SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Improving prediction of rare species' distribution from community data

Chongliang Zhang[1], Yong Chen[2], Binduo Xu[1], Ying Xue[1] & Yiping Ren[1,3,4]✉

Species distribution models (SDMs) have been increasingly used to predict the geographic distribution of a wide range of organisms; however, relatively fewer research efforts have concentrated on rare species despite their critical roles in biological conservation. The present study tested whether community data may improve modelling rare species by sharing information among common and rare ones. We chose six SDMs that treat community data in different ways, including two traditional single-species models (random forest and artificial neural network) and four joint species distribution models that incorporate species associations implicitly (multivariate random forest and multi-response artificial neural network) or explicitly (hierarchical modelling of species communities and generalized joint attribute model). In addition, we evaluated two approaches of data arrangement, species filtering and conditional prediction, to enhance the selected models. The model predictions were tested using cross validation based on empirical data collected from marine fisheries surveys, and the effects of community data were evaluated by comparing models for six selected rare species. The results demonstrated that the community data improved the predictions of rare species' distributions to certain extent but might also be unhelpful in some cases. The rare species could be appropriately predicted in terms of occurrence, whereas their abundance tended to be underestimated by most models. Species filtering and conditional predictions substantially benefited the predictive performances of multiple- and single-species models, respectively. We conclude that both the modelling algorithms and community data need to be carefully selected in order to deliver improvement in modelling rare species. The study highlights the opportunity and challenges to improve prediction of rare species' distribution by making the most of community data.

Species distribution model (SDMs) have been widely used to evaluate ecological niches and to predict geographic distribution of organisms across terrestrial, freshwater, and marine habitats[1–6]. A majority of SDMs have been developed for common and economically important species because of practical incentives, while predictive models are more challengeable for rare species due to methodological difficulties[7–9]. As most species are rare in natural biological communities[10,11], modeling common species cannot depict the full picture of biodiversity. In addition, rare species, characterized by low occurrence, are particularly vulnerable to environmental changes and human impacts thus deserve special concerns in biological conservation[8,12]. As such, there is a pressing need to predict the distribution of rare species for successful conservation in the practices of designing marine protected areas (MPAs) and identifying priorities for monitoring programs[13].

Accurate prediction of rare species is not easy. The difficulties come largely from the limits of data, as the observations of rare species are typically sparse in terms of spatial location and temporal frequency[14–16]. The sparse data imply that the number of presence observations is often small compared to the number of influential predictors, resulting in a critical problem of over-fitting in modelling[8,16,17]. Besides, occurrence or abundance of rare species are often vulnerable to sampling errors, which may lead to model misspecification, making it unfeasible to characterize species' niche space[18]. There are a few studies aiming to address the issue of rarity, e.g., by developing a large number of simple models averaged in an ensemble[8,9], and generating pseudo-absence from

[1]College of Fisheries, Ocean University of China, 216, Fisheries Hall, 5 Yushan Road, Qingdao 266003, China. [2]School of Marine Sciences, University of Maine, Libby Hall, Orono, ME 21604469, USA. [3]Field Observation and Research Station of Haizhou Bay Fishery Ecosystem, Ministry of Education, Qingdao 266003, China. [4]Laboratory for Marine Fisheries Science and Food Production Processes, Pilot National Laboratory for Marine Science and Technology (Qingdao), 1 Wenhai Road, Qingdao 266237, China. ✉email: renyip@ouc.edu.cn

a habitat suitability map[19–21]. In spite of the progress, many issues remain, such as species' nonlinear responses to environmental variables[2,22], unobserved/unknown driving forces[23], imperfect detection[16,24], among other outstanding difficulties[25].

With the development of modern statistics, technical advances provide powerful tools to estimate and predict species distributions, for example, machine learning methods and Bayesian hierarchical models are highly flexible to handle complex ecological responses and are promising for data-limited situations[26–28]. Some predictive methods have emerged to account for community information, leading to a new modelling approach known as community-level models[29] or joint species distribution models (JSDMs)[30–33].This modelling approach may benefit the prediction of rare species by borrowing strengths from community data[29,34–36], which include rich information of species correlations resulting from biological interactions or shared environmental gradients[30,37,38]. These factors have essential influences on species distributions thus may improve the predictive powers of species distribution models. That is, models that integrate community data may contribute to solving the 'rare-species modelling paradox'.

It should be acknowledged that this idea of community modelling is not quite new[39,40], and some studies have compared the performances between single- and multi-species models[41–43]. However, JSDMs remain underutilized to date[29,44,45], and there are limited understanding of their advantages and limitations. Although many studies suggest JSDMs may outperform single-SDMs (SSDMs), the advantage is not guaranteed[29], and JSDMs may lead to biased parameters if some species have responses to the environment very different from others. Therefore, the gains of adopting JSDMs need to be carefully considered.

This study tested the predictive performances of rare species distribution models, focusing on the hypothesis that community data may improve model prediction. We chose a range of SDMs that treat community data in different ways[29] and compared their performances using cross validation with survey data collected in the coastal water of Yellow sea, China. Both species occurrence and abundance were considered in the evaluation, as studies have concentrated on occurrence data but abundance data are better indicators of extinction risk[42,46]. In addition to comparing modelling algorithms, we evaluated two approaches of data arrangement, species filtering and conditional prediction, to enhance the predictive performances of the chosen models. These approaches were considered from a pragmatic viewpoint, i.e., available data and modelling techniques are often fixed and can be hardly improved in time, and improving model prediction, even to a limited extent, may be the only solution to account for the rare-species challenge. The goal of this study is to improve our ability to predict the spatial distribution of rare species for biological conservation.

