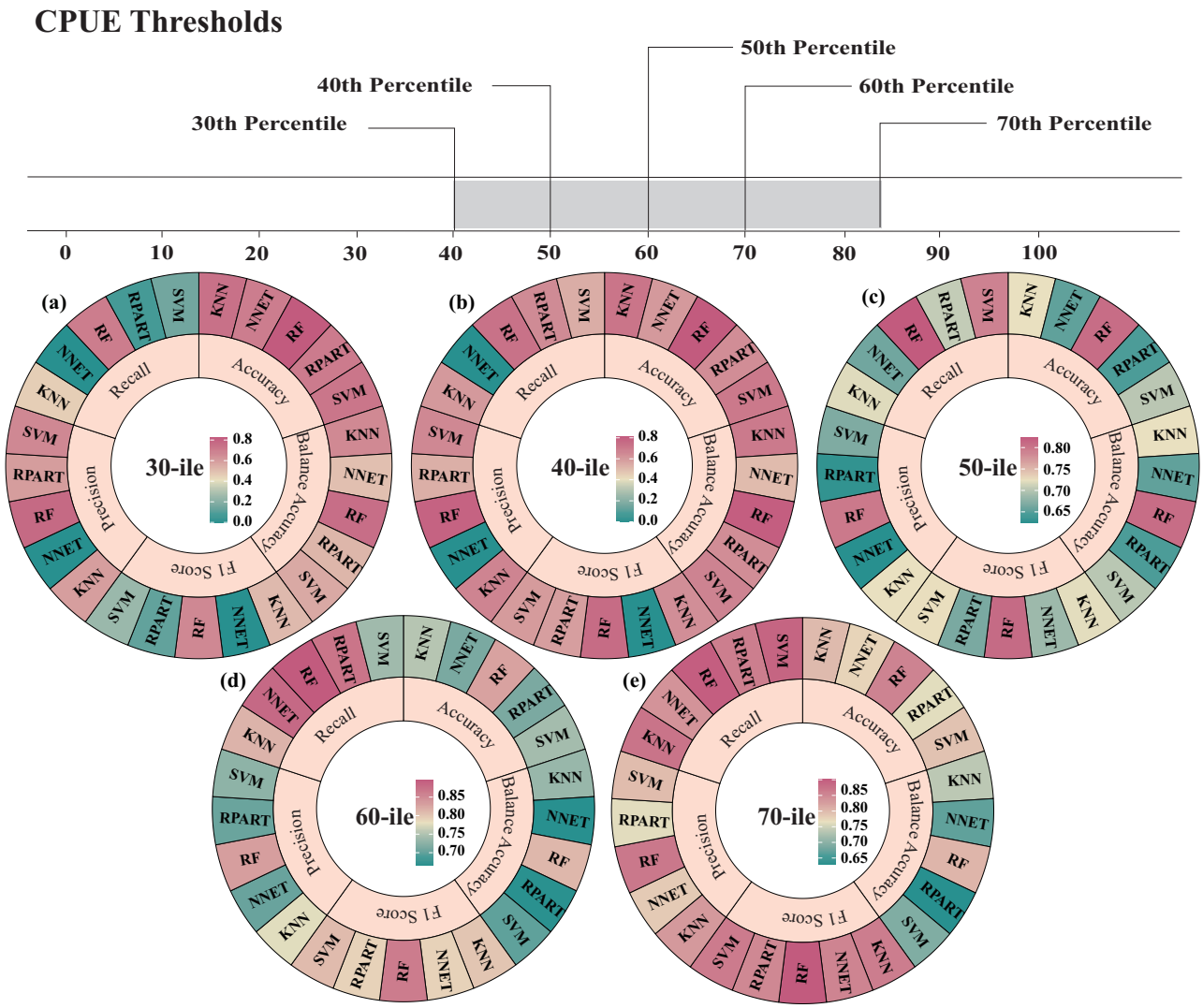

Figure 5. Comprehensive performance comparison of machine learning models across varying CPUE thresholds. The gray bar at the top represents the range of CPUE values corresponding to the number of individuals. The 30-ile represents the 30th percentile, the 40-ile represents the 40th percentile, the 50-ile represents the 50th percentile, the 60-ile represents the 60th percentile, the 70-ile represents the 70th percentile.

ically contributes positively to the predictions. However, the distribution density of points shows that blue points dominate, suggesting that being outside this region has a greater positive impact on the predictions. Similarly, for the remaining regions (SPSG, SSTC, TASM, and WARM), the points transition in color from red to blue, indicating that lower feature values have a modest positive impact on the predictions. Notably, for SSTC and TASM, the distribution density also reveals a dominance of blue points, further suggesting that being outside these regions has a stronger positive influence on the predictions.

ICE and partial dependence plot

The ICE and PDPs, presented in Fig. 9, offer a nuanced view of how individual forecasts vary with specific features. In contrast, PDPs provide an aggregate summary of the average impact of these features across all instances. In the left figure, the ICE plot shows the effect for each decile. Additionally, the original observation values are marked with semi-transparent circles on each ICE line. The gray frequency distribution histogram represents the distribution of the original data. The left-side PDP curve calculates the probability density or a fitted curve, showcasing the fluctuations of the data. The right-side PDP curve, on the other hand, does not calculate the probability density or a fitted curve but directly displays the trend of the data.

