

RESEARCH ARTICLE OPEN ACCESS

Factors Influencing Species Distribution Model Performance in Tropical Reef Fishes

Ying Liu¹ | Shuaishuai Liu^{2,3,4} | Ákos Bede-Fazekas^{5,6} | Stefano Mammola^{7,8,9}  | Liuyong Ding¹⁰ | Jiqi Gu¹¹ | Gabriel Nakamura¹² | Qiang Lin^{2,3,4} | Dazhi Wang¹ | Zhixin Zhang^{2,3,4} 

¹State Key Laboratory of Marine Environmental Science, College of the Environment and Ecology, Xiamen University, Xiamen, China | ²State Key Laboratory of Tropical Oceanography, Guangdong Provincial Key Laboratory of Applied Marine Biology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China | ³University of Chinese Academy of Sciences, Beijing, China | ⁴Marine Biodiversity and Ecological Evolution Research Center, South China Sea Institute of Oceanology, Guangzhou, People's Republic of China | ⁵HUN-REN Centre for Ecological Research, Institute of Ecology and Botany, Vácrátót, Hungary | ⁶Department of Environmental and Landscape Geography, ELTE Eötvös Loránd University, Institute of Geography and Earth Sciences, Budapest, Hungary | ⁷Molecular Ecology Group (MEG), Water Research Institute (IRSA), National Research Council (CNR), Verbania, Pallanza, Italy | ⁸Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland | ⁹NBFC, National Biodiversity Future Center, Palermo, Italy | ¹⁰Museum of Hydrobiological Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China | ¹¹State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Life Sciences of Beijing Normal University, Beijing, China | ¹²Programa de Pós-Graduação Em Ecologia e Evolução da Universidade Federal de Goiás, Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás, Brazil

Correspondence: Dazhi Wang (dzwang@xmu.edu.cn) | Zhixin Zhang (zxzhang@scsio.ac.cn)

Received: 20 May 2025 | **Revised:** 29 January 2026 | **Accepted:** 8 February 2026

Editor: Yoan Fourcade

Keywords: biodiversity conservation | marine biodiversity | predictive performance | range estimation | reef fish | species distribution model

ABSTRACT

Aim: Reliable biodiversity assessments using species distribution models (SDMs) are essential for effective conservation and management. Understanding factors influencing SDM performance is crucial for improving model reliability, yet such links remain underexplored in marine systems. To address this gap, we quantified the effects of geographical and species-level factors on SDM performance in tropical reef fishes, aiming to provide practical guidelines on which species are more likely to yield reliable predictions.

Location: Global tropical reef ecosystems.

Methods: We built ensemble SDMs for 1941 tropical reef fish species using occurrence records. We evaluated model predictive performance using the continuous Boyce index, the most suitable performance metric given that we lacked quality absence data, and two other commonly applied metrics (the area under the receiving operating characteristic curve and the true skill statistic). We compiled 10 factors related to species' geographical and ecological characteristics and assessed their influence on model performance using phylogenetic generalised linear models.

Results: Ensemble SDMs for tropical reef fishes exhibited high predictive performance based on the three evaluation metrics. Phylogenetic generalised linear models relating evaluation metrics to species geographical and ecological characteristics showed modest explanatory power, with R^2 varying from 0.253 to 0.341. Across evaluation metrics, SDM performance was strongly associated with species' latitude, proximity to shore, and environmental similarity between training and evaluation datasets. For continuous Boyce index, there were additional significant effects for range size, parental care, range coverage, and species description year.

Ying Liu and Shuaishuai Liu contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Diversity and Distributions* published by John Wiley & Sons Ltd.

Main Conclusions: Our study provides a practical framework for identifying tropical reef fish species more likely to yield reliable SDM predictions. The identified factors offer guidance for researchers to anticipate model reliability before undertaking extensive modelling efforts. This large-scale, multi-species comparative approach is broadly applicable to other marine taxa and regions, advancing our ability to model and conserve marine biodiversity.

1 | Introduction

Driven by multiple threats, global biodiversity is rapidly declining (Cowie et al. 2022; Jaureguiberry et al. 2022). Therefore, safeguarding biodiversity has become an urgent global priority. Central to this effort is a clear understanding of species' geographical distributions. Species distribution models (SDMs) have been widely used to estimate species' ranges by correlating occurrence records with environmental predictors (Guisan and Thuiller 2005; Elith et al. 2010). Thus far, SDMs have been applied in multiple disciplines, including conservation biology, biogeography, biological invasions, and global change biology (Araújo et al. 2019).

Before applying SDMs in biodiversity assessments, researchers should properly assess the reliability of SDM outputs, since range estimates by models with strong predictive performance are more likely to be informative and accurate. The predictive performance of SDMs can be influenced by various factors, including modelling algorithms (Valavi et al. 2022), the number of occurrence records (Wisiz et al. 2008), background data sampling method (Steen et al. 2024), and the number and type of environmental predictors used (e.g., Brun et al. 2020; Santini et al. 2021). Additionally, species-level intrinsic and extrinsic characteristics may also affect model performance. For example, range size is closely related to the quantity and quality of species distribution data (Meyer et al. 2016), which may subsequently influence the accuracy of SDM predictions (McPherson and Jetz 2007). To date, researchers have explored factors influencing model performance in several taxonomic groups in the terrestrial realm, including birds (e.g., Wiethase et al. 2024), mammals (e.g., Morán-Ordóñez et al. 2017), insects (e.g., Eskildsen et al. 2013), and plants (e.g., Guisan et al. 2007; Collart et al. 2023). However, just a handful of studies have addressed this question in marine systems (Soininen and Luoto 2014; Luan et al. 2020).

Marine and terrestrial ecosystems differ fundamentally in many key aspects, such as biota, environments, and availability of occurrence records (Webb 2012; Hughes et al. 2021; Zhang, Kass, et al. 2025), making it challenging to directly transfer findings from land to sea. Moreover, the few available studies on marine species focused on a limited number of species (e.g., 21 species in Luan et al. 2020), limiting the generalizability of findings. In addition, the effects of candidate factors on SDM performance have been inconsistent across studies (e.g., Stockwell and Peterson 2002; Soininen and Luoto 2014). The limited empirical evidence in marine systems, combined with the contradictory findings across studies, highlights the need for further investigation into how different factors influence SDM performance in marine species.

Tropical reef fishes represent an ideal taxon to fill this knowledge gap. Tropical reef fishes are globally distributed and exceptionally species-rich, comprising over 6000 species of considerable ecological, economic, cultural, and aesthetic importance to human

societies (Parravicini et al. 2014, 2021; Langlois et al. 2022; Mouquet et al. 2024). Moreover, tropical reef fishes exhibit remarkable diversity in ecological and life-history traits (Parravicini et al. 2014, 2021), which provides an excellent opportunity to explore factors influencing SDM performance in the marine realm.

We developed ensemble SDMs for nearly 2000 tropical reef fish species, evaluated their predictive performance using three widely used metrics, and identified factors associated with model performance. Candidate factors were selected based on the theoretical foundation of SDMs and scientific knowledge on the ecology of tropical reef fishes, and hypotheses on the impact of the candidate factors were postulated. By examining the factors influencing predictive performance of SDMs across a broad set of tropical reef fish species, we aim to provide a practical guideline for future model development. Such guidance is essential for improving the predictive utility of SDMs in macroecological analyses and for supporting more targeted conservation efforts in marine ecosystems.

2 | Materials and Methods

2.1 | Species List and Distribution Data

We obtained a comprehensive list of tropical reef fishes from Parravicini et al. (2013, 2014). We initially considered 6316 tropical reef fish species distributed across the Atlantic Ocean, the Indian Ocean, the Pacific Ocean, and the Tropical Eastern Pacific (Tables S1–S3). We verified scientific names by cross-referencing the World Register of Marine Species (<https://www.marinespecies.org>), which provides authoritative taxonomic information for marine organisms.

We developed SDMs for tropical reef fish species using opportunistic occurrence records from the Global Biodiversity Information Facility (GBIF.org 2025, retrieved 12 October 2025, available at <https://doi.org/10.15468/dl.w7zpa6>). Collection years of opportunistic occurrence records spanned from 1639 to 2025 (Figures S1–S2). To ensure a temporal match between species occurrences and marine data layers (between 2000 and 2014, see Marine Predictors for details) while retaining sufficient data for robust model development, we decided to consider records collected since 1990, deeming a 10-year temporal window acceptable for matching. This subset comprised 93% of available occurrences (Figure S1).

