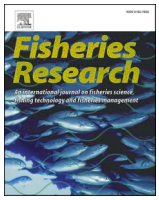
Fisheries Research 281 (2025) 107257

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Modeling the spatial distribution and abundance of deep-water red shrimps in the Mediterranean Sea: a machine learning approach

Elena Catucci [a,b,*], Diego Panzeri [c], Simone Libralato [c], Gianpiero Cossarini [c], Germana Garofalo [d], Irida Maina [e], Stefanos Kavadas [e], Federico Quattrocchi [d], Giulia Cipriano [f], Roberto Carlucci [f], Sergio Vitale [d], Chryssi Mytilineou [e], Fabio Fiorentino [d,g], Tommaso Russo [a,b]

[a] Laboratory of Experimental Ecology and Aquaculture, Department of Biology, University of Rome Tor Vergata, via della Ricerca Scientifica, Rome 00133, Italy
[b] National Inter-University Consortium for Marine Sciences (CoNISMa), Piazzale Flaminio, Rome, Italy
[c] National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy
[d] National Research Council (CNR) - Institute for Marine Biological Resources and Biotechnology (IRBIM), via L. Vaccara 61, Mazara del Vallo 91026, Italy
[e] Hellenic Centre for Marine Research (HCMR) - Institute of Marine Biological Resources and Inland Waters, 46.7 km Athens-Sounio Av., P.O. Box 712, Anavyssos, Attiki 19013, Greece
[f] Department of Biosciences, Biotechnologies and Environment – University of Bari Aldo Moro, Via Orabona, 4, Bari 70125, Italy
[g] Stazione Zoologica Anton Dohrn, Lungomare Cristoforo Colombo, Palermo 4521-90149, Italy

ARTICLE INFO

Keywords:
Modeling approaches
Random Forest
Niche overlap analysis
Spatial extrapolation
Fisheries management

ABSTRACT

Spatially-explicit models are invaluable tools for analyzing the species-environment interactions, even at scales beyond that of direct observations. In fisheries context, the observations on species usually consist of data derived from survey campaigns, such as the Mediterranean International Bottom Trawl Surveys (MEDITS) programme. MEDITS survey foresees the use of a standardized protocol for data acquisition on demersal species, such as the blue and red shrimp *Aristeus antennatus* and the giant red shrimp *Aristaeomorpha foliacea*. These two species are recognized as highly valuable marked resources accounting for about 5 % of the trawl fishing income in the Mediterranean basin. Here, we developed a modeling framework for the analysis of the MEDITS data on those species. Within our modeling framework we aimed at detecting the existence of a divergence in the spatial patterns that could guide the definition of targeted management actions for those two valuable fishing resources. A Random Forest (RF) machine learning approach has been used to model both the occurrence (i.e., presence/absence) and the biomass index (kg/km^2) of both species in four Geographical SubAreas (GSAs) located in the central part of the Mediterranean and the Ionian Sea. The RF showed high level of accuracy (i.e., K=0.83 and K=0.88, for *A. antennatus* and *A. foliacea*, respectively) in modeling species occurrence, and good level of performance (i.e., R^2=0.63 and R^2=0.74, respectively) in modeling their biomass index (kg/km^2). The niche overlap and statistical analyses we performed on the models outputs revealed the existence of a significant divergence in the spatial patterns between these species. This provides crucial ecological knowledge for the definition of targeted (i.e., species-related) management actions. Afterwards, the models have been extrapolated at the spatial scale of the Mediterranean Sea based on an approach we defined, called *hyperspace*. The *hyperspace* approach, while showing technical and ecological soundness, was meant to guarantee the reliability of model predictions in unknown areas. It reduces the need for a proper interpretation of "what is beyond a predicted value", offering a straightforward method for model extrapolation. Our effort aims to provide insights for prioritizing key areas in conservation strategies and marine spatial planning. It also represents an important contribution towards adopting an ecosystem-based approach to fishery resource management in the Mediterranean basin.

* Corresponding author at: Laboratory of Experimental Ecology and Aquaculture, Department of Biology, University of Rome Tor Vergata, via della Ricerca Scientifica, Rome 00133, Italy.
E-mail address: catucci.elena@gmail.com (E. Catucci).

https://doi.org/10.1016/j.fishres.2024.107257
Received 18 August 2023; Received in revised form 19 December 2024; Accepted 19 December 2024
Available online 8 January 2025

1. Introduction

A proper knowledge of the spatial distribution of species is of paramount importance not only for theoretical studies (Austin, 2002), but also for conservation and management planning (Elith et al., 2006; Guisan and Zimmermann, 2000; Hirzel and Le Lay, 2008). Modeling approaches are rooted in the quantification of the species-environment interactions, and they are critical for providing an ecological insight into the spatial patterns of a given species, to predict where is likely to occur, and to extrapolate such knowledge in unsampled locations (Pennino et al., 2019).

In the last decades, modeling approaches have become crucial for going beyond the available information, offering a way forward for the management of natural resources, especially in the context of fisheries (e.g., Aeberhard et al., 2018; Hazen et al., 2018; Keyl and Wolff, 2008). In the Mediterranean Sea the blue and red shrimp *Aristeus antennatus* (Risso, 1816) and the giant red shrimp *Aristaeomorpha foliacea* (Risso, 1827) rank among the most important commercial trawl resources exploited, accounting for about 5 % of the total fishing income in the whole basin (Podda et al., 2020). Since those species are distributed over the whole basin their analysis requires large-scale data which in turn demands massive and mostly multi-national efforts for their acquisition.

In this regard the Mediterranean International Bottom Trawl Survey (MEDITS) programme (Bertrand et al., 2002) results as an invaluable source of data. MEDITS is indeed a scientific survey carried out in the Mediterranean Sea since 1994 on a yearly basis, and is designed to acquire consistent biological data on the demersal biotic component through the use of a standardized method (Spedicato et al., 2019). The standardized protocol for gear and haul characteristics, sampling locations, file-formats of data, among other characteristics (Bertrand et al., 2002), has been defined jointly by the various partner countries. It was originally funded as a European Commission project for facilitating the cooperation among Member States of the European Union, yet over the years has been extended also to candidate countries for membership to European Union, e.g., Montenegro.

Since it is based on a common sampling procedure allowing to obtain consistent data, MEDITS survey provides a unique opportunity to increase our knowledge on distribution and abundance of Mediterranean species. Indeed, over the years MEDITS data have been widely used to analyze the spatial patterns of both *A. antennatus* and *A. foliacea*. For instance, Cau et al. (2002) analyzed abundance variability of both species in the Mediterranean basin in the 1994–1999 time range. They assessed the interannual variability in six 'Reference Areas' (RARs, hereafter), i.e., areas identified in the Mediterranean Sea considering the homogeneity of the environmental conditions (Cau et al., 2002). The authors pointed out a longitudinal gradient in which *A. antennatus* seemed to be more abundant in the RARs located in the western part of the Mediterranean Sea, while the opposite trend was found for *A. foliacea* which prevailed in RAR of the central and eastern part of the basin (Cau et al., 2002). Similarly, Guijarro et al. (2019) used the RAR-distinction of the Mediterranean basin to assess the spatial distribution of *A. antennatus* and *A. foliacea* and their sex ratio using MEDITS data acquired between the 1994 and 2015. Masnadi et al. (2018), using MEDITS data in 1994–2015, described the relationship between their biomass with the aim of estimating the dominance of one of the two species in the Tyrrhenian Sea. The authors claimed that *A. antennatus* showed a latitudinal gradient, being the dominant species in the northern part of the Tyrrhenian Sea, and it gradually decreased until a clear dominance of *A. foliacea* is reached in the Strait of Sicily (Masnadi et al., 2018).

MEDITS data were also used to conduct studies at local scale focusing on only one species. For instance, Orsi Relini et al. (2013) used data acquired in ten years of sampling (1994–2004) to analyze the population dynamics of *A. antennatus* in the Ligurian Sea. Podda et al. (2020), on the other hand, used MEDITS data on *A. foliacea* from 2009 to 2014 to determine the distribution of this species in the Sea of Sardinia in relation to environmental factors reflecting hydrographic conditions that can facilitate species movement, thus contributing to the peculiar spatial distributions.