Results

Variations in predictability.

The tested SSDMs and JSDMs had substantially different predictive abilities. Considering the results of Japanese seahorse (*Hippocampus mohnikei,* Sp4), AUCs (the area under curve of receiver operating characteristic) around 0.9 showed that occurrence of this species could be properly predicted by most models, except artificial neural network (ANN) (Fig. 1). The Cohen's κ coefficient indicated a similar pattern, whereas hierarchical modelling of species communities (HMSC) and generalized joint attribute model (GJAM) performed worse than those machine learning methods. The results of RMSE (root mean square error) were consistent with AUC, and ANN yielded RMSE larger than that simply assuming the absence of this species over survey areas (dash line). All the models had negative partial relative bias (PRB) on average, implying the tendency of underestimating abundance. The results of other five species showed a similar pattern but the values of performance metrics varied substantially (Supplementary Figure S5). In general, multivariate random forest (MRF) and random forest (RF) showed the best predictive powers for this species, followed by multi-response artificial neural network (MANN).

The divergences in the model performances were compared for other species. In terms of occurrences, MRF provided the best predictions of Sp1 (Brown croaker, *Miichthys miiuy*) and Sp3 (Blackhead seabream, *Acanthopagrus schlegelii*), and RF was optimal for Sp5 (Black scraper, *Erisphex pottii*). HMSC and MANN provided better predictions of Sp2 (Ocellate spot skate, *Raja porosa*) and Sp6 (Bartail flathead, *Platycephalus indicus*) in some measurements (Table 1). The cases of RMSE were complicated, i.e., HMSC and GJAM was the best for Sp1 and Sp2, respectively, RF best for Sp3 and Sp5, and MRF for Sp4 and Sp6. It should be noted that the discrepancies among models were relatively small in terms of the performance metrics, especially between RF and MRF. The predictions of abundance were poor for very rare species, and no model made better predictions than assuming all-zeros for Sp2. In addition, relative performances of the models were not consistent among species. Sp3, Sp4 and Sp6 were more readily predicted than the other species (Table 1). The occurrence of the rarest species in this study, Sp1, could be properly predicted, whereas Sp2 and Sp5 were less well predicted in terms of both occurrence and abundance.

Species filtering.

The increasing thresholds of species selection (filtering) led to less but strongly correlated species, which imposed different effects on the four JSDMs (Fig. 2). Among them, MRF tended to be less responsive to the changes of species selection, and the corresponding RMSE increased slightly only for Sp2 and Sp6 in LV3 (levels of species filtering, and LV3 denoted a small set of species selected). On the contrary, the predictions of MANN were substantially improved by reducing the number of species with decreasing RMSE, except for Sp6. HMSC was barely influenced in the cases of Sp1, Sp2 and Sp3 but benefited from specie selection for other species. GJAM also showed less responses to species selection for Sp1, 2, 3, but its performances decreased in terms of the other species. At LV3, MANN and HMSC tended to outperform the other models.

Conditional predictions.

Comparing to single-species RF, the predictive accuracy of conditional-RF (using ancillary species as predictive variables) was substantially improved for most species, indicated by the