Opportunistic occurrence records represent species' fine-scale distribution but are susceptible to biases and errors (Zizka et al. 2019; Hughes et al. 2021; Zhang, Ma, et al. 2025; Zhang, Kass, et al. 2025). Therefore, we filtered occurrence records before developing SDMs. We first removed erroneous records, including records on land and invalid records (e.g., zero coordinates). Next, we dealt with spatially clustered records, as they can adversely

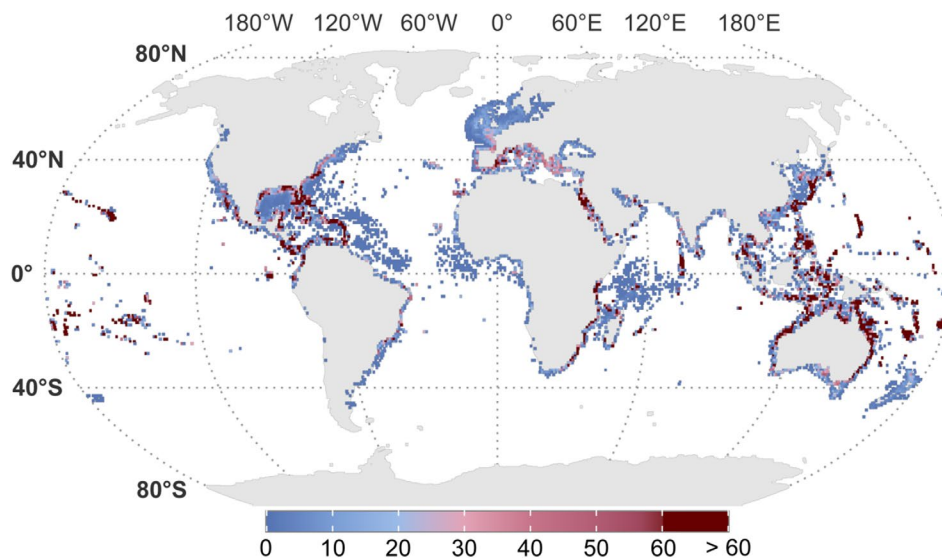


FIGURE 1 | Spatial distribution of occurrence records for 1941 tropical reef fish species used to develop species distribution models. Maps are displayed at a spatial resolution of 10 arcmin to enhance readability.

affect model performance by artificially inflating the importance of certain environmental conditions (Kramer-Schadt et al. 2013). We avoided clustered occurrence records by keeping one record per 5 arcmin grid cell, corresponding to the spatial resolution of our predictor layers (see Marine Predictors below), followed by a spatial thinning using a 20-km thinning distance (Hu et al. 2021).

After the above data filtering procedures, we were still at risk of including records beyond species' known geographical ranges. We avoided this issue with the help of expert range maps from the International Union for Conservation of Nature Red List of Threatened Species (hereafter IUCN Red List; <https://www.iucnredlist.org>). Expert range maps are constructed using taxonomic and ecological knowledge, describe species' geographic range boundaries, and have been widely applied in macroecological studies (e.g., Thuiller et al. 2019; Chevalier et al. 2024; Zhang, Ma, et al. 2025; Zhang, Kass, et al. 2025). For each species, we downloaded its expert range map from the IUCN Red List, created a 50-km buffer around the expert range polygon following Meyer et al. (2016), and retained only occurrence records located within this buffer area.

The number of occurrence records strongly influences the predictive performance of SDMs, with species having fewer records generally producing less reliable predictions (Wisz et al. 2008). Therefore, we filtered the species list to retain only those with at least 30 occurrences (Wisz et al. 2008). After filtering, we kept 1941 species for subsequent analyses, with occurrences ranging from 30 to 2353 records per species (Figure 1 and Figure S1; Table S1).

2.2 | Marine Predictors

As with macroecological studies involving multiple species (e.g., Thuiller et al. 2019; Chevalier et al. 2024), we used a uniform set of predictors to estimate species-environment relationships for nearly 2000 tropical reef fish species. While different species might have different environmental requirements and

species-specific predictor selection would be ideal, using a uniform predictor set ensures methodological consistency and comparability across species.

Evidence from the literature shows that a variety of environmental factors can influence the geographical distributions of tropical reef fish species, including climatic, physical, chemical, and topographical factors (Bradie and Leung 2017). Among these, marine topographic features, such as ocean depth, are widely recognised as key determinants of species distribution (Bradie and Leung 2017). In addition, physicochemical properties of the ocean, including temperature, salinity, current velocity, pH and dissolved oxygen, have been shown to significantly affect the survival and spatial patterns of marine fish (e.g., Bradie and Leung 2017; Cheung et al. 2021).

Based on this evidence, we initially sourced 14 marine predictors from the Bio-ORACLE version 2.2 database at a spatial resolution of 5 arcmin, which describes marine conditions for the period 2000–2014 (Assis et al. 2018) (Table S2). To reduce multicollinearity among predictors, we computed pairwise Pearson correlation (r) among predictors (Figure S3). Based on evidence from the literature (e.g., Bradie and Leung 2017; Mellin et al. 2010; Cheung et al. 2021) and our expert knowledge on tropical reef fish ecology, we selected the most ecologically meaningful predictor among highly correlated predictors (i.e., $|r| > 0.7$) (Dormann et al. 2013) (Table S2). Following this multicollinearity assessment, we retained five predictors: ocean depth, maximum sea surface temperature, pH, minimum sea surface salinity, and maximum current velocity (Figure S3, Table S2). Note that the maximum and minimum data layers represent the interannual minimum and maximum monthly values observed during 2000–2014 (Assis et al. 2018).

2.3 | Model Development

To ensure reproducibility of our modelling analyses, we reported all steps via a standard ODMAP (Zurell et al. 2020; Table S3).

The choice of modelling algorithms is a major source of uncertainties in biodiversity assessment (e.g., Thuiller et al. 2019; Valavi et al. 2022). To ensure the robustness of our results and reduce uncertainties associated with algorithm choice, we developed ensemble SDMs using three algorithms: generalised additive model (GAM), maximum entropy (MaxEnt) and random forest (RF). These three algorithms include regression and machine learning techniques, and they have been widely applied in marine biodiversity studies (Melo-Merino et al. 2020).

For each fish species, we generated a calibration area by applying a 1000 km buffer around occurrence records (Huang et al. 2024; Zhang, Ma, et al. 2025; Zhang, Kass, et al. 2025) and randomly selected pseudo-absence points (background data for MaxEnt) within this extent. There are several recommendations in the literature about the optimal numbers of pseudo-absences (e.g., Barbet-Massin et al. 2012; Steen et al. 2024; Whitford et al. 2024; Zbinden et al. 2024). We followed more traditional recommendations and generated 10,000 random points as pseudo-absences for all algorithms (Elith et al. 2011; Barbet-Massin et al. 2012; Valavi et al. 2021). Following best practice regarding SDMs (e.g., Araújo et al. 2019; Feng et al. 2019), we optimised model complexity to enhance accuracy and reduce the risk of overfitting. A complete description of the modelling workflow, including detailed information on model complexity optimization, is provided in the ODMAP protocol (Table S3).

We evaluated model performance using a spatially stratified four-fold cross-validation approach. For each species, occurrence and pseudo-absence records were divided into four spatial blocks using the `get.block` function in the ENMeval R package (Kass et al. 2021). Three blocks were randomly selected for model training, and the remaining block was used for model evaluation (Valavi et al. 2019; Huang et al. 2024). This step was repeated four times until all blocks were used for model evaluation. Model performance represents the average across all four cross-validation iterations. While spatial block cross-validation allows the use of more than four folds (e.g., Valavi et al. 2019), the lower the number of folds, the less biased performance measures are expected. Therefore, we selected four folds, which is typical in SDM studies (Kass et al. 2021).

Various metrics have been proposed to assess SDM performance. In this study, we evaluated the predictive performance of each modelling algorithm using three commonly used metrics: (a) the continuous Boyce index (CBI) (Hirzel et al. 2006), which is specifically designed for presence-only models; (b) the widely used AUC (Hanley and McNeil 1982); and (c) the prevalence-independent true skill statistic (TSS) (Hanssen and Kuipers 1965; Allouche et al. 2006). Although researchers have pointed out the limitations of AUC (e.g., Lobo et al. 2008) and TSS (e.g., Somodi et al. 2017; Leroy et al. 2018) as model evaluation metrics, we included them in our analyses for comparative purposes as they are widely used—although often improperly—in the SDM literature.

Finally, we assembled the outputs of the three algorithms to generate consensus predictions about species distributions. We constructed four ensemble models: one average ensemble model and three weighted (using CBI, AUC and TSS as weighting criteria) average ensemble models. We measured the predictive

performance of ensemble models using the abovementioned spatial cross-validation approach.

2.4 | Candidate Factors Influencing Model Performance

Building on existing evidence of factors influencing the predictive performance of SDM, and their potential roles in shaping species geographical distributions, we considered 10 candidate factors in our analyses (Table 1, Table S1), including:

- Range size: Range size represents a fundamental ecological and evolutionary characteristic (Brown et al. 1996; McPherson et al. 2004) and its effects on model performance has been frequently tested (e.g., McPherson et al. 2004; Wogan 2016). On the one hand, compared with narrow-ranging species, wide-ranging species typically have a higher number of occurrences, which should result in SDMs with better performance (e.g., Meyer et al. 2016; Yang et al. 2021). On the other hand, since range size is positively associated with species niche breadth (Galiana et al. 2023), wide-ranging species might have less accurate pseudo-absences than narrow-ranging species, potentially leading to poorer model performance (e.g., Hernandez et al. 2006; Newbold et al. 2009). We determined species range size by calculating the size of IUCN expert range maps.
- Latitude: There is a latitudinal gradient in knowledge of biodiversity information and low-latitude species generally have incomplete distribution information (Menegotto and Rangel 2018; Hughes et al. 2021). Therefore, we hypothesise that SDMs for low-latitude fish species might have poorer predictive performance. We determined species' latitude by calculating the median absolute latitude of species occurrence records.