For dissolved oxygen (o2_0) values in the range of ~210 mmol/m^3 to 233 mmol/m^3, the curve shows an upward trend, reaches a peak, and then levels off. For temperature (thetao_0), the overall trend of the curve initially rises, then falls, and finally stabilizes. At lower SSTs (15°C–18°C), the curve increases. When the temperature rises to the middle range (18°C to 20°C), it reaches its highest value. Beyond 20°C, the curve starts to decline. Once the temperature exceeds 30°C, the curve drops to its lowest value and gradually stabilizes thereafter. For mesoscale cyclones (NECR) and anticyclones (NEAR), the curves exhibit a similar overall trend: an initial rise, followed by a decline, and then stabilization. Peaks occur at 2.57 times the radius of the mesoscale cyclone and 3.37 times the radius of the mesoscale anticyclone. Beyond these values, the predicted values drop sharply, reach a minimum, and then stabilize.

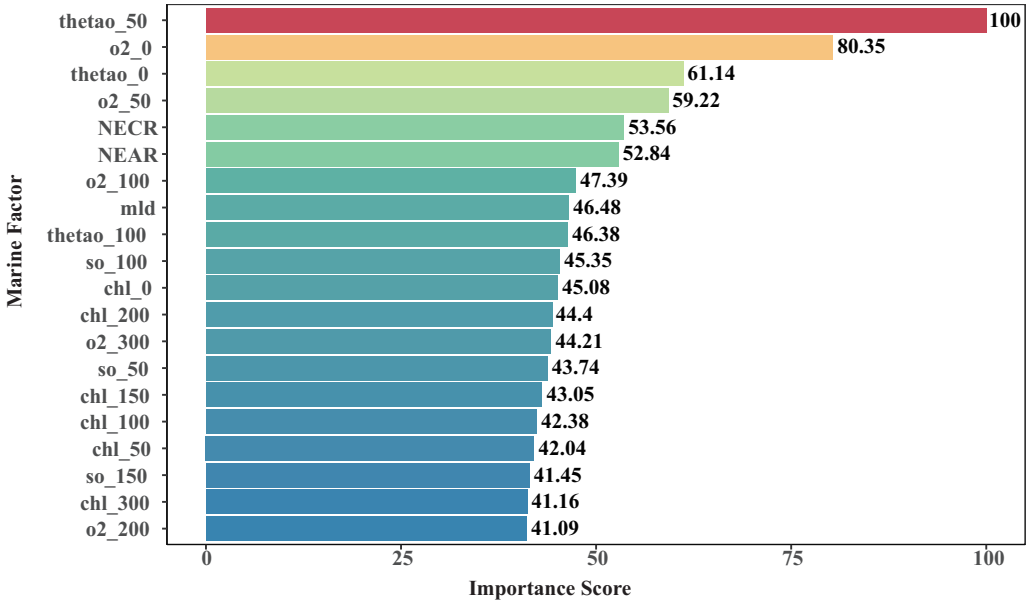

Figure 6. The importance of the top 20 predictors in the Random Forest model. NECR (Normalized distance to the cyclone eddy core radius) is the distance from the cyclone to the operating vessel divided by the cyclone radius. NEAR (Normalized distance to the anticyclone eddy core radius) is the distance from the anticyclone to the operating vessel divided by the anticyclone radius.

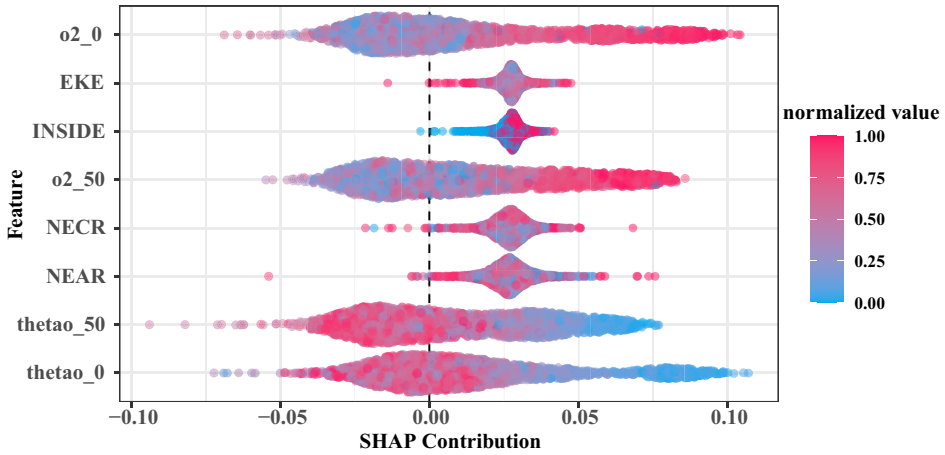

Figure 7. SHAP Summary Plot for o2_0, o2_50, thetao_0, thetao_50, EKE (eddy kinetic energy), INSIDE (whether the latitude and longitude of the operation are within 2.5 times the radius of the cyclone and anticyclone, represented by 0 for "no" and 1 for "yes"), NECR (Normalized distance to the cyclone eddy core radius) is the distance from the cyclone to the operating vessel divided by the cyclone radius. NEAR (Normalized distance to the anticyclone eddy core radius) is the distance from the anticyclone to the operating vessel divided by the anticyclone radius.