Therefore, MEDITS data have been widely used in research studies focusing on both *A. antennatus* and *A. foliacea* showing different aims (e.g., assessing the dominance of one species over the other *vs.* assessing populations dynamics), different spatial scale of analysis (e.g., RAR in the Mediterranean *vs.* regional sub-basins), while considering or not considering the interannual variability in the data. One of the main features of the MEDITS survey that can be further exploited is its provision of data on species in standardized format, allowing to compare the species spatial patterns (distribution and biomass index) in trawl fishing areas. Recognizing the above, we aimed at detecting the existence of a divergence in the spatial patterns through the analysis of MEDITS data that could guide the definition of targeted management actions for those two valuable fishing resources. We proposed a unifying modeling framework for exploiting MEDITS data on both *A. antennatus* and *A. foliacea* that is based on the MEDITS's standardized format of data. With the term 'unifying' we meant a procedure that directly uses the MEDITS data in the exact format in which they are provided, using environmental predictive variables available at Mediterranean spatial scale for modeling development. Together these features allow to apply the modeling framework to MEDITS data as soon as they become available regardless of their location of acquisition. This approach, in a wider perspective, would benefit from data obtained by MEDITS over the years, and would value the efforts of such survey.

Within our modeling framework, we aimed at (i) modeling the spatial distribution of these two species in the central Mediterranean and Ionian Sea using the Random Forest (RF) (Breiman, 2001a) machine learning method, and at (ii) providing important spatial information for management measures through models analyses. Moreover, (iii) we proposed a novel methodological approach for extrapolating the models at the Mediterranean scale with the purpose of enhancing the reliability of modeling outcomes for the management of natural resources.

The RF was used for performing a classification task for modeling the occurrence (i.e., presence/absence) of the species, and a regression task for modeling the biomass index, i.e., kg/km^2 (*sensu* MEDITS). We selected predictive variables thoroughly considering our aim of proposing a unifying modeling framework. In particular, we used both environmental variables, defined as Essential Ocean Variables (EOVs) (Lindstrom et al., 2012), and geo-morphological factors, along with haul depth recorded during the sampling. The EOVs are variables that account for the three-dimensionality of the marine environment providing crucial information on the ecological characteristics that could affect *A. antennatus* and *A. foliacea* (Melo-Merino et al., 2020; Miloslavich et al., 2018; Robinson et al., 2011). They are available through the EU Copernicus Marine Environment and Monitoring Service (CMEMS) database that provides free access to comprehensive, up-to-date and historical data on ocean conditions, such as monthly and daily values of ocean salinity and of dissolved oxygen. While the geo-morphological variables are freely available for the entire Mediterranean basin as a result of a completed European project, called MEDISEH project (Giannoulaki et al., 2013), the use of CMEMS opensource database guarantees the harmonization of the modeling framework, regardless of the temporal and/or spatial coverage of data, strengthening the rationale behind this study. Following a proper validation of the developed models, a niche overlap analysis has been applied with the purpose of potentially improving our understanding of the spatial patterns of the two species, and in turn their management. Finally, our approach for model spatial extrapolation, while showing technical and ecological soundness, has been used to identify the sites across the basin that showed consistency with those upon which our models were built and validated. The latter allows to guarantee the reliability of models predictions, thus their usefulness for decision-making in the context of fisheries.

2. Materials and methods

2.1. Data on A. antennatus and A. foliacea

We used MEDITS data on *A. antennatus* and *A. foliacea* acquired in four Geographical SubAreas (GSAs, hereafter), i.e., GSA16, GSA18, GSA19 and GSA20, located in the central part of the Mediterranean basin and the Ionian Sea, from surveys conducted between 1999 and 2020.

For GSA20 only, and only for the years 1999–2001 – accounting for only about 3 % of the total – the trawl data were obtained from different surveys, i.e., INTERREG, RESHIO and Deep Fishery projects (Mytilineou et al., 2005; Papaconstantinou and Kapiris, 2001; Politou et al., 2001). These surveys were performed based on a protocol largely similar to that of the MEDITS, in terms of sampling and depth limit 10–800 m, while covering the entire Eastern Ionian Sea.

Overall, the dataset used in this study included 5961 sampling records, which about 34 % have been acquired in GSA16 and GSA18, while about 26 % in GSA19 and the remaining (~6 %) in GSA20 (Fig. 1).

2.2. Predictive variables

The 16 predictive variables (Table 1) consist of haul depth recorded during the sampling (1st in Table 1), nine EOVs (2nd to 10th in Table 1) downloaded from the CMEMS (https://marine.copernicus.eu/access-data), and six geo-morphological time-invariant variables (11th to 16th in Table 1) derived from the MEDISEH project (Giannoulaki et al., 2013).

Among the EOVs, chlorophyll *a* ('Chl') and net primary production ('Npp') provide information on the productivity of the water column that may help describing trophic habitat features that can be related to the two analyzed species (Cartes et al., 2014). While temperature ('Temp'), oxygen concentration ('O2') and salinity ('Sal') at the sea bottom are used as proxy for the environmental conditions that are known to influence distribution and biomass of both *A. antennatus* and *A. foliacea* (Carbonell et al., 2017). In addition, four vertical layers of particulate organic carbon (POC) in the water column were used, namely 'POC' in the deepest layer, 'POC 1' and 'POC 2' in the second and third deepest layers, i.e., at about 10 m and 100 m above the bottom, respectively, and finally 'POC mean' in the 0–200 m depth range. These variables provide information on possible food sources where the species feed (Kapiris and Thessalou-Legaki, 2011).

For the EOVs, monthly mean values corresponding to the period of the surveys were used, so that the data are associated with conditions actually reflecting the environment of the sampling period. On the other hand, the time-invariant variables (i.e., 'Bottype', 'Gebasp', 'Distriver', 'Gebslo', 'Distport' and 'Distcoast'; Table 1) provide information on the morphological aspects of the seafloor as well as other characteristics of the marine environment.

While the EOVs have a spatial resolution of 1/24 degree (about 4 km), the time-invariant variables show a resolution of ¼ arcminute (about 400 m). We resampled the time-invariant data to the coarser resolution by means of the function *'resample'* of *'raster'* package (version 3.6–26 – Hijmans, 2018) in the R environment (R Development Core Team, 2022) using the method "bilinear" for the continuous variables and the "nearest neighbor" for the categorical ones. Accordingly, our resulting models have a spatial resolution of 1/24 degree, about 4 km in coordinate system WGS84 (EPSG:4326).

The rationale behind such an approach – i.e., interpolating from finer to coarser resolution – is to maintain consistency within the predictive variables without introducing artificial information. Since the EOVs derived from models being validated in their native resolution (Cossarini et al., 2021; Escudier et al., 2021, 2020; Teruzzi et al., 2021), opting for the coarser resolution resulted in the most conservative choice.

2.3. Random Forest

RF is a Machine Learning algorithm that is widely used for both classification and regression tasks (Biau and Scornet, 2016; Boulesteix et al., 2012; Breiman, 2001a; Catucci and Scardi, 2020a, 2020b; Criminisi et al., 2012; Wager, 2016). It creates an ensemble of Classification and Regression Trees (CTs) (Breiman et al., 1984) and combines their predictions into a single model (Breiman, 2001). Each CT is constructed using a random subset of data (i.e., bootstrap sample) and a random subset of predictive variables for the splitting procedure. While for classification tasks the predictions are based on majority voting, for

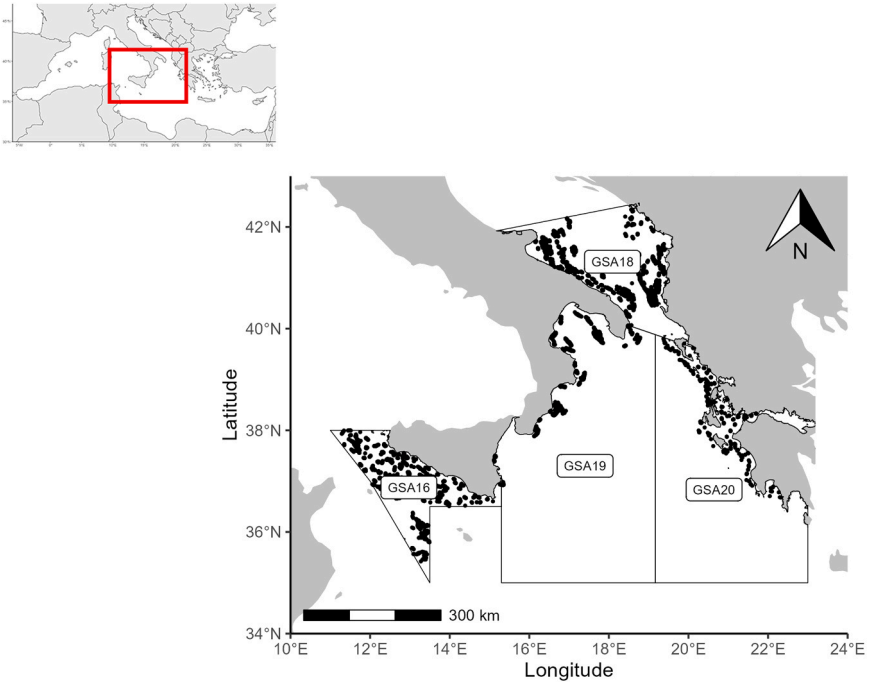


Fig. 1. Spatial distribution of trawl data on A. antennatus and A. foliacea gathered from 1999 to 2020.