TABLE 1 | Candidate factors that potentially affect predictive performance of species distribution models for tropical reef fishes.

Factor	Type	Value
Range size	Continuous	0.14– 225.54 × 10 ⁶ km ²
Latitude	Continuous	0.56°–57.38°
Environmental similarity	Continuous	44%–99%
Proximity to shore	Continuous	2–286 km
Range coverage	Continuous	–625 to –381 km
Data availability	Categorical	Data-deficient or data-sufficient
Species description year	Continuous	1758–2011
Body size	Continuous	1.2–539.2 cm
Trophic level	Continuous	1.9–4.5
Parental care	Categorical	Yes or no

- c. Environmental similarity between training and evaluation datasets: In SDM studies, it is problematic and unreliable to make predictions into novel environmental spaces (Elith et al. 2010; Nguyen and Leung 2022). We therefore hypothesise that model performance should be positively correlated with the degree of environmental similarity between training and evaluation datasets. As described earlier, we partitioned data into training and evaluation datasets via a four-fold spatial cross-validation approach. We measured environmental similarity between training and evaluation datasets using the multivariate environmental similarity surface (MESS) (Elith et al. 2010) and calculated the average proportion of positive MESS values across the four evaluation datasets. A higher proportion of positive MESS values indicates a greater level of environmental similarity between training and evaluation datasets.
- d. Proximity to shore: Sampling in marine systems is strongly constrained by accessibility and species near coastlines are generally better sampled (Hughes et al. 2021). Thus, we expect that SDMs for nearshore species will exhibit better performance. We quantified proximity to shore by calculating median distance of species occurrence records to shore.
- e. Range coverage: Distribution data are geographically biased, and such bias varies greatly among species (Meyer et al. 2016). Species with a higher level of bias in distribution data should produce poorer SDMs. Following Meyer et al. (2016), we measured species-level bias in occurrence records by calculating range coverage with the help of IUCN expert range maps and species occurrence records. For each species, (1) we randomly generated 1000 points within its IUCN expert range map, (2) for each random point, we calculated its minimum distance to species occurrence records, and (3) we determined range coverage as the negative mean minimum distance (see detailed explanation in Meyer et al. 2016). A larger value of range coverage indicates that the distribution data is less biased.
- f. Data availability: We retrieved conservation status of fish species from the IUCN Red List and classified them as either data-deficient (i.e., Data-Deficient) or data-sufficient (i.e., all other categories from Least Concern to Extinct). We hypothesise that data-sufficient species have received more research attention and monitoring, and thus their SDMs should perform better than those of data-deficient species.
- g. Species description year: Compared with newly described species, we should accumulate more abundant and complete distribution data for species described earlier. Hence, we hypothesise that SDMs for species described earlier should have higher performance. We extracted year of description information for each fish species from the World Register of Marine Species (<https://www.marinespecies.org>).
- h. Body size: The relationship between body size and model performance has been explored across taxa (Soininen and Luoto 2014). On the one hand, larger bodied species are expected to be more easily detected and monitored (McPherson and Jetz 2007), tend to have better-documented

distributions (Meyer et al. 2016), and allow for more accurate pseudo-absences generation, potentially improving model performance. On the other hand, larger-bodied fish are typically more mobile and more likely to escape from sampling efforts, which may result in incomplete detection of their true distributions and poor SDMs.

- i. Trophic level: The geographical distributions of species are influenced not only by abiotic environmental conditions but also by prey availability and food web dynamics (Soininen and Luoto 2014), which can be approximated with the trophic level. The effect of trophic level on SDM predictive performance remains inconclusive in the literature (e.g., Huntley et al. 2004; McPherson and Jetz 2007; Giannini et al. 2013; Luan et al. 2020). Though fish species at any trophic level are influenced by prey or food availability, we hypothesise that the distribution of higher-trophic-level species may be more difficult to model due to their dependence on complex food web dynamics, leading to reduced SDM performance.
- j. Spawning type: Fish species exhibiting parental care behaviour (e.g., mouthbrooding) often have more stable populations, higher offspring survival, and greater tolerance to environmental stressors (e.g., Gross and Sargent 1985). We therefore expect these species to be more easily monitored, resulting in improved SDM performance.

We obtained values for body size, trophic level and spawning type from Thorson et al. (2023), who predicted life-history attributes for over 32,000 fishes using phylogenetic comparative methods and structural equation models.

To explore the effects of these candidate factors on SDM performance, we fitted a generalised linear model (GLM) using a Bayesian framework. Continuous factors were z-transformed (zero mean and unit standard deviation) to improve model convergence and interpretability. All continuous predictors were included in the analysis as correlation among them was low (Figure S4). Although model performance metrics are not biological traits, phylogeny was included as a random effect to conservatively account for potential non-independence arising from shared, unobserved species attributes. Because closely related species often share similar characteristics (Li and Ives 2017), we accounted for phylogenetic non-independence by incorporating a random effect. We used the *r*trees R package (Li 2023) to construct a phylogeny for the tropical reef fish species based on a fish mega-tree (Rabosky et al. 2018) and calculated a variance–covariance matrix. For each model performance metric, we fitted a GLM using the *brms* R package, incorporating the variance–covariance matrix as a random effect (Bürkner 2017). We ran four independent chains of 20,000 iterations each, discarding the first 5000 as warm-up. We assessed the variance explained by each model using conditional R^2 . Factors were considered significant predictors if the 95% credible intervals of their coefficients did not overlap with zero. We assessed model convergence using the potential scale reduction factor ($\hat{R}$) and effective sample size (ESS). All parameters had $\hat{R}$ values close to 1 and sufficiently large ESS, indicating good convergence. We further computed the relative contribution of each factor in GLMs using

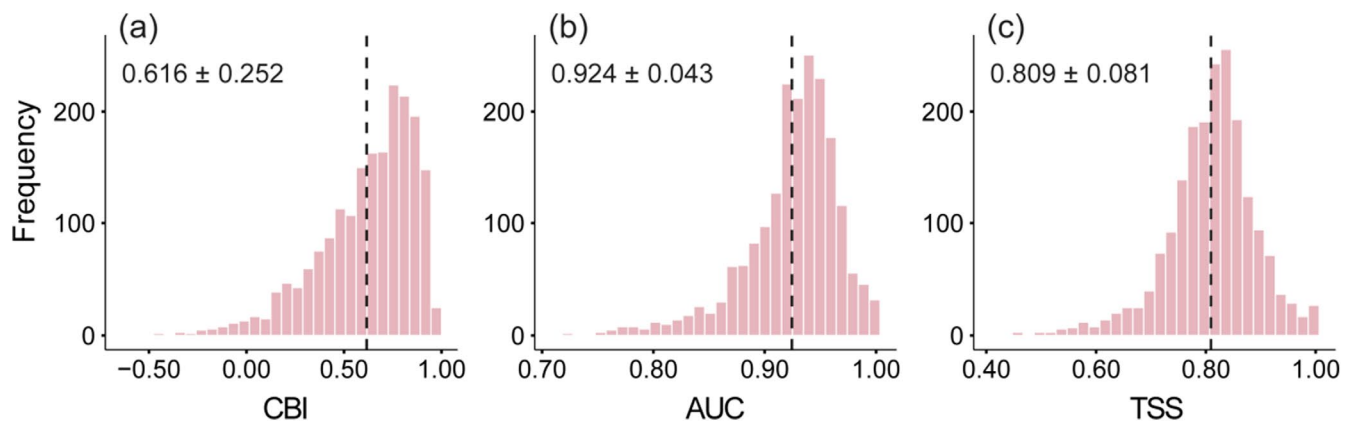


FIGURE 2 | Distributions of predictive performance for tropical reef fishes using the ensemble species distribution models. Model performance was assessed with three metrics: (a) continuous Boyce index, (b) area under the receiver operating characteristic curve, and (c) true skill statistic. The values in each panel represent the mean $\pm$ standard deviation of model performance. The vertical dashed lines indicate the mean values of model performance.

a hierarchical partitioning approach via the glmm.hp. R package (Lai et al. 2022).

3 | Results

3.1 | Ensemble Model Predictive Performance

Our results showed that regardless of evaluation metrics, the predictive performance of these four ensemble models was highly correlated (Pearson's r for CBI: 0.80–0.99; Pearson's r for AUC: 0.94–1.00; Pearson's r for TSS: 0.96–1.00; Figure 2 and Figures S5–S8). Moreover, the effects of candidate factors on model predictive performance were identical across four ensemble models (Figures 3–5 and Figures S9–S11). Therefore, for simplicity, we only presented results of average ensemble models in the main text.