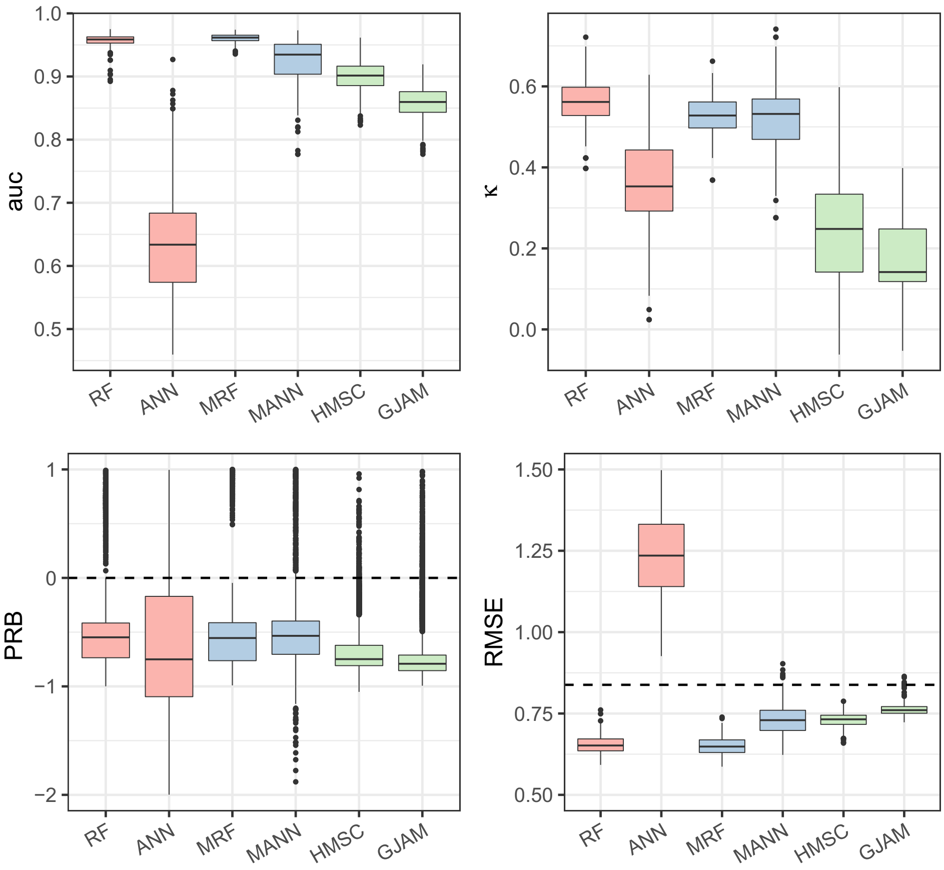

Figure 1. Predictive performances of models on the distribution of Japanese seahorse (*Hippocampus mohnikei*). The prediction of occurrence was evaluated by the area under the curve of receiver operating characteristic (auc) and Cohen's coefficient (κ), and prediction of abundance was evaluated by partial relative bias of non-zero data (PRB) and root mean square error (RMSE). The dash line in the last plot denotes a baseline of RMSE derived from all-zero predictions.

decreases in RMSE (ΔRMSE in Fig. 3). Predictions conditioning on observation data of ancillary species (RF-OBS) showed the most gains of accuracy; meanwhile, comparable improvement could be obtained with the help of JSDMs, i.e., conditional-RF based on JSDMs (using the outputs of JSDMs as predictors) could substantially improve RF, which performed better than MRF in many cases.

Conditional predictions also remarkably improved ANN to the performance similar to or better than MANN (Fig. 3). The degrees of improvement showed small differences between observation-based and model-based conditioning. However, the effects substantially differed among species, largest for Sp5 and Sp6 and least for Sp3.

Discussion

Given the global awareness of biodiversity loss with climate changes and anthropogenic pressures, it is not surprising that SDMs have been increasingly used in recent years. It is therefore of great concern how reliable the models are in their utility of predicting species distribution[47–49]. Here in this study, we examined the performances of a representative selection of modelling methods for rare species using a typical dataset available in marine fisheries surveys. Our results were generally mixed, that is, most species could be appropriately predicted in terms of occurrence, whereas non-zero abundance tended to be underestimated. Nevertheless, given the rather limited occurrence (mostly less than 10%), such performances were acceptable for rare species in a context of biological conservation. Although the conclusions may depend on specific objectives of studies and characteristics of targeted ecosystems, we highlight the opportunities of community data to address the 'rare-species modelling paradox'[30,35].

It is worth noting that this study covers a limited scope of SDMs in a continuous spectrum of complexity, and the potential of existing models may not be fully reflected. In particular, literature have concluded that the predictive abilities of SDMs may vary in different circumstances, depending on the type of organisms, their life-history trait, behavior, prevalence, data quality, spatial resolution and extent, and the impacts of human activities[17,25,50,51]. The target species in this study by no means represent the high diversity of marine organisms. In particular, the so-called 'rare species' may also diverge in definition, characterized by geographic range, habitat specificity and

Measures	Models	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
AUC	RF	0.875	0.644	0.949	0.959	0.800	0.911
	ANN	0.711	0.572	0.628	0.634	0.582	0.633
	MRF	0.893	0.618	0.956	0.962	0.784	0.926
	MANN	0.722	0.576	0.929	0.935	0.751	0.929
	HMSC	0.802	0.670	0.941	0.901	0.765	0.908
	GJAM	0.724	0.640	0.932	0.860	0.688	0.896
κ	RF	0.208	− 0.046	0.486	0.562	0.289	0.616
	ANN	0.134	0.030	0.358	0.353	0.102	0.320
	MRF	0.243	0.163	0.601	0.528	0.205	0.644
	MANN	0.088	0.030	0.486	0.532	0.243	0.634
	HMSC	0.041	− 0.034	0.493	0.248	0.126	0.541
	GJAM	0.046	− 0.034	0.408	0.142	0.069	0.497
RMSE	RF	0.299	0.419	0.377	0.652	0.569	0.588
	ANN	0.652	0.797	0.533	1.259	1.402	1.195
	MRF	0.300	0.416	0.414	0.648	0.577	0.559
	MANN	0.337	0.453	0.389	0.729	0.612	0.587
	HMSC	0.297	0.414	0.384	0.732	0.569	0.611
	GJAM	0.300	0.409	0.393	0.760	0.582	0.628
	Zero	0.300	0.397	0.418	0.838	0.601	0.776

Table 1. A summary of model predictive performances for target rare species. Each cell denotes the average values of the performance measures for a combination of species and models, respectively. Large values of AUC and κ represented high predictive accuracy of species occurrence and small values of RMSE represent high predictive accuracy of species abundance. The row of "Zero" denotes a baseline of RMSE when all predicted values are zeros.