Table 1
Predictive variables used for model development, along with specifications on their temporal coverage, temporal resolution, native spatial resolution and source. CMEMS: Copernicus Marine Environment and Monitoring Service.

	Variable	*Acronym*	*Unit*	*Temporal coverage*	*Temporal resolution*	*Native spatial resolution*	*Source*
1	Depth	Depth	m	Each haul	Each haul	Each haul	MEDITS
2	Bottom temperature	Temp	°C	1999–2020	monthly mean	1/24 degree	CMEMS
3	Bottom salinity	Sal	‰	1999–2020	monthly mean	1/24 degree	CMEMS
4	Dissolved oxygen at bottom	O2	mmol m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
5	Chlorophyll *a* in 0–200 m depth range	Chl	mg m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
6	Net primary production – integrated in 0–200 m depth	Npp	mg C m^{-3} d^{-1}	1999–2020	monthly mean	1/24 degree	CMEMS
7	Particulate organic carbon at the bottom	POC	mg C m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
8	Particulate organic carbon at the second to last depth layer	POC 1	mg C m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
9	Particulate organic carbon at the third to last depth layer	POC 2	mg C m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
10	Particulate organic carbon average in 0–200 m depth	POC mean	mg C m^{-3}	1999–2020	monthly mean	1/24 degree	CMEMS
11	Bottom type	Bottype	categorical variable	Time invariant		¼ arcminute	MEDISEH project
12	Seafloor aspect	Gebasp	categorical variable	Time invariant		¼ arcminute	MEDISEH project
13	Seafloor slope	Gebslo	%	Time invariant		¼ arcminute	MEDISEH project
14	Distance to river mouth	Distriver	km	Time invariant		¼ arcminute	MEDISEH project
15	Distance to harbors	Distport	km	Time invariant		¼ arcminute	MEDISEH project
16	Distance to coastline	Distcoast	km	Time invariant		¼ arcminute	MEDISEH project

regression applications the predictions are averaged for obtaining the final estimation.

The number of randomly selected predictive variables available at each split (*mtry)*, the minimum number of records contained in each leaf to stop the splitting procedure (*ndsize*), and the total number of CTs in the forest (*ntree*) are the main parameters that deeply affect the algorithm. For classification applications, *mtry* is usually set to $\sqrt{p}$ and to $p/3$ for regression, where p is the total number of available predictive variables, and tuned from half to twice its original value for model improvement (Cutler et al., 2012; Scornet, 2017). Both classification and regression tasks have been performed (see Section 2.4.2) and the RF has been tuned as follows:

- for classification models, *mtry* was tested in the 3–8 range, while the *ndsize* was tested in the 1–10 interval of values, setting *ntree* to 250;
- for regression models, *mtry* was tested in the 3–10 range, while the *ndsize* was tested in the 5–10 interval of values, setting *ntree* to 250;

The original RF algorithm implements two different Variable Importance Measures (VIM) for the assessment of the relative importance of the predictive variables, namely Gini and permutation measures. While the permutation measure is directly related to the prediction accuracy, the Gini one is strictly based on the splitting criterion (Boulesteix et al., 2012; Cutler et al., 2007; Gislason et al., 2006; Louppe, 2014). The modeling procedure was performed in the R environment (R Development Core Team, 2022) using the *'randomForest'* R package (Liaw and Wiener, 2002) which implements the original RF algorithm, i.e., the one provided by Breiman (2001).

2.4. Modeling framework

2.4.1. Data preprocessing

The MEDITS programme is aimed at collecting data on the entire community of demersal fish and shellfish rather than only on the two species considered in this study. As such, its sampling design may not provide a comprehensive representation of the entire spatial distribution of *A. antennatus* and *A. foliacea*. As a matter of fact, those species are known to occur more frequently at meso-bathyal depths, i.e. 500–800 m (Chan, 1998; D'Onghia et al., 2009; Komai et al., 2009; Sardà et al., 2004), with evidence of recruitment at even deeper depth especially for *A. antennatus* (Sardà and Cartes, 1997). Yet the majority (~70 %) of the MEDITS data were mostly collected in the upper part of continental slope (depth<500 m), meaning that the bulk of information refers to depths at which those species are hardly present. This may lead to a clear asymmetry in the distribution of the species occurrence in our dataset, especially regarding the absence records – the ones showing null biomass index (kg/km^2=0). Absence data indeed represented the majority, i.e., 75 % of the total data on average between the species, and they were mostly (~90 % on average) located in the upper part of continental slope (depth<500 m). Accordingly, the distribution of our input data resulted rather skewed and biased to allow the models to properly reconstruct and reproduce the potential actual distribution of the two species (e.g., Sardà and Company, 2012).

To overcome such limitation and to reduce the bias inherent of the MEDITS data, we incorporated pseudo-absence records into our dataset. Among the available methods, we adopted the background sampling approach (Elith et al., 2006) for generating pseudo-absence records. The latter foresees their selection based on ecological considerations, namely considering sites where the analyzed species is not expected to occur. The main ecological driver among those we considered that has the strongest and most established effect on the species spatial patterns is definitely depth (Massutí et al., 2008; Maynou, 2008; Rinelli et al., 2013). Depth is also the central parameter of the sampling protocol of the MEDITS survey, and in turn of the models that are developed using data coming from this programme. As such, we exploited the depth factor for selecting the pseudo-absence data since it guarantees that the rationale behind the selection of such type of data would consider the intrinsic nature of the species, of the MEDITS data, and of the models, integrating not only ecological considerations as the background approach requires (Barbet-Massin et al., 2012; Wisz and Guisan, 2009), but also methodological aspects. The selection of pseudo-absence data was made as follows.

Firstly, we obtained information on depth at Mediterranean scale from the CMEMS. The depth data show the same spatial resolution of

that of the EOVs (1/24 degree), accounting for more than 145,000 sites (i.e., raster pixels) over the whole basin ranging from about 5–5000 m. Then, within the four GSAs representing our study area, we selected the sites showing a depth value in the 1400–1500 m range, and we considered them as pseudo-absence data, as neither of the two species are expected to be present at this bathymetry. In total, we added 238 records to our dataset.

In a general perspective, these sites represent areas within which both *A. antennatus* and *A. foliacea* are hardily located, and they serve the purpose of rendering the distribution of the input data on species occurrence more symmetrical. Such integration could potentially enhance the overall quality of our modeling framework and related outcomes (e.g., Aarts et al., 2012; Stokland et al., 2011; Wisz and Guisan, 2009).

Finally, presence and absence records have been defined to perform the classification task as follows: if a given record showed non-null biomass index (i.e., kg/km^2>0), then that record was considered as 'presence' instance. On the contrary, all the records showing null biomass index were clearly identified as 'absence' instances. While for the regression models the data have been used in the format provided by the MEDITS survey, i.e., biomass index (kg/km^2). Indeed, MEDITS survey foresees that the data on catches (kg) are standardized based on the swept area (km^2) for providing information on species relative biomass index, i.e., kg/km^2.