Overall, ensemble SDMs for 1941 tropical reef fish species exhibited satisfactory predictive performance, with average ($\pm$ SD) CBI of 0.616 ($\pm$ 0.252), AUC of 0.924 ($\pm$ 0.043) and TSS of 0.809 ($\pm$ 0.081) (Figure 2). Across these three evaluation metrics, AUC and TSS were strongly positively correlated (Pearson's $r=0.94$, $p<0.001$), whereas CBI showed a weaker correlation with both AUC (Pearson's $r=0.44$, $p<0.001$) and TSS (Pearson's $r=0.23$, $p<0.001$) (Figure S12). Following the classification criteria of Engler et al. (2011) and Huang et al. (2024), models were considered good if CBI >0.4 , AUC >0.7 and TSS >0.4 . Based on these thresholds, we produced good models for 82% of tropical reef fishes (1591 out of 1941 species).

3.2 | Links Between Candidate Factors and Model Performance

The explanatory power of the GLMs relating model predictive performance to candidate factors varied across evaluation metrics. The corresponding conditional R^2 values were 0.341 for CBI, 0.254 for AUC and 0.253 for TSS, respectively. Across

these three evaluation metrics, three factors consistently emerged as significant predictors (Figures 3–5, Table S4). Environmental similarity between training and evaluation datasets was significantly positively correlated with predictive performance (i.e., the more similar the environments, the higher the predictive performance), while species' latitude (the higher the latitude, the better the model performance) and proximity to shore (the closer to shore, the better the model performance) were significantly negatively associated with predictive power (Figures 3–5).

Apart from these consistent patterns, we also observed contrasting results among evaluation metrics. The effect of range size and range coverage was significant across all three evaluation metrics, showing a positive association with CBI (Figure 3) but negative associations with AUC (Figure 4) and TSS (Figure 5). The effect of body size was significant in AUC only (Figure 4), while parental care behaviour exhibited a significant effect in CBI only (Figure 3). We found that species description year had a significant effect in both CBI and TSS, but in opposite directions (Figures 3 and 5). The relative contribution of these candidate factors in GLMs varied among evaluation metrics: range coverage was the most influential factor in GLM for CBI (Figure 3), whereas environmental similarity between training and evaluation datasets was the most important predictor in GLMs for AUC (Figure 4) and TSS (Figure 5).

4 | Discussion

We developed ensemble SDMs for nearly 2000 tropical reef fish species and assessed factors influencing model predictive performance. Proximity to shore, species' latitude, and environmental similarity between training and evaluation datasets consistently emerged as key predictors of SDM performance across all three evaluation metrics considered. The findings related to species characteristics offer a practical basis for identifying species that are more likely to yield reliable model predictions, with important implications for biodiversity analyses and the development of conservation strategies.

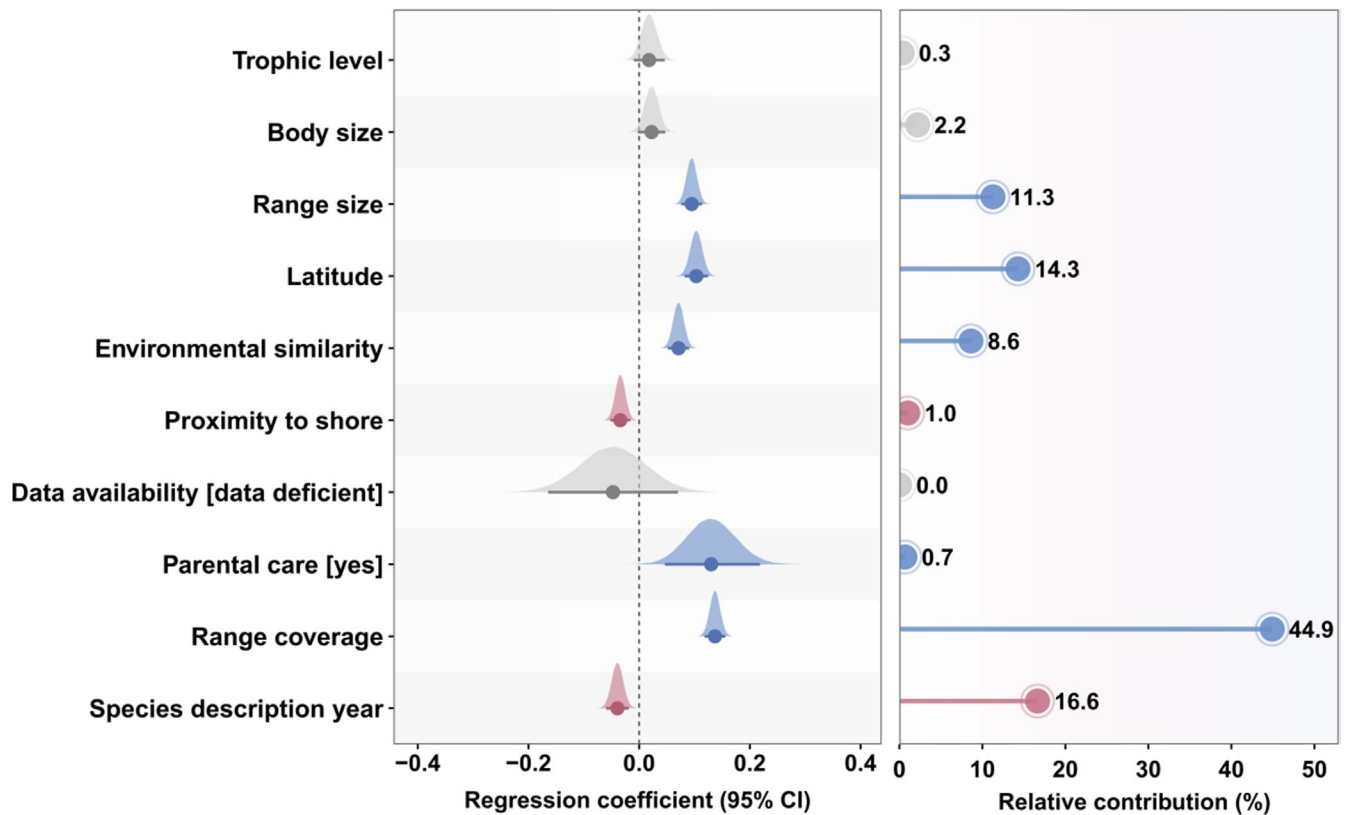


FIGURE 3 | Factors influencing the performance of the ensemble species distribution models for tropical reef fishes, measured by the continuous Boyce index. In the left panel, the dots represent the median values of the estimated posterior distributions, with horizontal lines indicating the 95% credible intervals. Significant effects are highlighted in blue (positive effect) and red (negative effect). The right panel illustrates the relative contribution of each factor to model performance.

4.1 | Factors Affecting Model Performance

The significant role of proximity to shore may result from biases inherent in occurrence data. Occurrence data for marine species are often concentrated in accessible nearshore coastal areas, whereas offshore regions are underrepresented due to logistical constraints (Hughes et al. 2021). Such sampling bias likely contributes to reduced SDM performance for offshore species. Apart from biases in occurrence data, the significant effect of proximity to shore might also reflect differences in site fidelity. Nearshore habitats are typically characterised by high structural complexity created by habitat-forming species such as corals and seagrasses, and fish inhabiting these environments tend to exhibit strong site fidelity and high habitat specialisation (e.g., Sale 1978; Streit et al. 2021). These characteristics increase the degree to which species distributions are constrained by environmental conditions (Brodie et al. 2021; Klaassen et al. 2025). As a result, the geographical distributions of nearshore species are expected to be more predictable from environmental variables than those of offshore species. Similarly, the positive relationship between species' latitude and SDM performance in tropical reef fishes can be partially attributed to the pronounced tropical gap in species occurrence records (e.g., Menegotto and Rangel 2018). Beyond data bias, the latitudinal variation in SDM performance may also reflect underlying gradients in biotic interactions across latitudes. For example, Longo et al. (2019) quantified latitudinal variation in trophic interactions in the Western Atlantic

and found that interactions were stronger in the tropics. This pattern suggests that biotic interactions exert a greater influence on the geographical distributions of low-latitude species than on those of high-latitude species. Consequently, SDMs ignoring biotic interactions are likely to have lower predictive accuracy for low-latitude species. This finding highlights the importance of considering biotic interactions in SDMs, especially for tropical species (e.g., Zhang et al. 2024; Qu et al. 2025).

Our results demonstrate that SDM predictive performance is positively correlated with the environmental similarity between training and evaluation datasets. This is expected because SDMs learn species-environment relationships from the training data, and these relationships can be reliably and confidently applied to evaluation datasets that occupy a similar environmental space (Elith et al. 2010). This result highlights the already-known inherent uncertainty in predicting species distributions under novel climatic conditions and serves as a caution against extrapolating SDMs across different spatial (e.g., biological invasions) and temporal (e.g., global change biology) scales (Araújo et al. 2019; Zurell et al. 2020).