Discussion

Comparative analysis on the five models

In this study, the RF model outperformed all other models across various metrics, particularly excelling in accuracy, balanced accuracy, and recall. While SVM and KNN performed well in precision and specificity, they lagged behind RF in recall and F1 score. NNET and RPART showed weaker performance, especially at higher threshold percentages, with NNET also underperforming in specificity and balanced accuracy. Overall, RF and SVM emerged as the top models, with RF achieving the highest area under the curve (AUC) in ROC analysis, indicating its strong ability to differentiate between positive and negative samples. KNN and SVM followed in AUC ranking, while NNET and RPART were less effective, highlighting their limitations with complex data. Previous studies support these findings. For example, Song et al. (2023) found that the RF model performed similarly to an optimized stacked ensemble learning (STK) model when predicting bigeye tuna fishing grounds in the tropical Atlantic. Mugo et al. (2020) and Zhang et al. (2023) reported comparable results for modeling skipjack and albacore tuna habitats, emphasizing RF's reliability in predicting tuna distribution. RF's noise resistance, which stems from its voting mechanism across multiple decision trees, reducing sensitivity to noise and outliers (Belgiu et al. 2016, Biau et al. 2016). This robustness is crucial for handling the inherent noise and uncertainties in marine environmental data (Rubbens et al. 2023). The extensive dataset used in this study, spanning a large temporal and spatial range, included diverse environmental factors from the ocean surface to depths of 300 m, allowing RF to effectively capture the complexity of these variables for more accurate predictions.

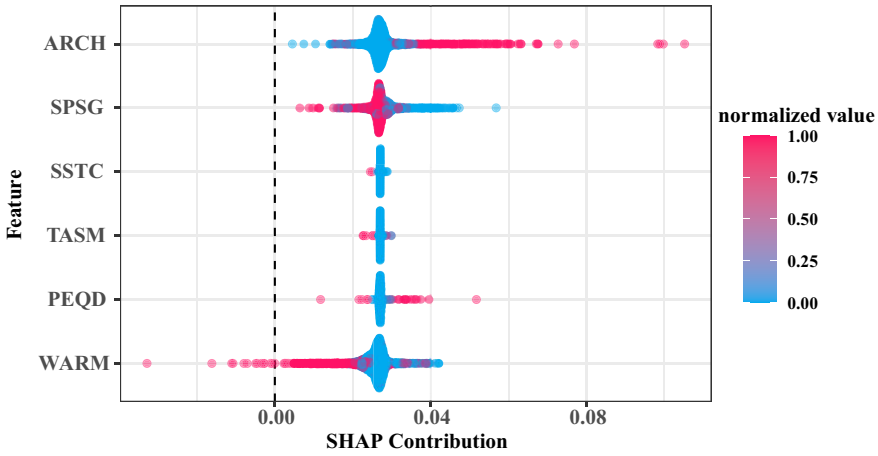

Figure 8. SHAP Summary Plot for ARCH (Trades—Archipelagic Deep Basins Province), PEQD (Trades—Pacific Equatorial Divergence Province), SPSG (Westerlies—S. Pacific Subtropical Gyre Province), SSTC (Westerlies—S. Subtropical Convergence Province), TASM (Westerlies—Tasman Sea Province), WARM (Trades—W. Pacific Warm Pool Province).

Analysis of the importance environmental factors

In the evolving landscape of machine learning, the complexity of models often results in "black box" predictions. Interpretable Machine Learning methods, such as SHAP, PDP, and ICE, are essential for making these models more understandable and transparent. The RF model assesses feature importance, identifying key environmental variables influencing albacore tuna habitat distribution. SHAP, ICE, and PDP analyses further enhance this understanding. The results indicate that dissolved oxygen and temperature at both the surface (o2_0, thetao_0) and 50-m depth (o2_50, thetao_50) are the most influential factors (Fig. 6), followed by the NECR and NEAR, as well as the MLD. These oceanographic variables play a significant role in shaping albacore tuna behavior and physiology.

Adequate oxygen supply is crucial for the physiological functions of organisms. As shown in Fig. 7, the color of the dissolved oxygen (o2_0) points gradually shifts from blue to red, further indicating that higher concentrations of dissolved oxygen (o2_0) provide favorable conditions for the survival of albacore tuna. In oxygen-based (Oxy) models, optimal habitat predictions are observed at a 0-meter depth (Su et al. 2024), which aligns with the results from our RF model. Dissolved oxygen also influences tuna's vertical distribution and spawning areas (Block et al. 1997). Our PDP and ICE analyses reveal a peak surface dissolved oxygen concentration of 233.3 mmol/m^3(∼5.22 ml/l), as shown in Fig. 9a. This finding is largely consistent with Lehodey et al. (2015), who identified oxygen levels above 4.5 ml/l as optimal for albacore tuna in the western Pacific. Additionally, our results align with those of Chang et al. (2021), who observed that albacore tuna in the South Pacific preferentially inhabit areas with dissolved oxygen concentrations ranging from 0.2 to 0.25 mmol/l. The difference of 0.72 ml/l between our results and those of Lehodey et al. (2015) may be due to differences in the temporal and spatial scales of the data. Suitable temperature is equally crucial for tuna survival, as it affects their metabolic functions and distribution. As seen in Fig. 7, the temperature (thetao_0 and thetao_50) points gradually shifts from red to blue, indicating that albacore tuna show a preference for cooler temperatures. Nugroho et al. (2022) found that SSTs between 15°C and 26°C significantly enhance catch rates (CPUE). Additionally, Collette, B. B. (1983) summarized that albacore tuna are most often found in surface waters with temperatures ranging from 15.6°C to 19.4°C, although larger individuals can be found in deeper waters with temperatures ranging from 13.5°C to 25.2°C. These findings are consistent with the results presented in Fig. 8B, which indicate that the optimal temperature is approximately between 18°C and 20°C.