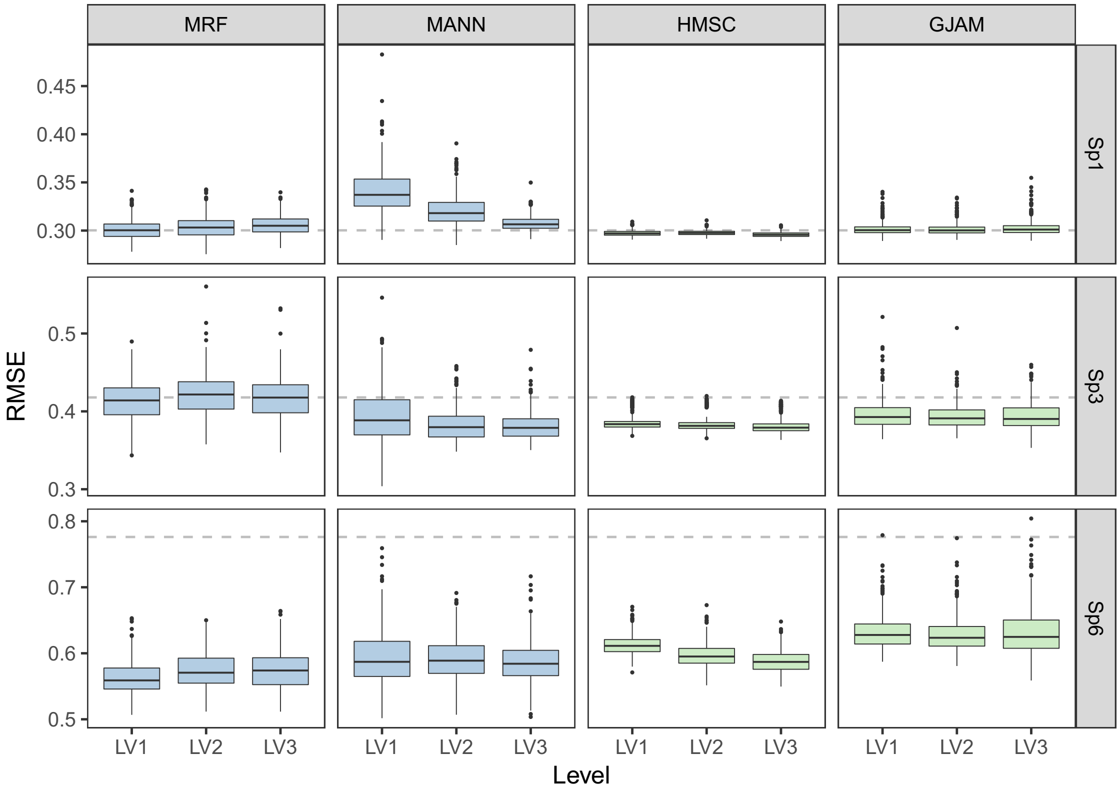

Figure 2. The influences of species filtering on the predictive performance of JSDMs. The levels in the X-axis denoted different thresholds of species correlation for selecting ancillary species (LV1 denoted a large set of species selected and LV3 denoted a small set. Three species are illustrated as examples and the full results are shown in Supplementary Information).

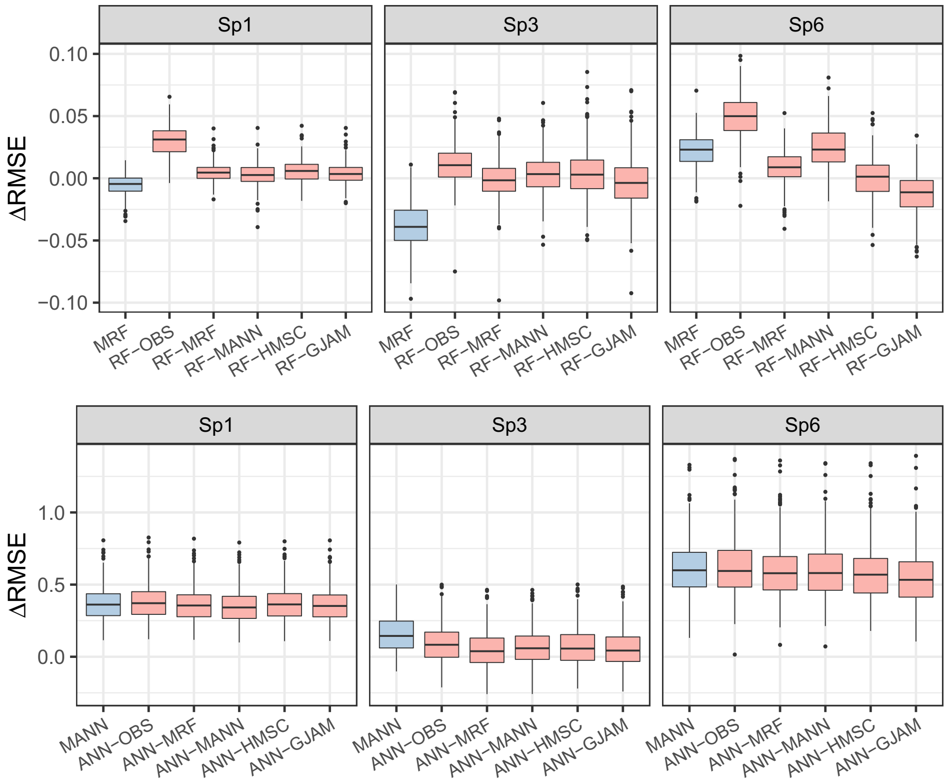

Figure 3. The effects of conditional prediction on improving predictive performances. The ΔRMSE indicates the decreases of RMSE in conditional models compared to that of single-species RF and ANN, respectively. RF-OBS and ANN-OBS denote the predictions conditioning on real observations (survey data), and others are conditioning on the prediction of JSDMs (Three species are illustrated as examples and the full results are shown in Supplementary Information).

local density, and different types of "rarity" may influence predictive models in different ways[16,52,53]. In general, substantial challenges still lie ahead on the road to predicting rare species.