We found great variation in the relative contribution of different factors across three evaluation metrics; besides, we also detected inconsistent effects of several factors (e.g., range size and body size) on model performance across evaluation metrics. These inconsistencies suggest that the choice of evaluation metric can

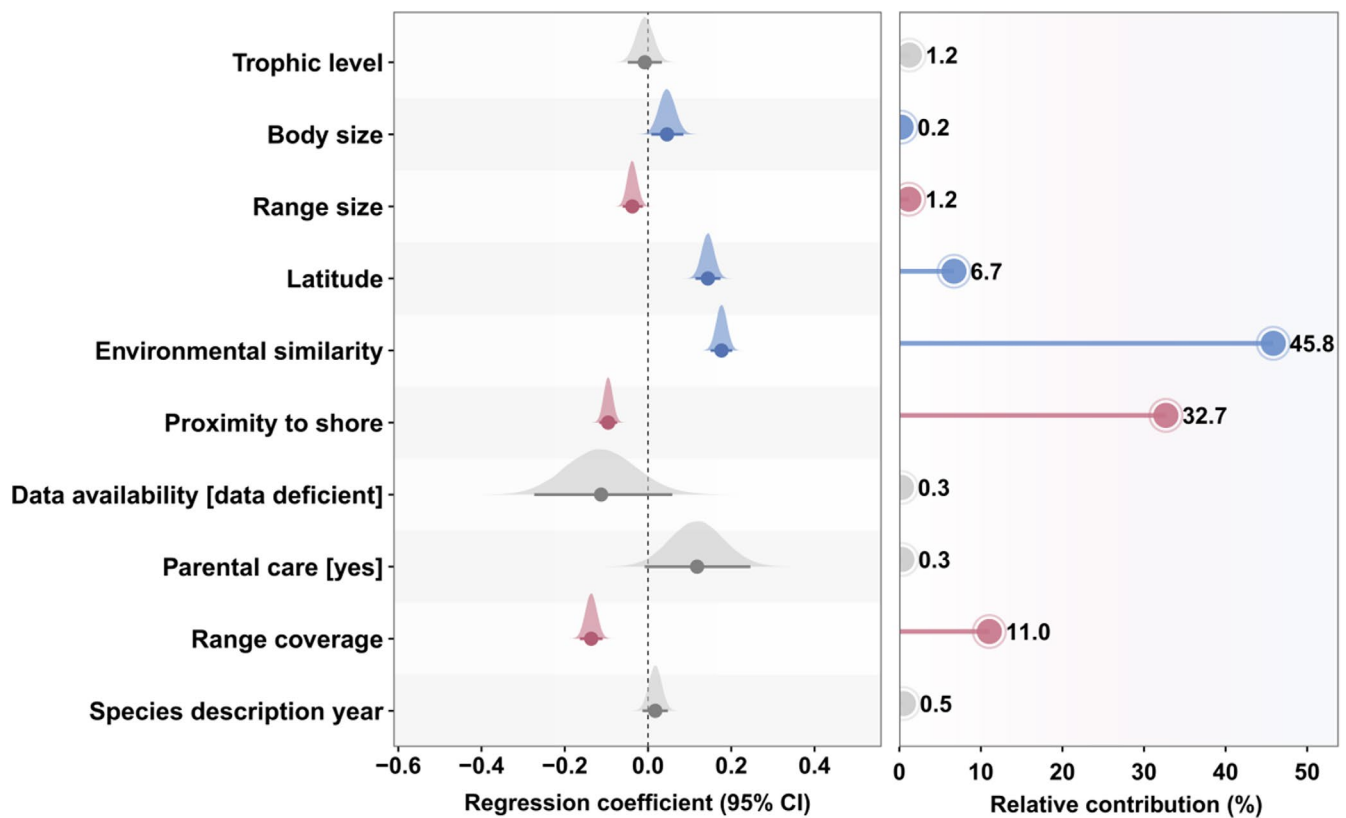


FIGURE 4 | Factors influencing the performance of the ensemble species distribution models for tropical reef fishes, measured by the area under the receiver operating characteristic curve. In the left panel, the dots represent the median values of the estimated posterior distributions, and the horizontal lines indicate the 95% credible intervals. Significant effects are highlighted in blue (positive effect) and red (negative effect). The right panel shows the relative contribution of each factor to model performance.

affect how species traits relate to model performance. To ensure robust inferences, we recommend using multiple evaluation metrics in future studies. In our study, we developed SDMs for tropical reef fish species using presence-background data due to the lack of quality absence data. Under such circumstances, evaluation metrics including AUC (Lobo et al. 2008) and TSS (Leroy et al. 2018) are misleading whereas CBI represents a reliable metric (Hirzel et al. 2006). The phylogenetic GLM for CBI further identified the significant effects of range size, parental care, range coverage and species description year on model performance. We detected a significant positive correlation between range size and CBI values. Researchers have proposed a variety of hypotheses to explain interspecific variation in species' range sizes, among which dispersal ability is considered a key determinant, with wide-ranging species generally exhibiting higher dispersal ability than narrow-ranging species (e.g., Brown et al. 1996). Consequently, compared with wide-ranging species, narrow-ranging species may occupy only a limited proportion of their environmentally suitable geographical space due to dispersal constraints. As a result, species-environment relationships inferred from SDMs for narrow-ranging species are more likely to be truncated, leading to lower predictive performance. Our results showed that SDMs for species with parental care had higher CBI. Parental care enhances offspring survival and buffers against harsh environmental conditions (Guindre-Parker and Rubenstein 2018; Gross and Sargent 1985), leading to greater temporal stability in population persistence (Takeuchi et al. 2009). As a consequence, occurrence records of

species with parental care are more consistently aligned with environmentally suitable conditions through time, capturing a stable species-environment signal that enhances the predictive performance of SDMs. Regarding species description year, as expected, SDMs for recently described tropical reef fish species exhibited poor predictive power. These newly described species are typically cryptic and their geographical range is restricted, confined to remote or poorly surveyed regions (Ceballos and Ehrlich 2009). Substantial gaps in ecological knowledge for newly described species might limit the ability of SDMs to accurately characterise their environmental requirements. Together with evidence that newly described species face elevated extinction risk (Liu et al. 2022), our finding highlights the need to prioritise conservation efforts for newly described species. We found a significantly positive association between range coverage and CBI score. This is reasonable because species with high range coverage have less biased occurrence records and are more likely to produce SDMs with high predictive performance.

Despite the robustness of our findings, some limitations should be noted. The explanatory power of our phylogenetic GLMs was modest ($R^2 < 0.35$), which is quite common in large-scale and multi-taxa ecological studies (e.g., McPherson and Jetz 2007). We hypothesise that different mechanisms might account for the modest explanatory power in our study, such as missing direct factors in GLMs and/or non-linear effects of these candidate factors. Besides, we built SDMs using a uniform predictor set, which might overlook species-specific predictors such as