Eddies play a crucial role in mixing the water column, influencing water stratification, and affecting nutrient distribution and primary productivity (McGillicuddy et al. 1998). Cyclonic eddies bring nutrient-rich deep water to the surface, enhancing phytoplankton growth and benefiting the entire food chain. By altering water stratification, these eddies impact the depth of the MLD, which, in turn, influences food availability and the vertical movement of albacore tuna (Lee et al. 2020). In our study, as shown in Fig. 7, the analysis of features such as EKE, INSIDE, NECR, and NEAR revealed that smaller values of EKE, NECR, and NEAR have a stronger positive influence on predictions. Additionally, being located within mesoscale eddies significantly enhances the predictions. Further SHAP analysis of NECR and NEAR (Figs. 9c and 9d) indicates that the influence of cyclonic and anticyclonic eddies peaks at 2.57 and 3.36 times the radius of the eddy center, respectively. These findings emphasize the importance of the eddy periphery regions as favorable habitat conditions for albacore tuna. This aligns with Brandt, S. B. (1981), who observed that eddy edges often form temperature gradients, attracting both cold-water and warm-water fish species. In areas with cyclonic upwelling, the increased surface nutrient content attracts marine life, including albacore tuna prey. Studies have demonstrated that eddies affect organisms across various trophic levels, from plankton to predators (Hsu et al. 2015, Setiawati et al. 2017). Additionally, anticyclonic eddies enhance primary production and biomass at their edges (Lima et al. 2002). Higher CPUE has been associated with SSH isopleths around 0.05 m at the periphery of anticyclonic eddies for albacore tuna in the South Pacific (Zhou et al. 2020). In the Mozambique Channel, tunas are frequently found near cyclonic eddies and their edges, where foraging conditions are more favorable (Tserpes et al. 2008). Moreover, mesoscale eddies, as significant physical phenomena in the ocean, are often associated with cold (cyclonic) or warm (anticyclonic) anoma-