2.4.2. Model development and validation

Following the pseudo-absence data inclusion, our dataset includes 6199 records (5961 plus 238), in which only about 17 % accounted for presence records of *A. antennatus* and about 28 % for the presence of *A. foliacea.* To properly develop and validate the models, the dataset has been partitioned into two subsets, called training and test sets. The subsets have been selected using the GFCM Statistical grid, available at https://www.fao.org/gfcm/data/maps/grid. For each cell of such a grid, 75 % of the data were randomly assigned to the training set while the 25 % to the test set. Such partitioning has been made also considering the necessity to maintain a similar observed proportion of species occurrence in both subsets. It is indeed paramount to have them as representative as possible of the modelled ecological process. Accordingly, in both training (n = 4610) and test (n = 1589) sets the portion of presence of both species was kept largely constant (i.e., ~17 % for *A. antennatus* and ~28 % for *A. foliacea*).

Furthermore, for dealing with the imbalance of the presence proportions showed by our dataset, the classification models were analyzed using the Receiver Operating Characteristic (ROC) curve before evaluating their accuracies by means of Cohen's Kappa (K, hereafter) (Cohen, 1960). The ROC curve, in addition to the value of the Area Under the Curve (AUC), provides the value of the optimal threshold (*t*) for discriminating between presence and absence instances, which is associated to the point on the curve at which the sum of sensitivity and specificity is maximized (Catucci and Scardi, 2022; Shatnawi, 2017). As the computation of K strictly depends on the true positive rate (TPR) and true negative rate (TNR), the selection of the optimal *t* represents a crucial step for obtaining models as unbiased as possible (Guisan et al., 2017). The K values range in the 0–1 interval, where 0 corresponds to a random agreement between predicted and observed data, while 1 indicates a perfect classification (Landis and Koch, 1977). While the ROC curve and optimal threshold analyses have been performed using the "*ROCR*" package (Sing et al., 2015), the K was computed using the "*psych*" package (Revelle, 2017).

On the other hand, the performance of regression models has been evaluated by computing the determination coefficient (R^2), which measures the proportion of target variance explained by the model, and the Root Mean Squared Error (RMSE). The models were optimized based on the K and R^2, respectively, so that the final models show the best predictive ability, i.e., their maximum values.

2.5. Niche overlap analysis

To provide an ecological insight into the results of our modeling framework for *A. antennatus* and *A. foliacea*, we performed a niche overlap analysis based on the Schoener index, called D (Schoener, 1974). D ranges in the 0–1 interval, where values close to 1 suggest significant niche overlap, while minimal overlapping for values that tend to 0. In a wider perspective, D quantifies the extent to which the species may interact, providing information about the similarity in their habitat requirements (Broennimann et al., 2012; Cardillo and L. Warren, 2016; Hurlbert, 1978). In other words, the niche overlap analysis offers a measure for estimating the degree to which the habitat requirements of the two analyzed species match (Dunham, 2013).

We performed such analysis on the RF predictions (i.e., model output, hereafter) of both the classification and the regression tasks. When applied to classification models outputs, D provides information on the degree of similarity of the habitat conditions that determine the occurrence of *A. antennatus* and *A. foliacea*. On the other hand, when applied to the regression models outputs, such similarity is quantified: the larger the D, the more similar are the conditions constraining the abundances of species in terms of biomass index (kg/km^2).

2.6. Extrapolation of models at Mediterranean spatial scale: hyperspace approach

Spatial extrapolation is the process of extending a model beyond the spatial scale of data on which it was originally developed. Its underlying assumption is that the relationships between the predictive variables and the response remain constant, regardless of space and scales. However, ecological conditions can vary widely, and any spatial extension of a model beyond the environmental domain over which it was built and validated involves uncertainty (Guillaumot et al., 2020; Merow et al., 2014; Owens et al., 2013).

In this study, we proposed an approach for model spatial extrapolation that is based on the concept of environmental domain, i.e., range of values (Barry and Elith, 2006). Such environmental domain is the multidimensional space (or hyper-space) consisting of the set of values of the predictive variables used for building the models (Barry and Elith, 2006). Consequently, we used the term '*hyperspace*' when referring to our approach for model spatial extrapolation.

The main idea of our *hyperspace* approach was to limit the necessity of a proper interpretation that is required when a model is extrapolated, favoring more ecological and methodological aspects. We indeed limited the spatial extrapolation of the models to only those sites across the basin that showed consistency with those upon which they were built and validated, as follows.

Firstly, we exploited the design of the MEDITS protocol for bathymetry for identifying the sites across the Mediterranean Sea that show a depth value consistent with that survey, i.e., within the 10–800 m range. Hence, based on the depth data for determining the pseudo-absence records (see Section 2.4.1), the number of sites (raster pixels) in the Mediterranean basin for spatially extrapolating the models has been reduced from about 145,000 to 52,372. This procedure allowed to maintain consistency between the input (MEDITS data in the four GSAs) and predicted (models extrapolation at Mediterranean scale) data.

We exploited the above-mentioned rationale also for all the other predictive variables in our dataset. For the quantitative ones (2nd to 10th and 13th to 16th in Table 1) we calculated their ranges of values, i.e., min-max. While for the qualitative ones (10th and 11th in Table 1) we considered the featured categories – i.e., the categories observed – in our dataset. This set of values (ranges plus categories) defined *de facto* the environmental domain of our data (*sensu* Barry and Elith, 2006), which we used for selecting the sites at which we extrapolated our models.

For each site in the Mediterranean basin already shown a depth value consistent with of the MEDITS survey (n = 52,372, see above), we

determined whether the relative environmental factors fall within the environmental domain. A site has been selected for model spatial extrapolation only if all the values of predictive variables ranged within the set of values of our data. On the contrary, even if only one predictive variable showed a value outside its range, that site was not considered for the model spatial extrapolation.

The last assumption for performing a technically sound spatial extrapolation was made on the temporal scale of the EOVs considering both the available data and the intrinsic nature of our models. In our dataset the extraction of EOV was performed to ensure alignment with the month and year corresponding to those in the MEDITS data, so that the MEDITS data acquired in a determined month and year are associated to EOVs reflecting the conditions of that month of that year (see Section 2.2). We observed that the largest amounts of records for both *A. antennatus* and *A. foliacea* have been acquired in the month of July for almost all the years. Indeed, about 40 % of the overall hauls have been carried out in July, while such percentage drops to about 8 % for the other months on average (Table 1 in supplementary material). Since July 2019 was the most representative sampling period over the whole dataset, we used the EOVs in such date. For the time-invariant predictive variables such consideration did not apply due to their stability over time by definition.

The selection of the most representative sampling period might help to reduce the impact of outliers that may be present in the MEDITS data compared to, for instance, the use of the most recent one, i.e., November 2020, during which the entire world was facing the COVID-19 pandemic that heavily affected also the fisheries (e.g., Russo et al., 2022), thus most likely leading to misinforming predictions (Pau et al., 2011). This time period might also result the most interesting from an ecological perspective. It is indeed the most recent available case for the July month, thus the more valuable for the natural resources management. Therefore, in order to maintain consistency also from a temporal perspective, we favor constraining the EOVs data to the most representative period of sampling, i.e., July 2019.

Moreover, the selection of a specific time period was also determined considering the modeling perspective. Indeed, due to the machine learning nature of the RF algorithm our models are totally data-driven (Breiman, 2001b). These type of models exploit the data for automatically detecting unidentified patterns, and use them for predicting new or unknown conditions (Murphy, 2012). In this view, July 2019 was the most representative "data" from which in all the likelihood the patterns have been detected by the RF.

3. Results

3.1. Modeling framework using Random Forest

The RF showed a high level of accuracy in modeling both the occurrence and the biomass index for both species (Table 2). The final RF configurations of the classification models of *A. antennatus* and *A. foliacea* were both based on 250 *ntree*, presenting 8 and 7 as *mtry*, and 6 and 3 as *ndsize* value, respectively. The final RF configurations of the regression models were based on 250 *ntree*, presenting 9 and 8 as *mtry* value for *A. antennatus* and *A. foliacea*, respectively, while the *ndsize* value was 7 for both species. The results on the RF training for both tasks and both species are shown in Table 2 in the supplementary material.