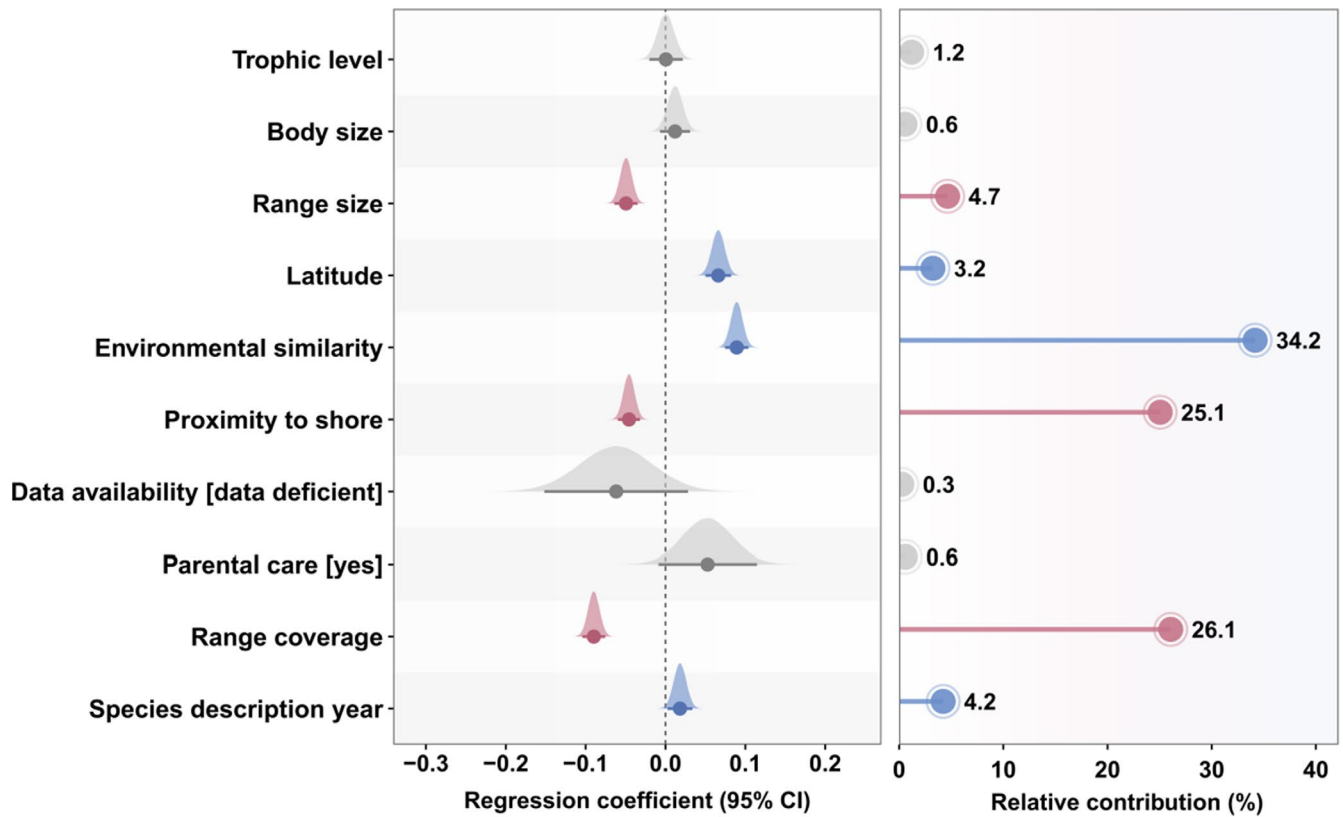


FIGURE 5 | Factors influencing the performance of the ensemble species distribution models for tropical reef fishes, measured by the true skill statistic. In the left panel, the dots represent the median values of the estimated posterior distributions, and the horizontal lines indicate the 95% credible intervals. Significant effects are highlighted in blue (positive effect) and red (negative effect). The right panel shows the relative contribution of each factor to model performance.

habitat and biotic interactions. We encourage future studies to account for these predictors if possible. Following best-practice guidelines, we built SDMs by generating 10,000 random pseudo-absence data. It has been demonstrated that pseudo-absence data generation greatly influences the predictive performance of SDMs (e.g., VanDerWal et al. 2009; Barbet-Massin et al. 2012; Steen et al. 2024), which may in turn alter the associations between candidate factors and model performance. We encourage further studies to clarify these issues.

4.2 | Implications for Marine Biodiversity Studies

Our findings yield two key recommendations for improving marine biodiversity modelling and conservation.

First, given the negative effects of proximity to shore and positive effects of species' latitude on SDM performance, expanding data collection in less-accessible marine areas (e.g., offshore and low-latitude areas) is essential. Currently, most available data come from easily surveyed areas (i.e., nearshore shallow waters), charismatic taxa, and developed countries. For example, marine species are better documented near research stations (Rubio-López et al. 2023), while soft-bodied or microscopic organisms remain understudied (Martínez et al. 2025). Since data availability often influences conservation funding (e.g., in birds; Correia et al. 2024), we emphasise the importance of targeting underexplored taxa and regions.

To address these disparities, we recommend increasing targeted research funding, promoting interdisciplinary collaboration, and leveraging policy mechanisms to support biodiversity data collection. Additionally, emerging technologies and community-based efforts can greatly expand monitoring coverage. For example, environmental DNA has proven to be a non-invasive and effective tool for biodiversity monitoring, even in deep-sea habitats (see review by Lu et al. 2024). Likewise, citizen science platforms such as iNaturalist (www.inaturalist.org) provide high volumes of high-quality observations, contributing to biodiversity knowledge and conservation efforts.

Second, broader biodiversity conservation strategies are needed to address systematic biases in data availability. Our findings showed that range coverage, a metric measuring bias in distribution data, has a significant effect (though different effects across metrics) on model accuracy. Combined with prior reports of geographic and taxonomic biases in biodiversity data (e.g., Meyer et al. 2016; Hughes et al. 2021) and unequal conservation funding across taxa (e.g., Mammola et al. 2020; Correia et al. 2024; Guénard et al. 2025), these results underscore the urgency of pursuing more equitable biodiversity targets.

Future efforts should focus on understanding the origins of bias, improving data coverage for poorly studied species, and optimising the allocation of conservation resources. In marine regions where biodiversity data are particularly scarce or inaccessible, recent developments in artificial intelligence offer

promising solutions. For example, large language models have shown great potential for extracting biodiversity data from unstructured text (e.g., Castro et al. 2024; Branco et al. 2025). These tools can not only fill information gaps in biodiversity data and enhance spatial predictions but also generate novel hypotheses that guide future field surveys, as highlighted by Pollock et al. (2025). Only through such integrated and equitable strategies can we achieve meaningful progress in safeguarding marine biodiversity.

Author Contributions

Z.Z. and D.W. conceived the research idea and designed the study. Y.L. and S.L. collected data with help from L.D., S.M., and G.N. Y.L. led data analyses with critical suggestions from Á.B.-F., S.M., L.D. Z.Z., Y.L., and D.W. drafted the manuscript with critical suggestions from all co-authors, especially Á.B.-F. and S.M.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (42230409 and 42276126), the National Key Research and Development Program of China (2023YFC3108800), and the Guangdong Basic and Applied Basic Research Foundation (2024A151010604). Á.B.-F. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. This work has been implemented by the National Multidisciplinary Laboratory for Climate Change (RRF-2.3.1-21-2022-00014) project within the framework of Hungary's National Recovery and Resilience Plan supported by the Recovery and Resilience Facility of the European Union.

Funding

This work was funded by the National Natural Science Foundation of China (42230409 and 42276126), the National Key Research and Development Program of China (2023YFC3108800), and the Guangdong Basic and Applied Basic Research Foundation (2024A151010604). Á.B.-F. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. This work has been implemented by the National Multidisciplinary Laboratory for Climate Change (RRF-2.3.1-21-2022-00014) project within the framework of Hungary's National Recovery and Resilience Plan supported by the Recovery and Resilience Facility of the European Union.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Species distribution data were retrieved from the Global Biodiversity Information Facility (<https://www.gbif.org>) and are available at <https://doi.org/10.15468/dl.w7zpa6>. Marine data layers were extracted from the Bio-ORACLE database version 2.2 (<http://www.bio-oracle.org>). Candidate factors influencing model performance are available in Table S1.

Peer Review

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ddi.70160>.

References

Allouche, O., A. Tsoar, and R. Kadmon. 2006. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." *Journal of Applied Ecology* 43: 1223–1232.

Araújo, M. B., R. P. Anderson, A. Márcia Barbosa, et al. 2019. "Standards for Distribution Models in Biodiversity Assessments." *Science Advances* 5: eaat4858.

Assis, J., L. Tyberghein, S. Bosch, H. Verbruggen, E. A. Serrão, and O. De Clerck. 2018. "Bio-ORACLE v2. 0: Extending Marine Data Layers for Bioclimatic Modelling." *Global Ecology and Biogeography* 27: 277–284.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. "Selecting Pseudo-Absences for Species Distribution Models: How, Where and How Many?" *Methods in Ecology and Evolution* 3: 327–338.

Bradie, J., and B. Leung. 2017. "A Quantitative Synthesis of the Importance of Variables Used in MaxEnt Species Distribution Models." *Journal of Biogeography* 44: 1344–1361.

Branco, V. V., J. Benedek, L. Pivovarov, L. Correia, and P. Cardoso. 2025. "ARETE: An R Package for Automated Retrieval From TExt With Large Language Models." arXiv preprint arXiv:2511.04573.

Brodie, S., B. Abrahms, S. J. Bograd, et al. 2021. "Exploring Timescales of Predictability in Species Distributions." *Ecography* 44: 832–844.

Brown, J. H., G. C. Stevens, and D. M. Kaufman. 1996. "The Geographic Range: Size, Shape, Boundaries, and Internal Structure." *Annual Review of Ecology and Systematics* 27: 597–623.

Brun, P., W. Thuiller, Y. Chauvier, et al. 2020. "Model Complexity Affects Species Distribution Projections Under Climate Change." *Journal of Biogeography* 47: 130–142.

Bürkner, P. C. 2017. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80: 1–28.

Castro, A., J. Pinto, L. Reino, P. Pipek, and C. Capinha. 2024. "Large Language Models Overcome the Challenges of Unstructured Text Data in Ecology." *Ecological Informatics* 82: 102742.

Ceballos, G., and P. R. Ehrlich. 2009. "Discoveries of New Mammal Species and Their Implications for Conservation and Ecosystem Services." *Proceedings of the National Academy of Sciences* 106: 3841–3846.

Cheung, W. W., T. L. Frölicher, V. W. Lam, et al. 2021. "Marine High Temperature Extremes Amplify the Impacts of Climate Change on Fish and Fisheries." *Science Advances* 7: eabh0895.

Chevalier, M., O. Broennimann, and A. Guisan. 2024. "Climate Change May Reveal Currently Unavailable Parts of Species' Ecological Niches." *Nature Ecology & Evolution* 8: 1–13.

Collart, F., O. Broennimann, A. Guisan, and A. Vanderpoorten. 2023. "Ecological and Biological Indicators of the Accuracy of Species Distribution Models: Lessons From European Bryophytes." *Ecography* 2023: e06721.

Correia, R. A., J. E. Brommer, A. Haukka, et al. 2024. "Conservation Needs and Opportunities Drive LIFE Funding Allocation for European Birds." *Biological Conservation* 300: 110833.