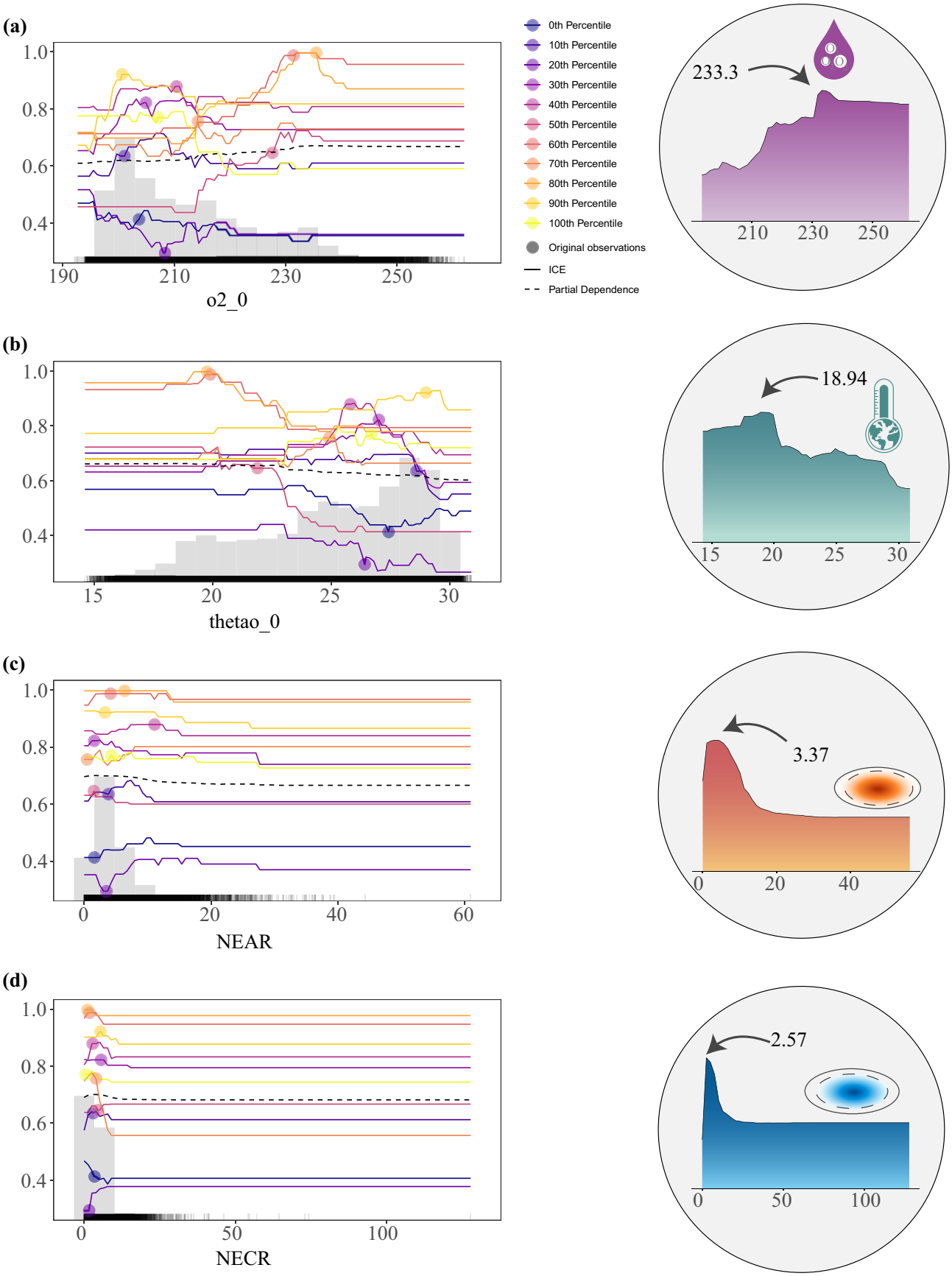

Figure 9. The ICE and PDP curves for dissolved oxygen (o2_0), temperature (thetao_0), NECR (normalized distance to the cyclone eddy core radius, which represents the distance from the cyclone center to the operating vessel divided by the cyclone radius), and NEAR (normalized distance to the anticyclone eddy core radius, which represents the distance from the anticyclone center to the operating vessel divided by the anticyclone radius).

lies at the sea surface. However, a substantial number of eddies exhibit anomalous SSTs, such as warm cyclonic eddies and cold anticyclonic eddies (Sun et al. 2019, Liu et al. 2021, Ni et al. 2021). Early studies were limited by low temporal and spatial resolution, focusing on the upwelling effect in cold eddies and the associated high productivity or highlighting the more suitable temperature in warm eddies as better habitats. As research progressed, this dichotomy has been increasingly challenged. Chelton et al. (2011) showed that both cold and warm eddies can significantly influence primary productivity by modulating chlorophyll distribution along their edges during horizontal movement.

Ecological zone analysis

The spatial variation of ecological regions significantly influences the predictive capacity of the model, as reflected in Fig. 8. Among the six regions analyzed, only the WARM region demonstrates red points distributed on the left side of the x-axis, suggesting that lower feature values within this region positively contribute to predictions. Conversely, points from other regions are concentrated on the right side, indicating a generally positive influence of being either within or outside an ecological region. The ARCH region exhibits a prominent color gradient from blue to red, with red points positioned farther from the x-axis, underscoring its strong positive impact on predictions. For the PEQD region, the red-to-blue-to-red transition highlights the dual influence of this region, where being inside or outside typically enhances predictions. However, the dominance of blue points in the distribution suggests that being outside PEQD contributes more positively. Similarly, for SPSG, SSTC, TASM, and WARM regions, lower feature values show a modest positive impact, with SSTC and TASM regions particularly revealing a higher density of blue points, indicating that areas outside these regions exert a stronger positive influence on predictions. These results suggest that while temporal dynamics may be implicitly captured by the oceanographic data, spatial variation remains a key factor. The contrasting impacts of different regions highlight the complexity of ecological interactions and the need for models to account for spatial heterogeneity when predicting albacore tuna distribution. Future studies should explore more detailed spatial patterns and the interplay between ecological regions and environmental features to refine predictive accuracy.

Prospects and limitations

This study standardized the mesoscale eddy spatial radius data, which to some extent addresses the limitations of previous studies that relied on EKE as a proxy for eddy dynamics. This limitation may prevent the machine learning models from fully capturing key features and variations necessary for comprehensive predictions. Some of the environmental variables included in our analysis may be significantly correlated with each other. This multicollinearity could affect the interpretability of the results, as it may mask the independent effects of individual variables. In future analyses, we will use methods such as variance inflation factors to address multicollinearity and more accurately isolate the effects of each variable. Our study employed interpretable models such as RF, SHAP, ICE, and PDP, which yielded good results. However, the analysis did not fully account for the details of spatial distribution, which could introduce some bias. While temporal dynamics at the seasonal scale may be implicit in the oceanographic data, spatial variation presents a more challenging factor. In machine learning, many models assume feature independence, but in reality, spatial data often exhibit spatial correlation. Future research could consider incorporating spatial statistical methods, such as Spatial Autoregressive Models (SARs), and introducing finer spatial resolutions. These approaches can effectively capture spatial variation and correlation, thereby improving predictive accuracy.

Our model provides actionable insights for fisheries management, particularly in optimizing fishing operations and minimizing ecological impacts. By identifying key areas with favorable environmental conditions for albacore tuna, such as optimal oxygen levels and temperatures, and proximity to mesoscale eddies, our approach allows fisheries to target high-efficiency zones. This can significantly improve catch rates while reducing bycatch, especially in regions with abundant prey near eddy peripheries. The study emphasizes the critical role of mesoscale eddies, not just as feeding grounds but as vital habitats for albacore tuna throughout the year. As such, we recommend incorporating mesoscale eddy data into real-time fisheries monitoring systems to better predict albacore tuna distribution and adjust fishing strategies accordingly. To further improve management, fisheries could integrate predictive models into daily operations, allowing for real-time adjustments to fishing locations, quotas, and regulations based on up-to-date oceanographic conditions. This dynamic, data-driven approach will not only optimize albacore tuna catch efficiency but also the protection of albacore tuna habitats, contributing to the overall health of marine ecosystems.

Acknowledgements

Additionally, we appreciate the assistance provided by our colleagues and the support of Shanghai Ocean University throughout the course of this study. We would like to express our gratitude to Dr David M. Kaplan and another anonymous reviewer for their valuable feedback and constructive suggestions, which have greatly improved the quality of this manuscript.