Table 2
Performance of the final models. K and AUC are measures for the classification models. R^2 and RMSE are measures for the regression models. t = optimal threshold value provided by the ROC analysis. The results refer to the test set data only.

	Classification model			Regression model	
	K	AUC	t	R^2	RMSE
A. antennatus	0.83	0.98	0.34	0.63	1.86
A. foliacea	0.88	0.98	0.48	0.74	1.04

For the classification models, K value resulted equal to 0.83 for *A. antennatus*, while that of *A. foliacea* to 0.88 (Table 2) indicating high level of models accuracy. The K values have been computed following the optimization of the RF outputs using only the data included in test set by means of the ROC analyses (Fig. 1 in supplementary materials). The ROC analyses provided an optimal *t* of 0.34 and AUC value of 0.98 for *A. antennatus*, and a $t = 0.48$ and AUC= 0.98 for *A. foliacea*, respectively (Table 2). Table 3 showed the confusion matrices of the classification models following the threshold optimization.

The RF showed a largely good level of performance also for the regression task (R^2=0.63 and R^2=0.77 for *A. antennatus* and for *A. foliacea,* respectively) (Table 2). The RF was able to explain more than 60 % of *A. antennatus* variance, and more than 75 % for *A. foliacea*. The regression models exhibited a rather small error (i.e., 2.10 kg/km^2 on average) which resulted quite negligible when compared to the inner variability of the biomass index in our dataset, ranging from 1 to about 150 kg/km^2 for *A. antennatus* and from 1 to ~300 kg/km^2 for *A. foliacea* (Fig. 2 in supplementary material).

Overall, the RF proved to be quite effective in reconstructing the spatial distribution and biomass index of the two species in the four GSAs (Fig. 2). The spatial patterns observed in the input data (Figs. 2a and 2b) were largely reflected in both the classification (Figs. 2c and 2d) and the regression (Figs. 2e and 2f) models. For instance, *A. antennatus* showed the largest portion of presence records (about 33 %) and largest value of biomass index (about 15.76 kg/km^2) in GSA19 representing the central part of our study area, while *A. foliacea* was predominant (about 33 % and 72.11 kg/km^2) in the easternmost part, i.e., GSA20. Those spatial patterns, although with some differences in terms of pure values, are well-recognized and revealed in the RF models outputs. In fact, *A. antennatus* showed the largest predicted values (9.54 kg/km^2) of biomass index in GSA19, where the habitat conditions resulted largely suitable (mean $p = 0.75$ – being p the probability of presence). While *A. foliacea* exhibited the largest predicted biomass index value (42.70 kg/km^2) in GSA20, the latter showing a high level of habitat suitability (mean $p = 0.79$).

3.2. Relative importance of predictive variables

The relative importance of the predictive variables was assessed using both Gini and permutation measures (Fig. 3). For all the models, depth was the predictive variable showing the largest relative importance, regardless of the measure used. The only exception to this is the result of the classification model for *A. foliacea* regarding the permutation measure (Fig. 3b – the plot on the right panel), in which 'POC 2' resulted the relative most important variable. Yet, even in the latter case, depth had a contribution as large as ~90 % on respect to that of the 'POC 2'.

These results underline the fundamental role of depth in the tree-building process during the RF training, most probably because of its crucial function in reconstructing the spatial patterns of the two species. It is worth noticing that those results reflect the high-order relationships assessed during the tree-building process, rather than merely causality (Fabbrizzi et al., 2020; Catucci and Scardi, 2020a; Louppe et al., 2013). Indeed, the RF exploits the spatial resemblance between the response and the predictive variables, but also that among the predictive variables themselves. For example, it stands to reason that the 'POC 2' shows a strong spatial resemblance to the depth since it reflects the gravitational settling of the organic carbon. In other words, not surprisingly depth showed the largest relative importance in our modeling framework, since this variable is the main ecological driver that encompasses the environmental factors at the surrounding, including the other predictive variables that we considered. Therefore, the relative importance of the predictive variables needs always be considered as an expression of multifaceted interactions that the RF handles that go beyond the linearity – a one-order function by definition (Breiman et al., 1984).

Table 3
Final classification models results for A. antennatus and for A. foliacea. The results refer to the data of the test set only.

A. antennatus				*A. foliacea*			
		observed				*observed*	
		absence	**presence**			**absence**	**presence**
predicted	**absence**	1250	18	*predicted*	**absence**	1098	44
	presence	71	250		**presence**	53	410

3.3. Niche overlap analysis

To detect the potential divergences in spatial patterns that might appear in the RF predictions in relation to the types of models (classification *vs.* regression) and species being modelled, we performed a niche overlap analysis. While the classification models outputs of the two species showed a partial niche overlap (mean D=0.68), the regression ones exhibited a rather limited niche overlap (mean D=0.40).

For providing an in-depth insight into the above, we further performed the niche overlap analysis at different depth strata (Table 4). In doing so, we considered also the pseudo-absence data, i.e., the ones at 1400–1500 m depth, for the sake of completeness. The classification models outputs of *A. antennatus* and *A. foliacea* the maximum degree of niche overlap (D=0.84) was at the depth stratum 700–800 m, while minimum degree (D=0.20) at the lowest one, i.e., 10–100 m (Table 4). With regards to the regression models outputs, the maximum level of niche overlap (D=0.71) was observed at intermediate depth, i.e. 200–300 m, and minimum level (D=0.19) at 600–700 m depth stratum (Table 4).

3.4. Models spatial extrapolation in Mediterranean Sea

In Table 5 we reported the ranges of values of the predictive variables that composed the environmental domain. Using those values, we were able to select 34,517 sites in the Mediterranean basin out of 52,372 – about 65 % (Fig. 4). The sites identified using our *hyperspace* approach mostly cover the central part of the Mediterranean Sea, being GSA13 and GSA15 the areas with largest portion of selected sites, i.e., about 85 % on average, while GSA11.1 the one with the smallest portion, i.e., ~3 % (Table 3 in the supplementary material). Not even a single site located in GSA21 has been selected for the model spatial extrapolation over about 7000 candidates (Table 3 in the supplementary materials).

Both the classification and the regression models have been extended at the spatial scale of the Mediterranean basin (Fig. 5) in only those sites selected by our *hyperspace* approach. In our results we observed that for both *A. antennatus* and *A. foliacea* the largest mean predicted values ($p = 0.60$ on average) of probability of presence was in GSA20, while smallest ones ($p = 0.03$ and $p = 0.06$) were found in GSA17 and GSA14, respectively. With regard to the regression models extrapolation, the largest mean predicted value of biomass index (7.83 kg/km^2) of *A. antennatus* was observed in GSA27, while the smallest one (1.29 kg/km^2) in GSA14. For *A. foliacea* the largest mean predicted value of biomass index (33.44 kg/km^2) was observed in GSA20 and the smallest one (1.45 kg/km^2) in GSA14.

4. Discussion

4.1. Unifying modeling framework and spatial analyses

This study proposed a modeling framework for exploiting MEDITS data to assess the spatial distribution and biomass index of *A. antennatus* and *A. foliacea*, two species that are known to rank among the most valuable trawl resources over the whole Mediterranean basin. By applying a RF machine learning approach, we achieved high accuracy in predicting species occurrence (K=0.83 for *A. antennatus* and K=0.88 for *A. foliacea*, respectively) and largely good performance in modeling their biomass index (R^2=0.63 and R^2=0.74, respectively). Hence, the RF resulted effective in modeling the spatial patterns of species, thus in detecting and in reproducing the spatial distribution and biomass index of both *A. antennatus* and *A. foliacea*.

Interestingly, despite being based on the same predictive variables, the RF provided differences in terms of purely predicted values. Those differences were related not only to the task (classification *vs.* regression), but also to the species being modelled. This suggests an intrinsic divergence between the spatial patterns of *A. antennatus* and of *A. foliacea* that the RF was able to detect and reproduce in its output.

To examine such divergence, we analyzed the distributions of the models outputs in relation to the depth – considering only the test set data. For both the performed tasks, the models outputs showed quite different values and distributions for the two species in relation to the depth strata (Fig. 6). Not surprisingly, the differences were mainly related to the deeper strata (depth>400 m), thus where *A. antennatus* and *A. foliacea* actually occurred. For the sake of completeness and consistency with the niche overlap analyses, also the pseudo-absence data are shown, i.e., the ones at 1400–1500 m refer to as depth stratum> 800 m (Fig. 6).