Cowie, R. H., P. Bouchet, and B. Fontaine. 2022. "The Sixth Mass Extinction: Fact, Fiction or Speculation?" *Biological Reviews* 97: 640–663.

Dormann, C. F., J. Elith, S. Bacher, et al. 2013. "Collinearity: A Review of Methods to Deal With It and a Simulation Study Evaluating Their Performance." *Ecography* 36: 27–46.

Elith, J., M. Kearney, and S. Phillips. 2010. "The Art of Modelling Range-Shifting Species." *Methods in Ecology and Evolution* 1: 330–342.

Elith, J., S. J. Phillips, T. Hastie, M. Dudik, Y. E. Chee, and C. J. Yates. 2011. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions* 17: 43–57.

Engler, R., C. F. Randin, W. Thuiller, et al. 2011. "21st Century Climate Change Threatens Mountain Flora Unequally Across Europe." *Global Change Biology* 17: 2330–2341.

- Eskildsen, A., P. C. le Roux, R. K. Heikkinen, et al. 2013. "Testing Species Distribution Models Across Space and Time: High Latitude Butterflies and Recent Warming." *Global Ecology and Biogeography* 22: 1293–1303.
- Feng, X., D. S. Park, C. Walker, A. T. Peterson, C. Merow, and M. Papeš. 2019. "A Checklist for Maximizing Reproducibility of Ecological Niche Models." *Nature Ecology & Evolution* 3: 1382–1395.
- Galiana, N., M. Lurgi, J. M. Montoya, M. B. Araújo, and E. D. Galbraith. 2023. "Climate or Diet? The Importance of Biotic Interactions in Determining Species Range Size." *Global Ecology and Biogeography* 32: 1178–1188.
- GBIF.org. 2025. "GBIF Occurrence." <https://doi.org/10.15468/dl.w7zpa6>.
- Giannini, T. C., D. S. Chapman, A. M. Saraiva, I. Alves-dos-Santos, and J. C. Biesmeijer. 2013. "Improving Species Distribution Models Using Biotic Interactions: A Case Study of Parasites, Pollinators and Plants." *Ecography* 36: 649–656.
- Gross, M. R., and R. C. Sargent. 1985. "The Evolution of Male and Female Parental Care in Fishes." *American Zoologist* 25: 807–822.
- Guénard, B., A. C. Hughes, C. Lainé, S. Cannicci, B. D. Russell, and G. A. Williams. 2025. "Limited and Biased Global Conservation Funding Means Most Threatened Species Remain Unsupported." *Proceedings of the National Academy of Sciences* 122: e2412479122.
- Guindre-Parker, S., and D. R. Rubenstein. 2018. "Multiple Benefits of Alloparental Care in a Fluctuating Environment." *Royal Society Open Science* 5: 172406.
- Guisan, A., and W. Thuiller. 2005. "Predicting Species Distribution: Offering More Than Simple Habitat Models." *Ecology Letters* 8: 993–1009.
- Guisan, A., N. E. Zimmermann, J. Elith, C. H. Graham, S. Phillips, and A. T. Peterson. 2007. "What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics?" *Ecological Monographs* 77: 615–630.
- Hanley, J. A., and B. J. McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology* 143: 29–36.
- Hanssen, A. W., and W. J. A. Kuipers. 1965. "On the Relationship Between the Frequency of Rain and Various Meteorological Parameters." *Koninklijk Nederlands Meteorologisch Instituut* 81: 2–25.
- Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. "The Effect of Sample Size and Species Characteristics on Performance of Different Species Distribution Modeling Methods." *Ecography* 29: 773–785.
- Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin, and A. Guisan. 2006. "Evaluating the Ability of Habitat Suitability Models to Predict Species Presences." *Ecological Modelling* 199: 142–152.
- Hu, Z. M., Q. S. Zhang, J. Zhang, et al. 2021. "Intraspecific Genetic Variation Matters When Predicting Seagrass Distribution Under Climate Change." *Molecular Ecology* 30: 3840–3855.
- Huang, H., Z. Zhang, Á. Bede-Fazekas, et al. 2024. "Cross-Validation Matters in Species Distribution Models: A Case Study With Goatfish Species." *Ecography* 2025: e07354.
- Hughes, A. C., M. C. Orr, K. Ma, et al. 2021. "Sampling Biases Shape Our View of the Natural World." *Ecography* 44: 1259–1269.
- Huntley, B., R. E. Green, Y. C. Collingham, et al. 2004. "The Performance of Models Relating Species Geographical Distributions to Climate Is Independent of Trophic Level." *Ecology Letters* 7: 417–426.
- Jaureguiberry, P., N. Titeux, M. Wiemers, et al. 2022. "The Direct Drivers of Recent Global Anthropogenic Biodiversity Loss." *Science Advances* 8: eabm9982.
- Kass, J. M., R. Muscarella, P. J. Galante, et al. 2021. "ENMeval 2.0: Redesigned for Customizable and Reproducible Modeling of Species' Niches and Distributions." *Methods in Ecology and Evolution* 12: 1602–1608.
- Klaassen, M., T. A. Marques, F. Alves, and M. Fernandez. 2025. "Trends in Marine Species Distribution Models: A Review of Methodological Advances and Future Challenges." *Ecography* 48: e07702.
- Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, et al. 2013. "The Importance of Correcting for Sampling Bias in MaxEnt Species Distribution Models." *Diversity and Distributions* 19: 1366–1379.
- Lai, J., Y. Zou, S. Zhang, X. Zhang, and L. Mao. 2022. "Glm.hp: An R Package for Computing Individual Effect of Predictors in Generalized Linear Mixed Models." *Journal of Plant Ecology* 15: 1302–1307.
- Langlois, J., F. Guilhaumon, F. Baletaud, et al. 2022. "The Aesthetic Value of Reef Fishes Is Globally Mismatched to Their Conservation Priorities." *PLoS Biology* 20: e3001640.
- Leroy, B., R. Delsol, B. Hugué, et al. 2018. "Without Quality Presence–Absence Data, Discrimination Metrics Such as TSS Can Be Misleading Measures of Model Performance." *Journal of Biogeography* 45: 1994–2002.
- Li, D. 2023. "Rtrees: An R Package to Assemble Phylogenetic Trees From Megatrees." *Ecography* 2023: e06643.
- Li, D., and A. R. Ives. 2017. "The Statistical Need to Include Phylogeny in Trait-Based Analyses of Community Composition." *Methods in Ecology and Evolution* 8: 1192–1199.
- Liu, J., F. Slik, S. Zheng, and D. B. Lindenmayer. 2022. "Undescribed Species Have Higher Extinction Risk Than Known Species." *Conservation Letters* 15: e12876.
- Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." *Global Ecology and Biogeography* 17: 145–151.
- Longo, G. O., M. E. Hay, C. E. Ferreira, and S. R. Floeter. 2019. "Trophic Interactions Across 61 Degrees of Latitude in the Western Atlantic." *Global Ecology and Biogeography* 28: 107–117.
- Lu, S., H. Zeng, F. Xiong, M. Yao, and S. He. 2024. "Advances in Environmental DNA Monitoring: Standardization, Automation, and Emerging Technologies in Aquatic Ecosystems." *Science China Life Sciences* 67: 1368–1384.
- Luan, J., C. Zhang, B. Xu, Y. Xue, and Y. Ren. 2020. "The Predictive Performances of Random Forest Models With Limited Sample Size and Different Species Traits." *Fisheries Research* 227: 105534.
- Mammola, S., N. Riccardi, V. Prié, et al. 2020. "Towards a Taxonomically Unbiased European Union Biodiversity Strategy for 2030." *Proceedings of the Royal Society B* 287: 20202166.
- Martínez, A., S. Bonaglia, M. Di Domenico, et al. 2025. "Fundamental Questions in Meiofauna Research Highlight How Small but Ubiquitous Animals Can Improve Our Understanding of Nature." *Communications Biology* 8: 449.
- McPherson, J. M., and W. Jetz. 2007. "Effects of Species' Ecology on the Accuracy of Distribution Models." *Ecography* 30: 135–151.
- McPherson, J. M., W. Jetz, and D. J. Rogers. 2004. "The Effects of Species' Range Sizes on the Accuracy of Distribution Models: Ecological Phenomenon or Statistical Artefact?" *Journal of Applied Ecology* 41: 811–823.
- Mellin, C., C. J. A. Bradshaw, M. G. Meekan, and M. J. Caley. 2010. "Environmental and Spatial Predictors of Species Richness and Abundance in Coral Reef Fishes." *Global Ecology and Biogeography* 19: 212–222.
- Melo-Merino, S. M., H. Reyes-Bonilla, and A. Lira-Noriega. 2020. "Ecological Niche Models and Species Distribution Models in Marine