Author contributions

Liwen Liu: Data curation, data analysis, drafting the original manuscript. Rong Wan: Supervision, project administration, funding acquisition, manuscript review and editing. Feng Wu: Visualization, software development, data analysis. Yucheng Wang: Data curation, resources, methodology. Yonghan Zhu: Investigation, validation, resources. Cheng Zhou: Conceptualization, methodology development, project administration, manuscript review and editing.

Conflict of interest: The authors declare no conflicts of interest. The illustrations in this manuscript are partly hand-painted and partly royalty-free vector graphics.

Funding

This study was funded by the National Key R&D Program of China (Project No. 2023YFD2401301 and 2023YFD2401305) and the Program on the Survey, Monitoring and Assessment of Global Fishery Resources sponsored by the Ministry of Agriculture and Rural Affairs (Project No. D-8025-24-5001). We gratefully acknowledge the financial support that enabled us to conduct this study.

Data availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request. These datasets are subject to restrictions due to data access agreements.

References

Arostegui MC, Gaube P, Woodworth-Jefcoats PA *et al.* Anticyclonic eddies aggregate pelagic predators in a subtropical gyre. *Nature* 2022;**609**:535–40. https://doi.org/10.1038/s41586-022-05162-6

Arrizabalaga H, Dufour F, Kell L *et al.* Global habitat preferences of commercially valuable tuna. *Deep Sea Res Part II* 2015;**113**:102–12. https://doi.org/10.1016/j.dsr2.2014.07.001

Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J Photogramm Remote Sens* 2016;**114**:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bertrand A, Josse E, Bach P *et al.* Hydrological and trophic characteristics of tuna habitat: consequences on tuna distribution and longline catchability. *Can J Fish AquatSci* 2002;**59**:1002–13. https://doi.org/10.1139/f02-073

Biau G, Scornet E. A random forest guided tour. *Test* 2016;**25**:197–227. https://doi.org/10.1007/s11749-016-0481-7

Block BA, Keen JE, Castillo B *et al.* Environmental preferences of yellowfin tuna (*Thunnus albacares*) at the northern extent of its range. *Mar Biol* 1997;**130**:119–32. https://doi.org/10.1007/s002270050231

Brandt SB. Effects of a Warm-Core Eddy on Fish Distributions. *Mar Ecol Progr Ser* 1981;**6**:19–33.

Chang YJ, Hsu J, Lai PK *et al.* Evaluation of the impacts of climate change on albacore distribution in the South Pacific Ocean by using ensemble forecast. *Front Mar Sci* 2021;**8**:731950. https://doi.org/10.3389/fmars.2021.731950

Chelton DB, Gaube P, Schlax MG *et al.* The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science* 2011;**334**:328–32. https://doi.org/10.1126/science.1208897

CHEN IC, LEE PF, TZENG WN. Distribution of albacore (*Thunnus alalunga*) in the Indian Ocean and its relation to environmental factors. *Fisher Oceanogr* 2005;**14**:71–80. https://doi.org/10.1111/j.1365-2419.2004.00322.x

Collette BB, Nauen CE. Scombrids of the world: an annotated and illustrated catalogue of tunas, mackerels, bonitos, and related species known to date. v. 2. 1983. https://www.fao.org/4/ac478e/ac478e00.htm (2 October 2023, date last accessed).

Dong C, Liu L, Nencioli F *et al.* The near-global ocean mesoscale eddy atmospheric-oceanic-biological interaction observational dataset. *Sci Data* 2022;**9**:436. https://doi.org/10.1038/s41597-022-01550-9

Elith J, Graham H, P. Anderson C *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 2006;**29**:129–51. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Fitrianah D, Hidayanto AN, Gaol JL *et al.* A spatio-temporal data-mining approach for identification of potential fishing zones based on oceanographic characteristics in the eastern Indian Ocean. *IEEE J Select Topics Appl Earth Observ Remote Sens* 2015;**9**:3720–8. https://doi.org/10.1109/JSTARS.2015.2492982

Gavaris S. Use of a multiplicative model to estimate catch rate and effort from commercial data. *Can J Fish AquatSci* 1980;**37**:2272–5. https://doi.org/10.1139/f80-273

Godø OR, Samuelsen A, Macaulay GJ *et al.* Mesoscale eddies are oases for higher trophic marine life. *PLoS One* 2012;**7**:e30161. https://doi.org/10.1371/journal.pone.0030161

Goldstein A, Kapelner A, Bleich J *et al.* 2017. Package 'ICEbox'. https://cran.ma.imperial.ac.uk/web/packages/ICEbox/ICEbox.pdf (2 September 2024, date last accessed).