In particular, for the classification models outputs the Kolmogorov-Smirnov test pointed out a statistically difference (p-value<0.001, on average) between the distributions of the models outputs of the two species for the strata ranging from 200 m to 700 m depth (Fig. 6a). While the Mann-Whitney test suggested a statistically significant difference between their median values (p-value<0.001 on average – using Bonferroni correction) for all the strata from 200 m depth to 800 m, including the 700–800 m one (Fig. 6a).

Similar results were observed for the regression models outputs (Fig. 6b). The distributions of regression models outputs of the two species differ at all depths< 700 m (p-value<0.001 on average) according to the Kolmogorov-Smirnov test. For Mann-Whitney test statistically significant difference between their median values (p-value<0.001 on average – using Bonferroni correction) for all the strata from 200 m depth to 700 m, but not for the 700–800 m one (p-value=0.355) (Fig. 6b). None a statistically significant difference has been observed at depth> 800 m – accounting for the pseudo-absence data at 1400–1500 m – for both the performed tasks.

These results suggested that the species showed spatial distributions that are statistically different in relation to the depth. Indeed, *A. foliacea* is mostly located and more abundant at depth in the 500–600 m range, while *A. antennatus* is mostly located in the 700–800 m range. The latter is a well-known pattern distribution (e.g., D'Onghia et al., 2009; Rinelli et al., 2013; Sardà and Cartes, 1997), thus observations from literature confirm our results.

In addition to this, our results further suggested a general spatially-explicit pattern of the two species according to which *A. foliacea* tended to find its optimum – i.e., largest mean predicted values of both probability of presence and biomass index – at intermediate depth 500–700 m, showing a decrease around 800 m depth. While *A. antennatus,* on the other hand, showed a tendency to constantly increase with the depth, having its optimum at 800 m.

When combining these results with those provided by the niche overlap analyses the following can be drawn. The habitat conditions determining the suitability for the occurrence of *A. antennatus* and *A. foliacea*, expressed by the classification models outputs in terms of

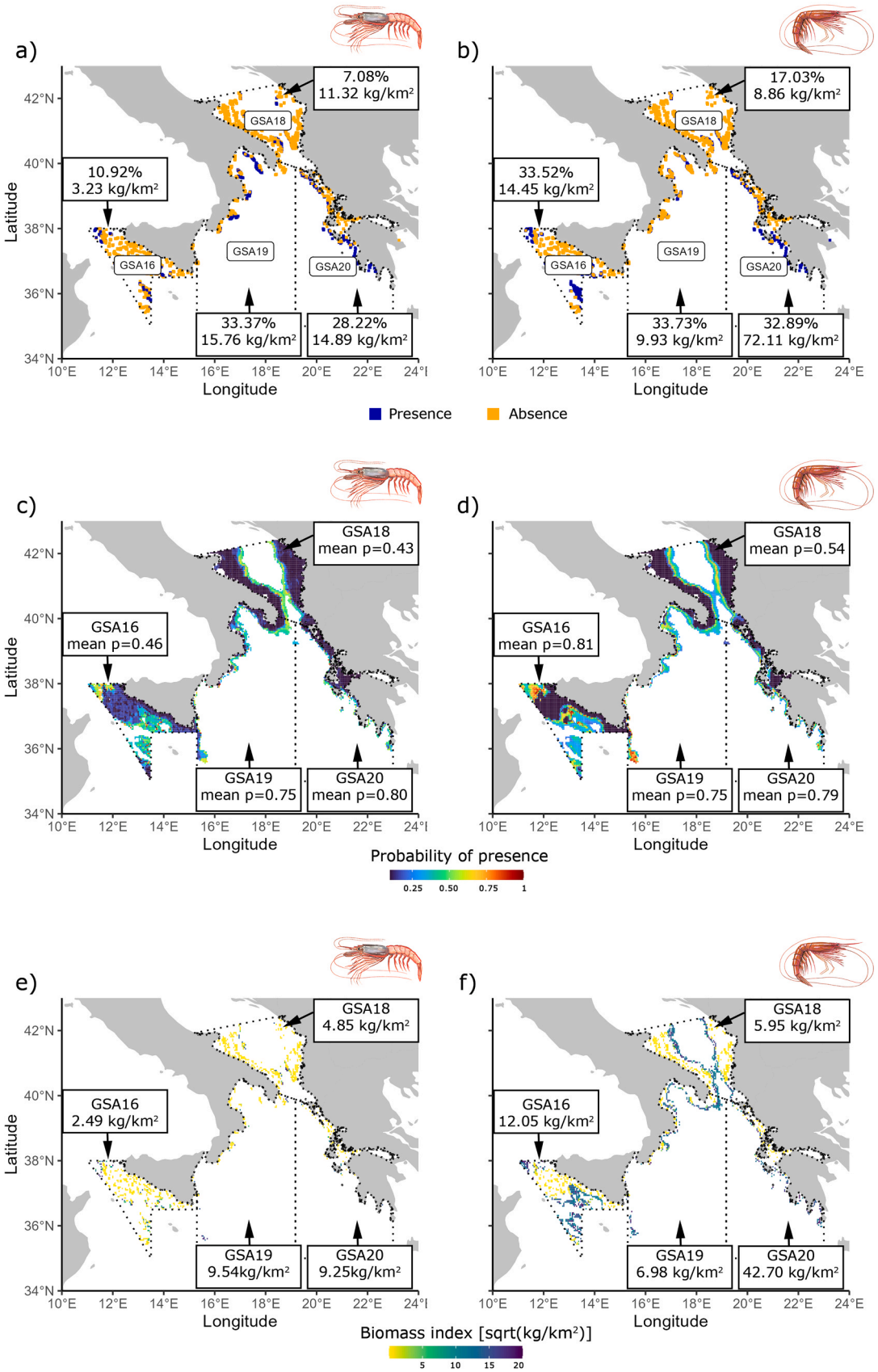


Fig. 2. Maps of spatial patterns of species. Input data of a) A. antennatus and b) A. foliacea. Classification model output of c) A. antennatus and d) A. foliacea. The rectangles report the mean predicted values of probability of presence (p). Regression model of e) A. antennatus and f) A. foliacea. The rectangles report the mean predicted value of biomass index (kg/km^2). For the regression models only the sites with biomass index$\geq$ 1 kg/km^2 are shown.