In our evaluation, the six models had divergent performances when evaluated with different objectives, measures and target species. In general, the models using RF algorithms had better predictive ability than ANN- and regression-based models for both occurrence and abundance. The advantage could be largely attributed to the successful control of overfitting by model ensembles and internal cross-validation[54]. On the other hand, ANN easily led to overfitting under the circumstance of sampling errors and environmental noise[55]. Nevertheless, the predictive power was substantially improved in MANN and conditional ANN, implying that the overfitting issue was effectively alleviated by borrowing information from common species. On the other hand, the regression algorithm adopted by HMSC and GJAM implied that they were less flexible to non-linear relationships[30] and at the same time less vulnerable to overfitting[56]. Whereas, the regression-based JSDMs tended to be 'conservative" for rare species in terms of PRB. We highlight that model ensemble and internal cross-validation should be considered in the future development of SDMs, and particularly the capacity to account for non-linearity and overfitting for JSDMs[57].

Considering the overall performances of the SDMs, our evaluations generally find better predictive powers in the category of machine-learning JSDMs and conditional SSDMs, suggesting that community information are useful for the prediction of rare species[36], although the extent of improvement depends on the statistical algorithms adopted. It is well established that such gains could be attributed to the covariations in species distribution, as a result of (dis)similar environmental requirements, biotic interactions such as competition and predation, human impact such as fishing, and other stochastic processes such as observation/sampling errors[29,30,32]. Our results are consistent with this conclusion, i.e., species less correlated with the others (Sp2) tend to be poorly predicted while the well predicted one (Sp3, Sp4 and Sp6) show relative high correlations in the raw data (Supplementary Fig. S2). Meanwhile, it should be noted that SSDMs, specifically RF, may outperform the community models when predicting rare species, implying that community information are not helpful in certain circumstances. This is because the underlying driving forces may be idiosyncratic for the target species and others[29,58]. In this case, the distributional patterns of rare species reflected by the limited data may be concealed by the relatively large amount of data of common species, and increasing species number may make the situation worse for model fitting. Such a result was evident in the species selection processes in MANN and HMSC, both of which tended to have improved predictive powers when the number of species was reduced. On the other hand, MRF showed less responses to species selection because the RF algorithm could effectively suppress predictor species with loose correlations[54]. The declining performance of GJAM might also be attributed to the predicting algorithm, which generated latent variables randomly from a multivariate normal distribution according to species covariance matrices[59]. In this case, a strong correlation matrix might lead to larger prediction of latent variables and increased RMSE for rare species. Our results highlight the critical role of species selection in the implementation of JSDMs especially MANN and HMSC.

This study provides suggestions for the application of SDMs for rare species. First, MRF, conditional RF and HMSC are recommended provided the models properly tuned in structure and input variables. Conditional RF should be most powerful for modelling rare species when the distribution of common species are known in the locations of interest (RF-OBS). These results may contribute to extending the scope of species that can be statistically modelled and facilitating studies of similar backgrounds in the cases of rare species or limited data. In future studies, in addition to the improvement of data quality and quantity, algorithmic development is still in need to address the multiple issues raised by rarity. As no models is likely to be superior in all circumstances, diverse types of SSDMs and JSDMs with different features should be combined to address different situations of biological characteristics, rarity and available data, for which better understanding of potential and shortcoming of the existing models are required. Finally, regarding the challenges far from solved, we highlight the need of research efforts in the field of modelling rare species to deliver successful ecosystem management and biodiversity conservation.

Methods

Study area and data.

A marine fisheries survey was conducted in the north Yellow Sea, China to collect data. A modified systematic survey design was implemented with a total of 118 sampling stations in 2017 (Supporting information, Supplementary Fig. S1). In each station, an otter trawl which has the net width of 15 m and cod-end mesh size of 20 mm was towed for around 1 h at a speed of nearly 3 knots. Catch data were standardized to the same sampling efforts (trawling speed *time) for modelling. The survey and analysis methods were carried out in accordance with the ethics and guideline of the China law and the experimental protocol is approved by Ethical committee of Ocean University of China.

A total of 145 fish, shrimp and cephalopod species, in addition to benthos, were identified in the survey. As this study concentrated on rare species, only species occurring in less than 15% of the survey stations were selected as target species. As a result, six species with the occurrence frequency ranging from 3 to 12% were selected, including Brown croaker (*Miichthys miiuy,* Sp1, 3.5%), Ocellate spot skate (*Raja porosa,* Sp2, 4.3%), Blackhead seabream (*Acanthopagrus schlegelii,* Sp3, 6.1%), Japanese seahorse (*Hippocampus mohnikei,* Sp4, 8.8%), Black scraper (*Erisphex pottii,* Sp5, 9.6%), and Bartail flathead (*Platycephalus indicus,* Sp6, 12.3%) (Supplementary Table S1 in Supporting Information). In addition, 31 most prevalent species with occurrence frequency ranging from 23 to 87% were used as ancillary species (Supplementary Fig. S2) to help the prediction of target species. Commonly available hydrological variables in marine surveys were measured, including bottom water temperature, salinity, and depth (details are shown in Supplementary Table S2; Supplementary Fig. S3), using a CTD system (XR-420) in the same sampling stations after hauling.

Predictive models.

We selected six SDMs following three approaches in terms of how species associations are utilized. The first modelling approach is single-species distribution models (SSDM), which refer to the traditional methods that exclude community data. Two commonly used models, random forest (RF)[60] and artificial neural network (ANN)[61] are adopted. The two models are selected because they are powerful and can automatically deal with non-linear relationships that are prevalent in ecological studies[62,63]. The two models are used as references to evaluate how community information may improve the prediction of rare species distribution.