- Environments: A Literature Review and Spatial Analysis of Evidence." *Ecological Modelling* 415: 108837.
- Menegotto, A., and T. F. Rangel. 2018. "Mapping Knowledge Gaps in Marine Diversity Reveals a Latitudinal Gradient of Missing Species Richness." *Nature Communications* 9: 4713.
- Meyer, C., W. Jetz, R. P. Guralnick, S. A. Fritz, and H. Kreft. 2016. "Range Geometry and Socio-Economics Dominate Species-Level Biases in Occurrence Information." *Global Ecology and Biogeography* 25: 1181–1193.
- Morán-Ordóñez, A., J. J. Lahoz-Monfort, J. Elith, and B. A. Wintle. 2017. "Evaluating 318 Continental-Scale Species Distribution Models Over a 60-Year Prediction Horizon: What Factors Influence the Reliability of Predictions?" *Global Ecology and Biogeography* 26: 371–384.
- Mouquet, N., J. Langlois, N. Casajus, et al. 2024. "Low Human Interest for the Most At-Risk Reef Fishes Worldwide." *Science Advances* 10: ead9510.
- Newbold, T., T. Reader, S. Zalut, A. El-Gabbas, and F. Gilbert. 2009. "Effect of Characteristics of Butterfly Species on the Accuracy of Distribution Models in an Arid Environment." *Biodiversity and Conservation* 18: 3629–3641.
- Nguyen, D., and B. Leung. 2022. "How Well Do Species Distribution Models Predict Occurrences in Exotic Ranges?" *Global Ecology and Biogeography* 31: 1051–1065.
- Parravicini, V., M. G. Bender, S. Villéger, et al. 2021. "Coral Reef Fishes Reveal Strong Divergence in the Prevalence of Traits Along the Global Diversity Gradient." *Proceedings of the Royal Society B: Biological Sciences* 288: 20211712.
- Parravicini, V., M. Kulbicki, D. R. Bellwood, et al. 2013. "Global Patterns and Predictors of Tropical Reef Fish Species Richness." *Ecography* 36: 1254–1262.
- Parravicini, V., S. Villéger, T. R. McClanahan, et al. 2014. "Global Mismatch Between Species Richness and Vulnerability of Reef Fish Assemblages." *Ecology Letters* 17: 1101–1110.
- Pollock, L. J., J. Kitzes, S. Beery, et al. 2025. "Harnessing Artificial Intelligence to Fill Global Shortfalls in Biodiversity Knowledge." *Nature Reviews Biodiversity* 1: 17.
- Qu, J., Z. Zhang, S. Ma, et al. 2025. "Habitat-Forming Species: Buffers or Amplifiers for Mutualistic Organisms in Response to Climate Change?" *Journal of Biogeography* 52: e15174.
- Rabosky, D. L., J. Chang, P. F. Cowman, et al. 2018. "An Inverse Latitudinal Gradient in Speciation Rate for Marine Fishes." *Nature* 559: 392–395.
- Rubio-López, I., F. Pardos, D. Fontaneto, A. Martínez, and G. García-Gómez. 2023. "Biases and Distribution Patterns in Hard-Bodied Microscopic Animals (Acari: *Halacaridae*): Size Does Not Matter, but Generalism and Sampling Effort Do." *Diversity and Distributions* 29: 821–833.
- Sale, P. F. 1978. "Coexistence of Coral Reef Fishes—A Lottery for Living Space." *Environmental Biology of Fishes* 3: 85–102.
- Santini, L., A. Benítez-López, L. Maiorano, M. Čengić, and M. A. Huijbregts. 2021. "Assessing the Reliability of Species Distribution Projections in Climate Change Research." *Diversity and Distributions* 27: 1035–1050.
- Soininen, J., and M. Luoto. 2014. "Predictability in Species Distributions: A Global Analysis Across Organisms and Ecosystems." *Global Ecology and Biogeography* 23: 1264–1274.
- Somodi, I., N. Lepesi, and Z. Botta-Dukát. 2017. "Prevalence Dependence in Model Goodness Measures With Special Emphasis on True Skill Statistics." *Ecology and Evolution* 7: 863–872.
- Steen, B., O. Broennimann, L. Maiorano, and A. Guisan. 2024. "How Sensitive Are Species Distribution Models to Different Background Point Selection Strategies? A Test With Species at Various Equilibrium Levels." *Ecological Modelling* 493: 110754.
- Stockwell, D. R., and A. T. Peterson. 2002. "Effects of Sample Size on Accuracy of Species Distribution Models." *Ecological Modelling* 148: 1–13.
- Streit, R. P., C. R. Hemingson, G. S. Cumming, and D. R. Bellwood. 2021. "How Flexible Are Habitat Specialists? Short-Term Space Use in Obligate Coral-Dwelling Damselfishes." *Reviews in Fish Biology and Fisheries* 31: 381–398.
- Takeuchi, Y., W. Wang, S. Nakaoka, and S. Iwami. 2009. "Dynamical Adaptation of Parental Care." *Bulletin of Mathematical Biology* 71: 931–951.
- Thorson, J. T., A. A. Maureaud, R. Frelat, et al. 2023. "Identifying Direct and Indirect Associations Among Traits by Merging Phylogenetic Comparative Methods and Structural Equation Models." *Methods in Ecology and Evolution* 14: 1259–1275.
- Thuiller, W., M. Guéguen, J. Renaud, D. N. Karger, and N. E. Zimmermann. 2019. "Uncertainty in Ensembles of Global Biodiversity Scenarios." *Nature Communications* 10: 1446.
- Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Aroita. 2019. "blockCV: An R Package for Generating Spatially or Environmentally Separated Folds for *k*-Fold Cross-Validation of Species Distribution Models." *Methods in Ecology and Evolution* 10: 225–232.
- Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Aroita. 2021. "Modelling Species Presence-Only Data With Random Forests." *Ecography* 44: 1731–1742.
- Valavi, R., G. Guillera-Aroita, J. J. Lahoz-Monfort, and J. Elith. 2022. "Predictive Performance of Presence-Only Species Distribution Models: A Benchmark Study With Reproducible Code." *Ecological Monographs* 92: e01486.
- VanDerWal, J., L. P. Shoo, C. Graham, and S. E. Williams. 2009. "Selecting Pseudo-Absence Data for Presence-Only Distribution Modeling: How Far Should You Stray From What You Know?" *Ecological Modelling* 220: 589–594.
- Webb, T. J. 2012. "Marine and Terrestrial Ecology: Unifying Concepts, Revealing Differences." *Trends in Ecology & Evolution* 27: 535–541.
- Whitford, A. M., B. R. Shipley, and J. L. McGuire. 2024. "The Influence of the Number and Distribution of Background Points in Presence-Background Species Distribution Models." *Ecological Modelling* 488: 110604.
- Wiethase, J. H., P. S. Mostert, C. R. Cooney, R. B. O'Hara, and C. M. Beale. 2024. "Spatio-Temporal Integrated Bayesian Species Distribution Models Reveal Lack of Broad Relationships Between Traits and Range Shifts." *Global Ecology and Biogeography* 33: e13819.
- Wisz, M. S., R. J. Hijmans, J. Li, et al. 2008. "Effects of Sample Size on the Performance of Species Distribution Models." *Diversity and Distributions* 14: 763–773.
- Wogan, G. O. 2016. "Life History Traits and Niche Instability Impact Accuracy and Temporal Transferability for Historically Calibrated Distribution Models of North American Birds." *PLoS One* 11: e0151024.
- Yang, W., D. Liu, Q. You, et al. 2021. "Taxonomic Bias in Occurrence Information of Angiosperm Species in China." *Science China. Life Sciences* 64: 584–592.
- Zbinden, R., N. Van Tiel, B. Kellenberger, L. Hughes, and D. Tuia. 2024. "On the Selection and Effectiveness of Pseudo-Absences for Species Distribution Modeling With Deep Learning." *Ecological Informatics* 81: 102623.
- Zhang, Z., J. M. Kass, Á. Bede-Fazekas, et al. 2025. "Differences in Predictions of Marine Species Distribution Models Based on Expert Maps and Opportunistic Occurrences." *Conservation Biology* 39: e70015.

Zhang, Z., S. Ma, Á. Bede-Fazekas, et al. 2024. "Considering Biotic Interactions Exacerbates the Predicted Impacts of Climate Change on Coral-Dwelling Species." *Journal of Biogeography* 51: 769–782.

Zhang, Z., S. Ma, Á. Bede-Fazekas, et al. 2025. "Integrating Expert Range Maps and Opportunistic Occurrence Records of Marine Fish Species in Range Estimates." *Conservation Biology* 39: e70154.

Zizka, A., D. Silvestro, T. Andermann, et al. 2019. "CoordinateCleaner: Standardized Cleaning of Occurrence Records From Biological Collection Databases." *Methods in Ecology and Evolution* 10: 744–751.

Zurell, D., J. Franklin, C. König, et al. 2020. "A Standard Protocol for Reporting Species Distribution Models." *Ecography* 43: 1261–1277.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** ddi70160-sup-0001-Supinfo1.pdf. **Figures S1–S12:** ddi70160-sup-0002-Figures.pdf. **Table S1:** ddi70160-sup-0003-TableS1.pdf. **Table S2:** ddi70160-sup-0004-TableS2.pdf. **Table S3:** ddi70160-sup-0005-TableS3.pdf. **Table S4:** ddi70160-sup-0006-TableS4.pdf.