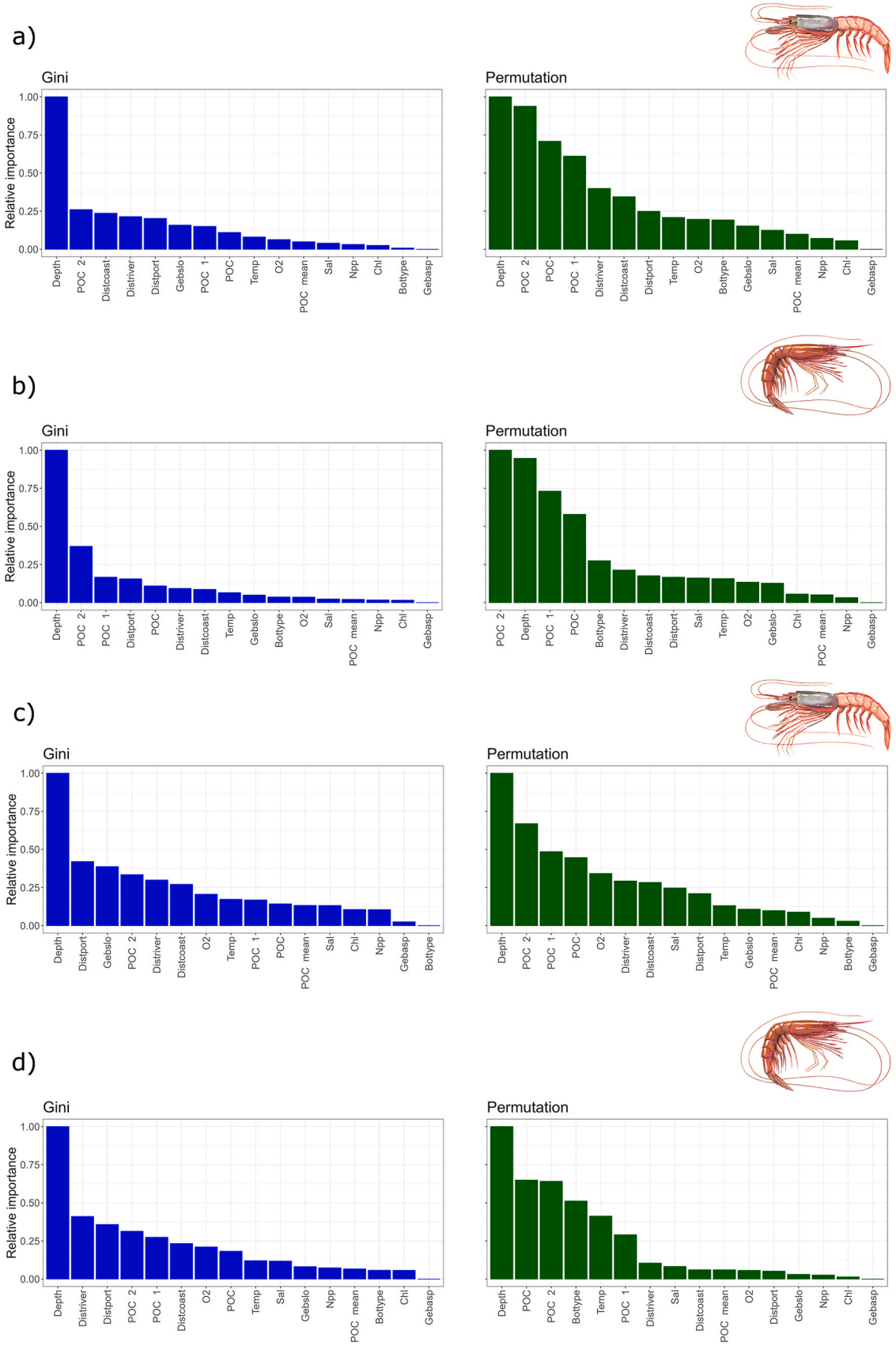


Fig. 3. Relative importance of predictive variables. The results are normalized based on the variable showing the largest relative importance. The subplots show the classification models for a) A. antennatus and b) A. foliacea, and the regression models for c) A. antennatus and d) A. foliacea.

Table 4
Niche overlap analysis of both classification and regression models in relation to different depth strata. D refers to the Schoener Index. The results refer to the test set data only, and for the sake of completeness also the pseudo-absence data (refer to as at depth>800 m) are analyzed.

Niche overlap analysis		
Depth (m)	*D index for classification models outputs*	*D index for regression models outputs*
10–100	0.20	0.22
100–200	0.63	0.39
200–300	0.48	0.71
300–400	0.70	0.42
400–500	0.73	0.37
500–600	0.53	0.59
600–700	0.60	0.19
700–800	0.84	0.56
> 800	0.20	0.22

probability of presence, largely matched in the 300–800 m interval of depth. On the other hand, the environmental factors influencing their abundance, expressed by the regression models outputs in terms of biomass index, differ to some extent between the two species especially at 300–500 m and in 600–700 m depth. This suggests that while for their occurrence *A. antennatus* and *A. foliacea* have rather similar habitat requirements, the conditions that actually defined their abundance diverge. In other words, the degree of similarity of the habitat conditions determining the suitability for species occurrence is larger than those constraining their biomass index.

Clearly, the assessment of which environmental variables contribute the most to the definition of such (more or less divergent) patterns may be interesting, yet it goes beyond the main objectives of the present study. It would indeed require an assessment of the cause-effect relationships between predictive variables and the two species, while the RF mainly assessed the relative importance of environmental factors based on non-linear interactions. Nonetheless, to the best of our knowledge and at the time writing, our work represents the first effort in such direction, and our outcomes could be useful to guide future research studies, and to design more targeted strategies for natural resources management (see Section 4.3).

Moreover, it is worth noting that precisely because establishing with absolute certainty which environmental factors are causally related to these species is still an open debate (Deval, 2019; Rinelli et al., 2013), the use of the RF in our modeling framework represented a rather remarkable advantage. It is largely known as the algorithm most robust to the inclusion of predictive variables that may be more or less relevant to the modeling (Louppe et al., 2013). Such feature allows us to explore the use of predictive variables showing different levels of relevance, such as EOVs and the geo-morphological ones, to the modeling of *A. antennatus* and *A. foliacea*, which resulted pivotal to achieve our aim of obtaining accurate and ecologically sound models. In a wider perspective, our overall approach can be considered an important step forward in providing a unifying modeling framework for the analysis of MEDITS data, that yields ecological insights for such valuable species and their management.

4.2. Hyperspace approach for model spatial extrapolation

Following our *hyperspace* approach, we were able to extrapolate the models in about 65 % of the sites, within the 10–800 m depth, across the Mediterranean Sea. Those sites were selected because the predictive variables showed values that ranged in the environmental domain, i.e., set of values and categories observed in the data.

As previously noted, the environmental domain represents the multidimensional space composed of the range of values of the environmental variables that are used as predictors in developing species

Table 5
Environmental domain of our data. The set of values of the continuous predictive variables are reported on the left, while the categories of the categorical variables on the right.

	Depth (m)	*Temp (°C)*	*Sal (‰)*	*O2 (mmol m^{-3})*	*Chl (mg m^{-3})*	*Npp (mg C m^{-3} d^{-1})*	*Poc (mg C m^{-3})*	*Poc_1 (mg C m^{-3})*	*Poc_2 (mg C m^{-3})*	*Poc_mean (mg C m^{-3})*	*Distriver (km)*	*Gebslo (%)*	*Distport (km)*	*Distcoast (km)*	*Bottype*	*Gebasp*
min	10.00	11.94	32.57	190.61	0.03	40.59	0	0	0	8.35	55.55	0	12.10	1.12	observed categories	1, 2, 3, 4, 5, 6	0, 90, 180, 270
max	800.00	27.81	39.19	270.68	0.61	925.75	89.38	89.04	89.23	87.09	844.43	24.84	366.66	239.99			

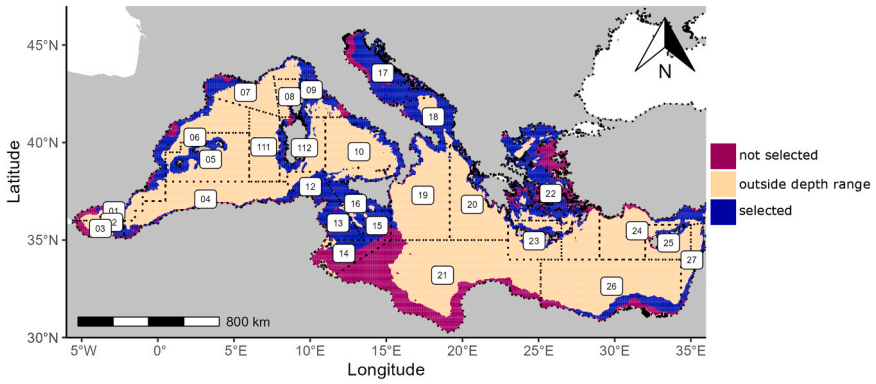


Fig. 4. Hyperspace approach for model spatial extrapolation. In blue, the sites selected for model spatial extrapolation. In magenta the sites not selected. In light orange the sites with a depth value outside that of the MEDITS. The numbers in the white squares refer to the GSAs. The map refers to July 2019.

distribution models. Hence, variable selection results critical in enhancing models predictive ability. Indeed, it should encompass all the environmental conditions affecting species occurrence in order to capture the complexity of species-environment interactions (Barry and Elith, 2006; Elith and Leathwick, 2009). Yet, the availability of (high--resolution) data, their temporal and spatial coverage, and the necessity to synthetize a natural process in modeling terms showing the best trade-off between accuracy and generality (Guisan et al., 2017), should always be considered in this selection. As such, models usually rely on a subset of environmental conditions that the modeler selects considering the final goal of its research study.

Here, the goal of developing a unifying modeling framework for MEDITS data analysis and of proposing a novel method, the *hyperspace* approach, to spatially extrapolate the models while guaranteeing the reliability of predictions – namely, limiting the subjective interpretability and favoring the methodological and ecological aspects – were the main drivers.