Handegard NO, Buisson LD, Brehmer P *et al.* Towards an acoustic-based coupled observation and modelling system for monitoring and predicting ecosystem dynamics of the open ocean. *Fish Fisher* 2013;**14**:605–15. https://doi.org/10.1111/j.1467-2979.2012.00480.x

Hsu AC, Boustany AM, Roberts JJ *et al.* Tuna and swordfish catch in the US northwest Atlantic longline fishery in relation to mesoscale eddies. *Fisheries Oceanography* 2015;**24**:508–20. https://doi.org/10.1111/fog.12125

Hu Z, Tan Y, Song X *et al.* Influence of mesoscale eddies on primary production in the South China Sea during spring inter-monsoon period. *Acta Oceanolog Sin* 2014;**33**:118–28. https://doi.org/10.1007/s13131-014-0431-8

Lee MA, Weng JS, Lan KW *et al.* Empirical habitat suitability model for immature albacore tuna in the North Pacific Ocean obtained using multisatellite remote sensing data. *Int J Remote Sens* 2020;**41**:5819–37. https://doi.org/10.1080/01431161.2019.1666317

Lehodey P, Senina I, Nicol S *et al.* Modelling the impact of climate change on South Pacific albacore tuna. *Deep Sea Res Part II* 2015;**113**:246–59. https://doi.org/10.1016/j.dsr2.2014.10.028

Lima ID, Olson DB, Doney SC. Biological response to frontal dynamics and mesoscale variability in oligotrophic environments: Biological production and community structure. *J Geophys Res Oceans* 2002;**107**:25–1. https://doi.org/10.1029/2000JC000393

Liu Y, Zheng Q, Li X. Characteristics of global ocean abnormal mesoscale eddies derived from the fusion of sea surface height and temperature data by deep learning. *Geophys Res Lett* 2021;**48**:e2021GL094772. https://doi.org/10.1029/2021GL094772

Longhurst A, Sathyendranath S, Platt T *et al.* An estimate of global primary production in the ocean from satellite radiometer data. *J Plankton Res* 1995;**17**:1245–71. https://doi.org/10.1093/plankt/17.6.1245

Lopez J, Moreno G, Lennert-Cody C *et al.* Environmental preferences of tuna and non-tuna species associated with drifting fish aggregating devices (DFADs) in the Atlantic Ocean, ascertained through fishers' echo-sounder buoys. *Deep Sea Res Part II* 2017;**140**:127–38. https://doi.org/10.1016/j.dsr2.2017.02.007

Maury O. An overview of APECOSM, a spatialized mass balanced "Apex Predators ECOSystem Model" to study physiologically structured tuna population dynamics in their ecosystem. *Prog Oceanogr* 2010;**84**:113–7. https://doi.org/10.1016/j.pocean.2009.09.013

McGillicuddy Jr DJ, Robinson AR, Siegel DA *et al.* Influence of mesoscale eddies on new production in the Sargasso Sea. *Nature* 1998;**394**:263–6. https://doi.org/10.1038/28367

Mohri M, Nishida T. Distribution of bigeye tuna and its relationship to the environmental conditions in the Indian Ocean based on the Japanese longline fisheries information. *Res Dev* 1999;**299**:319pp. https://www.fao.org/fishery/docs/CDrom/IOTC_Proceedings(1999-2002)/files/proceedings/proceedings2/wptt/TT99-11.pdf (17 June 2024, date last accessed).

Mondal S, Wang YC, Lee MA *et al.* Ensemble three-dimensional habitat modeling of Indian Ocean immature albacore tuna (*Thunnus alalunga*) using remote sensing data. *Remote Sens* 2022;**14**:5278. https://doi.org/10.3390/rs14205278

Mugo R, Saitoh SI. Ensemble modelling of skipjack tuna (Katsuwonus pelamis) habitats in the western north pacific using satellite remotely sensed data; a comparative analysis using machine-learning models. *Remote Sens* 2020;**12**:2591. https://doi.org/10.3390/rs12162591

Ni Q, Zhai X, Jiang X *et al.* Abundant cold anticyclonic eddies and warm cyclonic eddies in the global ocean. *J Phys Oceanogr* 2021;**51**:2793–806. https://doi.org/10.1175/JPO-D-21-0010.1

Nikolic N, Morandeau G, Hoarau L *et al.* Review of albacore tuna, *Thunnus alalunga*, biology, fisheries and management. *Rev Fish Biol Fisher* 2017;**27**:775–810. https://doi.org/10.1007/s11160-016-9453-y

Nugroho SC, Setiawan RY, Setiawati MD *et al.* Estimation of Albacore Tuna Potential Fishing Grounds in the Southeastern Indian Ocean. *IEEE Access* 2022;**11**:1141–7. https://doi.org/10.1109/ACCESS.2022.3233353

Okamoto H, Miyabe N, Matsumoto T. GLM analyses for standardization of Japanese longline CPUE for bigeye tuna in the Indian Ocean applying environmental factors. In *IOTC Proceedings*. 2001;

(Vol. **4**, pp.491–522). ⟨?PMU?⟩https://www.fao.org/fishery/docs/CDrom/IOTC_Proceedings(1999-2002)/files/proceedings/proceedings4/wptt/TT01-21.pdf (11 July 2024, date last accessed).

Pauly D, Christensen V, Guénette S *et al.* Towards sustainability in world fisheries. *Nature* 2002;**418**:689–95. https://doi.org/10.1038/nature01017

Pitcher TJ, Preikshot D. RAPFISH: A rapid appraisal technique to evaluate the sustainability status of fisheries. *Fish Res* 2001;**49**:255–70. https://doi.org/10.1016/S0165-7836(00)00205-8

Rubbens P, Brodie S, Cordier T *et al.* Machine learning in marine ecology: an overview of techniques and applications. *ICES J Mar Sci* 2023;**80**:1829–53. https://doi.org/10.1093/icesjms/fsad100

Schaefer KM, Fuller DW. Movements, behavior, and habitat selection of bigeye tuna (*Thunnus obesus*) in the eastern equatorial Pacific, ascertained through archival tags. 2002. https://aquadocs.org/handle/1834/31101 (11 September 2024, date last accessed).