The second approach includes multivariate random forest (MRF) and multi-response artificial neural network (MANN), which are extensions of RF and ANN to account for multiple response variables, respectively. The former is analog to RF in term of bootstrap resampling but the split function is modified to minimize species compositional similarity within groups[64,65]. The latter MANN shares the same algorithm with ANN whereas its output layer has multiple neurons[66]. The connection coefficients between input and hidden layers affect all species collectively in MANN. Although both MRF and MANN are designed for modelling community data, their algorithms account for the information of species associations implicitly (c.f. the following category).

The third approach accounts for species associations explicitly, including two JSDMs that adopt the Bayesian hierarchical framework. The first is a versatile statistical framework of hierarchical modelling of species communities (HMSC)[32], which uses latent variables to incorporate information of species associations[32,67]. The other is generalized joint attribute model (GJAM), designed to accommodate multifarious data types flexibly, such as presence-absence, ordinal, continuous, discrete, composition and censored data[59,68]. The model represents species responses using a latent continuous variable, which can be censored to the discrete space of observations.

All the models were implemented on the R platform (version 3.5.1), using packages “randomForest”, “nnet”, “MultivariateRandomForest”, “HMSC”, and “gjam”, respectively. A summary of the models was provided in Table 2, and additional technical details were shown in Supporting Information.

Prediction improvement.

We tested two approaches to improving predictions of JSDMs and SSDMs, using species filtering and conditional prediction, respectively. It should be noted that the “improved” models used the same algorithms as above, whereas the variables used for model fitting varied. The first approach followed the concern that community models might not benefit predictions when the response variables were poorly correlated[29]. To avoid the undue influences, we selected ancillary species from the 31 common ones according to their correlations with target species. Three levels of species filtering were considered, level-1 (LV1) included all 31 common species, level-2 (LV2) included two-third species of the highest correlations, and level-3

Categories	Models	Full names	How to address species associations	R packages	References
SSDM	RF	Random forest	None	randomForest (v4.6-14)	Breiman[60]
	ANN	Artificial neural network	None	nnet (v7.3-12)	Basheer and Hajmeer[73]
Machine-learning JSDM	MRF	Multivariate random forest	Implicitly incorporated from compositional similarity	MultivariateRandomForest (v1.1.5)	Segal and Xiao[64]
	MANN	Multiresponse artificial neural network	Implicitly incorporated from neuron connections	nnet (v7.3-12)	Olden[66]
Regression-based JSDM	HMSC	Hierarchical Modelling of Species Communities	Explicitly incorporated with latent variables	HMSC (v2.2-0)[a]	Ovaskainen et al.[32]
	GJAM	Generalized Joint Attribute Model	Explicitly incorporated with a covariance matrix	gjam (v2.2.6)	Clark et al.[59]

Table 2. A summary of predictive models used in this study. [a]R codes of HMSC are available on Github (https://github.com/guiblanchet/HMSC), and others are available on CRAN.

(LV3) with the first third of the highest correlations. The process of species selection was conducted for each target species, and JSDMs were fitted with target species and their corresponding ancillary species at different levels of thresholds (LV), respectively.

The second approach, conditional prediction, was designed to improve the SSDMs using ancillary species directly as predictive variables[69]. The ancillary species were considered in two scenarios, one that ancillary species were observed in all sampling sites, and the other that they were predicted from JSDMs. Obtained from either way, the information of ancillary species were used in SSDMs as predictive variables. To suppress noise and reduce the number of predictive variables, principal component analyses (PCA) were conducted on ancillary species data prior to model fitting, and only PCs with eigenvalues above one were included in the conditional models[70].

Evaluation procedures.

A four-fold cross validation procedure was used to evaluate models' predictive performances. The total data were split into four equal sized subsamples, in which 75% were used for model training and the remaining 25% for testing, iteratively. To avoid potential failures with all-zero training/testing dataset, the nonzero data of rare species were randomly assigned to the four subsamples to ensure that each had equal number of occurrence of target species. Specifically, data splitting was conducted separately for samples with and without target species, and a permutation process was used to assign the survey data to four subsamples.

The predictive performances for species abundance were measured by root mean square error (RMSE) between observations and model predictions, RMSE $=\sqrt{\sum_i^N (P_i - O_i)^2/N}$, where P*i* and O*i* were the prediction and observation of abundance in sampling site *i*, respectively (RMSE thus has the same unit as abundance and the unit is omitted in the texts). In addition, we concerned the models' predictive power for non-zero observations and used partial relative bias (PRB) to measure predictive accuracy in the sampling sites where target species were present, i.e., PRB $=(P_p - O_p)/O_p$, where O_p was non-zero observations and P_p was the prediction in the corresponding sampling site.

Performances on predicting species occurrence were measured by the area under curve (AUC) of receiver operating characteristic and Cohen's κ coefficient[17]. The former has been commonly used for model evaluation of presence-absence, and the latter is used to indicate the chance-corrected agreement between predictions and observations[71]. A random guess of occurrence leads to 0.5 and zero in AUC and Cohen's κ, respectively. Additionally, True Skill Statistics[72] were calculated and shown in the Supporting Information. Given that low detectability of rare species might lead to zero observations, a species-specific threshold, mean abundance in the whole area, was used to determine species occurrence from predicted abundance. Data splitting, model fitting, prediction, and evaluation were conducted for each of the target species, and the processes of cross-validation were repeated 500 times.

Data availability

Data and R codes may be available from the Dryad Digital Repository.

Received: 19 February 2020; Accepted: 29 June 2020
Published online: 22 July 2020

References

1. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecol. Lett.* **16**, 1424–1435 (2013).
2. Elith, J. & Leathwick, J. R. Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697 (2009).
3. Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E. & Lundquist, C. J. A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Front. Mar. Sci.* **4**, 421 (2017).
4. Sofaer, H. R. *et al.* Development and delivery of species distribution models to inform decision-making. *Bioscience* **69**, 480–480 (2019).
5. Guisan, A. & Thuiller, W. Predicting species distribution: Offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009 (2005).

6. Hao, T., Elith, J., Guillera-Arroita, G. & Lahoz-Monfort, J. J. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Divers. Distrib.* https://doi.org/10.1111/DDI.12892 (2019).
7. Gogol-Prokurat, M. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecol. Appl.* **21**, 33–47 (2011).
8. Lomba, A. *et al.* Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. *Biol. Conserv.* **143**, 2647–2657 (2010).
9. Breiner, F. T., Guisan, A., Bergamini, A. & Nobis, M. P. Overcoming limitations of modelling rare species by using ensembles of small models. *Methods Ecol. Evol.* **6**, 1210–1218 (2015).
10. Magurran, A. E. & Henderson, P. A. Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**, 714–716 (2003).
11. Cao, Y., Larsen, D. P. & Thorne, R.S.-J.J. Rare species in multivariate analysis for bioassessment: Some considerations. *J. N. Am. Benthol. Soc.* **20**, 144–153 (2001).
12. Foden, W. B. *et al.* Climate change vulnerability assessment of species. *Wiley Interdiscip. Rev. Clim. Chang.* **10**, e551 (2019).
13. Guisan, A. *et al.* Using niche-based models to improve the sampling of rare species. *Conserv. Biol.* **20**, 501–511 (2006).
14. Ancillotto, L. *et al.* An African bat in Europe, *Plecotus gaisleri*: Biogeographic and ecological insights from molecular taxonomy and Species Distribution Models. *Ecol. Evol.* https://doi.org/10.1002/ece3.6317 (2020).
15. Della Rocca, F., Bogliani, G., Breiner, F. T. & Milanesi, P. Identifying hotspots for rare species under climate change scenarios: Improving saproxylic beetle conservation in Italy. *Biodivers. Conserv.* **28**, 433–449 (2019).
16. Cunningham, R. B. & Lindenmayer, D. B. Modeling count data of rare species: Some statistical issues. *Ecology* **86**, 1135–1142 (2005).
17. Vaughan, I. P. & Ormerod, S. J. The continuing challenges of testing species distribution models. *J. Appl. Ecol.* **42**, 720–730 (2005).
18. Franklin, J., Wejnert, K. E., Hathaway, S. A., Rochester, C. J. & Fisher, R. N. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Divers. Distrib.* **15**, 167–177 (2009).
19. Engler, R., Guisan, A. & Rechsteiner, L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *J. Appl. Ecol.* **41**, 263–274 (2004).
20. Chefaoui, R. M. & Lobo, J. M. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecol. Modell.* **210**, 478–486 (2008).
21. Phillips, S. J. *et al.* Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181–197 (2009).
22. Meynard, C. N. & Quinn, J. F. Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *J. Biogeogr.* **34**, 1455–1469 (2007).
23. Royle, J. A., Nichols, J. D. & Kéry, M. Modelling occurrence and abundance of species when detection is imperfect. *Oikos* **110**, 353–359 (2005).
24. Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecol. Modell.* **88**, 297–308 (1996).
25. Yates, K. L. *et al.* Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* **33**, 790–802 (2018).
26. Williams, J. N. *et al.* Using species distribution models to predict new occurrences for rare plants. *Divers. Distrib.* **15**, 565–576 (2009).
27. Rufener, M.-C., Kinas, P. G., Nóbrega, M. F. & Lins Oliveira, J. E. Bayesian spatial predictive models for data-poor fisheries. *Ecol. Modell.* **348**, 125–134 (2017).
28. Blangiardo, M. & Cameletti, M. Spatial and spatial-temporal bayesian models with R-INLA. *Spat Spat. Epidemiol.* **4**, 33–49 (2013).
29. Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W. & Fitzpatrick, M. C. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods Ecol. Evol.* **9**, 834–848 (2018).
30. Warton, D. I. *et al.* So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.* **30**, 766–779 (2015).
31. Thorson, J. T., Pinsky, M. L. & Ward, E. J. Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. *Methods Ecol. Evol.* https://doi.org/10.1111/2041-210X.12567 (2016).
32. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20**, 561–576 (2017).
33. Hui, F. K. C. Boral-Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods Ecol. Evol.* **7**, 744–750 (2016).
34. Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J. & Dunstan, P. K. Model-based thinking for community ecology. *Plant Ecol.* **216**, 669–682 (2015).
35. Ovaskainen, O. & Soininen, J. Making more out of sparse data: Hierarchical modeling of species communities. *Ecology* **92**, 289–295 (2011).
36. Hui, F. K. C., Warton, D. I., Foster, S. D. & Dunstan, P. K. To mix or not to mix: Comparing the predictive performance of mixture models vs separate species distribution models. *Ecology* **94**, 1913–1919 (2013).
37. Leach, K., Montgomery, W. I. & Reid, N. Modelling the influence of biotic factors on species distribution patterns. *Ecol. Modell.* **337**, 96–106 (2016).
38. Anderson, R. P. When and how should biotic interactions be considered in models of species niches and distributions?. *J. Biogeogr.* **44**, 8–17 (2017).
39. D'Amen, M., Rahbek, C., Zimmermann, N. E. & Guisan, A. Spatial predictions at the community level: From current approaches to future frameworks. *Biol. Rev.* **92**, 169–187 (2017).
40. Kindsvater, H. K. *et al.* Overcoming the data crisis in biodiversity conservation. *Trends Ecol. Evol.* **33**, 676–688 (2018).
41. Thorson, J. T., Kell, L. T., De Oliveira, J. A. A., Sampson, D. B. & Punt, A. E. Introduction to data-poor stock assessment. *Fish. Res.* **171**, 1–3 (2015).
42. Schliep, E. M. *et al.* Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. *Glob. Ecol. Biogeogr.* **27**, 142–155 (2018).
43. Maguire, K. C. *et al.* Controlled comparison of species- and community-level models across novel climates and communities. *Proc. R. Soc. B Biol. Sci.* **283**, 20152817 (2016).
44. Zhang, C., Chen, Y., Xu, B., Xue, Y. & Ren, Y. Comparing the prediction of joint species distribution models with respect to characteristics of sampling data. *Ecography (Cop.)* **41**, 1876–1887 (2018).
45. Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M. A. A comparison of joint species distribution models for presence–absence data. *Methods Ecol. Evol.* **10**, 198–211 (2019).
46. Ehrlén, J. & Morris, W. F. Predicting changes in the distribution and abundance of species under environmental change. *Ecol. Lett.* **18**, 303–314 (2015).
47. Smeraldo, S. *et al.* Modelling risks posed by wind turbines and power lines to soaring birds: The black stork (*Ciconia nigra*) in Italy as a case study. *Biodivers. Conserv.* **29**, 1959–1976 (2020).
48. Rizvanovic, M., Kennedy, J. D., Nogués-Bravo, D. & Marske, K. A. Persistence of genetic diversity and phylogeographic structure of three New Zealand forest beetles under climate change. *Divers. Distrib.* **25**, 142–153 (2019).
49. Guillera-Arroita, G. *et al.* Is my species distribution model fit for purpose? Matching data and models to applications. *Glob. Ecol. Biogeogr.* **24**, 276–292 (2015).

50. Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773–785 (2006).
51. Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. & Guisan, A. Measuring the relative effect of factors affecting species distribution model predictions. *Methods Ecol. Evol.* **5**, 947–955 (2014).
52. Rabinowitz, D., Cairns, S. & Dillon, T. Seven forms of rarity and their frequency in the flora of the British Isles. In *Conservation Biology: The Science of Scarcity and Diversity* 182–204 (Sinauer, 1986).
53. Gaston, K. J. *What is Rarity? In The Biology of Rarity: Causes and Consequences of Rare-Common Differences* 30–47 (Chapman and Hall, New York, 1997).
54. Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2**, 493–507 (2012).
55. Özesmi, S. L., Tan, C. O. & Özesmi, U. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Modell.* **195**, 83–93 (2006).
56. Norberg, A. *et al.* A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecol. Monogr.* **89**, e01370 (2019).
57. Harris, D. J. Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.* **6**, 465–473 (2015).
58. Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151 (2006).
59. Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J. & Zhang, S. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87**, 34–56 (2017).
60. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
61. Suryanarayana, I. *et al.* Neural networks in fisheries research. *Fish. Res.* **92**, 115–139 (2008).
62. Brun, P., Kiørboe, T., Licandro, P. & Payne, M. R. The predictive skill of species distribution models for plankton in a changing climate. *Glob. Chang. Biol.* **22**, 3170–3181 (2016).
63. Smoliński, S. & Radtke, K. Spatial prediction of demersal fish diversity in the Baltic Sea: Comparison of machine learning and regression-based techniques. *ICES J. Mar. Sci. J. Cons.* **74**, 102–111 (2017).
64. Segal, M. & Xiao, Y. Multivariate random forests. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**, 80–87 (2011).
65. Rahman, R., Otridge, J. & Pal, R. IntegratedMRF: Random forest-based framework for integrating prediction from different data types. *Bioinformatics* **33**, 1407–1410 (2017).
66. Olden, J. D. A species-specific approach to modeling biological communities and its potential for conservation. *Conserv. Biol.* **17**, 854–863 (2003).
67. Ovaskainen, O., Roy, D. B., Fox, R. & Anderson, B. J. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods Ecol. Evol.* **7**, 428–436 (2016).
68. Clark, J. S. Why species tell more about traits than traits about species: Predictive analysis. *Ecology* **97**, 1979–1993 (2016).
69. Araújo, M. B. & Luoto, M. The importance of biotic interactions for modelling species distributions under climate change. *Glob. Ecol. Biogeogr.* **16**, 743–753 (2007).
70. Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* **49**, 974–997 (2005).
71. Fielding, A. H. & Bell, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **24**, 38–49 (1997).
72. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223–1232 (2006).
73. Basheer, I. & Hajmeer, M. Artificial neural networks: Fundamentals, computing, design, and application. *J. Microbiol. Methods* **43**, 3–31 (2000).

Acknowledgements

The authors acknowledge all colleagues who have contributed to the marine surveys, data collection and analyses. Funding for this study was provided by the National Key R&D Program of China (2018YFD0900906) and the National Natural Science Foundation of China (31802301, 31772852).

Author contributions

C.Z. conceived the ideas, designed the study and wrote the first draft of the manuscript. Y.C. contributed ideas for the interpretation of the analyses and revised the manuscript. B.X., Y.X. and Y.R. collected the data and performed the analyses. All authors gave final approval for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69157-x.

Correspondence and requests for materials should be addressed to Y.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.