Setiawati MD, Sambah AB, Miura F *et al.* Characterization of bigeye tuna habitat in the Southern Waters off Java–Bali using remote sensing data. *Adv Space Res* 2015;**55**:732–46. https://doi.org/10.1016/j.asr.2014.10.007

Setiawati MD, Tanaka T. Utilization of scatterplot smoothers to understand the environmental preference of Bigeye Tuna in the southern waters off Java-Bali: Satellite remote sensing approach. *Fishes* 2017;**2**:2. https://doi.org/10.3390/fishes2010002

Song L, Li T, Zhang T *et al.* Comparison of machine learning models within different spatial resolutions for predicting the bigeye tuna fishing grounds in tropical waters of the Atlantic Ocean. *Fisher Oceanogr* 2023;**32**:509–26. https://doi.org/10.1111/fog.12643

Su S, Mao Q, Li Y *et al.* Deep learning-based fishing ground prediction for albacore and yellowfin tuna in the Western and Central Pacific Ocean. *Fish Res* 2024;**278**:107103. https://doi.org/10.1016/j.fishres.2024.107103

Sun W, Dong C, Tan W *et al.* Statistical characteristics of cyclonic warm-core eddies and anticyclonic cold-core eddies in the North Pacific based on remote sensing data. *Remote Sens* 2019;**11**:208. https://doi.org/10.3390/rs11020208

Sund PN, Blackburn M, Williams F. Tunas and their environment in the Pacific Ocean: a review. *Oceanogr Mar Biol Ann Rev* 1981;**19**:443–512. https://swfsc-publications.fisheries.noaa.gov/publications/CR/1981/8161.PDF (11 March 2024, date last accessed).

Teears T, Castillo-Jordán C, Davies N *et al.* Western and Central Pacific Fisheries Commission (WCPFC). *2024 South Pacific Albacore Tuna Stock Assessment (C20-SA-WP-02, Rev. 3)*. 2024. https://meetings.wcpfc.int/file/15784/download (20 May 2024, date last accessed).

Tserpes G, Peristeraki P, Valavanis VD. Distribution of swordfish in the eastern Mediterranean, in relation to environmental factors and the species biology. *Hydrobiologia* 2008;**612**:241–50. https://doi.org/10.1007/s10750-008-9499-5

Wang G, Su J, Chu PC. Mesoscale eddies in the South China Sea observed with altimeter data. *Geophys Res Lett* 2003;**30**. https://doi.org/10.1029/2003GL018532

Williams AJ, Allain V, Nicol SJ *et al.* Vertical behavior and diet of albacore tuna (*Thunnus alalunga*) vary with latitude in the South Pacific Ocean. *Deep Sea Res Part II* 2015;**113**:154–69. https://doi.org/10.1016/j.dsr2.2014.03.010

Xu Y, Nieto K, Teo SL *et al.* Influence of fronts on the spatial distribution of albacore tuna (*Thunnus alalunga*) in the Northeast Pacific over the past 30 years (1982–2011). *Prog Oceanogr* 2017;**150**:72–8. https://doi.org/10.1016/j.pocean.2015.04.013

Yang S, Song L, Zhang Y *et al.* The potential vertical distribution of bigeye tuna (*Thunnus obesus*) and its influence on the spatial distribution of CPUEs in the Tropical Atlantic Ocean. *J Ocean Univ China* 2020;**19**:669–80. https://doi.org/10.1007/s11802-020-4264-0

Zagaglia CR, Lorenzzetti JA, Stech JL. Remote sensing data and longline catches of yellowfin tuna (*Thunnus albacares*) in the equatorial Atlantic. *Remote Sens Environ,* 2004;**93**:267–81. https://doi.org/10.1016/j.rse.2004.07.015

Zainuddin M, Kiyofuji H, Saitoh K *et al.* Using multi-sensor satellite remote sensing and catch data to detect ocean hot spots for albacore (*Thunnus alalunga*) in the northwestern North Pacific. *Deep Sea Res Part II* 2006;**53**:419–31. https://doi.org/10.1016/j.dsr2.2006.01.007

Zainuddin M, Safruddin S, Selamat MB *et al.* Prediction of potential fishing zones for skipjack tuna during the northwest monsoon using remotely sensed satellite data. *Mar Sci* 2017;**22**:59–66. https://doi.org/10.14710/ik.ijms.22.2.59-66

Zainuddin M, Saitoh K, SAITOH SI. Albacore (*Thunnus alalunga*) fishing ground in relation to oceanographic conditions in the western North Pacific Ocean using remotely sensed satellite data. *Fisher Oceanogr* 2008;**17**:61–73. https://doi.org/10.1111/j.1365-2419.2008.00461.x

Zhang J, Fan D, He H *et al.* Forecasting Albacore (*Thunnus alalunga*) fishing grounds in the South Pacific based on machine learning algorithms and ensemble learning model. *Appl Sci* 2023;**13**:5485. https://doi.org/10.3390/app13095485

Zhou C, He P, Xu L *et al.* The effects of mesoscale oceanographic structures and ambient conditions on the catch of albacore tuna in the South Pacific longline fishery. *Fisher Oceanogr* 2020;**29**:238–51. https://doi.org/10.1111/fog.12467

Handling Editor: David M. Kaplan