Accordingly, we ensured that our *hyperspace* approach would rely on: (i) ecological informative predictive variables, that would be available at Mediterranean spatial scale considering that MEDITS is carried out by European countries. This assures the possibility of applying it to all MEDITS data gathered across the Mediterranean basin, regardless of their spatial and/or temporal coverage; and (ii) considering the range of values of only the available input data for maintaining consistency between the properties of input and extrapolation sites. Since the models are trained and validated on determined data, i.e., input data, that refer to specific sites, when a given model is extrapolated in areas showing consistent conditions with such data, its response can be properly analyzed without any speculation.

Widening the data sources, i.e., integrate new MEDITS data as soon as they become available, is not only possible and straightforward in our modeling framework, but it would also enlarge the environmental domain which is the main ecological concept on which our *hyperspace* approach is built on. As the environmental domain expands, a larger number of sites may be selected for spatially extrapolating the models. The greater the number of selected sites, the wider is the area covered by our predictions that could guide the management more effectively because of their ecological basis.

From a management perspective, we recognize that extrapolating the models in 65 % of Mediterranean sites could limit, to some extent, the possibility of adopting a holistic approach for resources management for fisheries at the basin level. For instance, not even a single site among more than 7000 candidates of the GSA21 has been selected for the models spatial extrapolation. This outcome, rather than only depending on our *hyperspace* approach, highlights that our current level of knowledge in this geographical area is profoundly limited and it cannot be inferred from the current available data. As it stands, any kind of analysis or management strategy for fisheries in this GSA is not possible, let alone effective.

It is worth noting that GSA21 is managed by a not-EU, thus is not included in the EU Data Collection Framework (https://dcf.ec.europa.eu/index_en). This framework foresees the use of a systematic and common method for gathering and sharing data – such as the MEDITS programme – to ensure a holistic approach to natural resources management. Gathering data in this GSA, as well as in other areas that are not or poorly covered by the EU Data Collection Framework, could be needed to improve to improve our scientific knowledge on natural resources for their effective management in the context of fisheries.

In conclusion, it is worth stress that our modeling framework can be easily applied to other GSAs as soon as new MEDITS data become available, and even to different species for which the MEDITS provides standardized data, since we already pre-processed the predictive variables at Mediterranean spatial scale. This allows to overcome the necessity of building the models from scratch since our results are already validated – in this sense we defined our modeling framework unifying. Similarly, the *hyperspace* approach can be applied to a wide range of ecological applications, and it is not limited to the MEDITS data. It can be easily adapted to different ranges of data and to different tasks, while maintaining its main crucial feature, i.e., it guarantees the reliability of modeling outcomes because of the consistency between input and extrapolation sites. This applies especially to the data-driven models based on a machine learning algorithm, such as the ones we developed.

4.3. *Natural resources management based on a unifying modeling framework and hyperspace extrapolation: caveats and potentialities*

Spatially-explicit models are invaluable tools for the definition of a wide range of conservation and management actions (Foley et al., 2010) since their outcomes are largely straightforward, thus easy to understand and to communicate. They have substantially advanced the ecosystems and natural resources management (Catucci et al., 2022; Coll et al., 2020; Garofalo et al., 2011; Geary et al., 2020; Howell et al., 2021; Link et al., 2020; Panzeri et al., 2023, 2021; Townsend et al., 2019).

Modeling allows to analyze the species-environment interactions at scales beyond that of direct observations (Duarte et al., 2003). In fisheries context, usually but not exclusively, these direct observations refer to data derived from survey campaigns, such as MEDITS. Even though they typically reflect a rather specific population of the species, these data are pivotal to increase our ecological understanding on the demersal biotic component.

As MEDITS sampling is performed in in a rather specific temporal range, i.e., mostly during summer, the possibility of providing a holistic temporal analysis on the biology (such as intra-annual seasonal variation) of *A. antennatus* and *A. foliacea* is quite hindered. Clearly, we are not claiming that a temporal analysis based on the MEDITS data cannot be done, yet the inherent limits to the available information must be

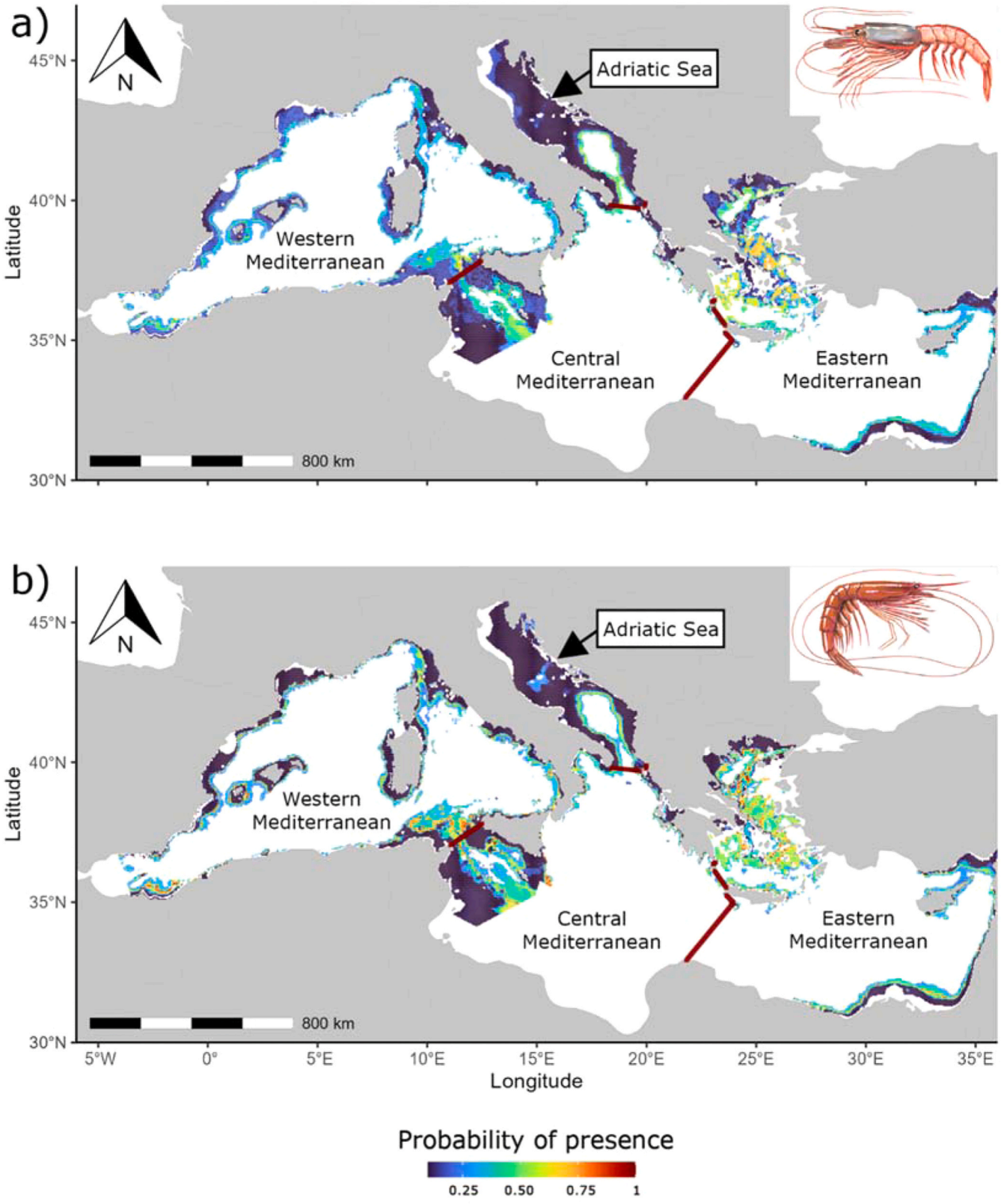


Fig. 5. Models spatial extrapolation at Mediterranean scale based on our hyperspace approach. Classification models extrapolation of a) A. antennatus and b) of A. foliacea. Regression models of c) A. antennatus and d) A. foliacea. Regarding the regression task, for the sake of clarity, only the RF predictions with a biomass index$\geq$ 1 kg/km^2 are shown. The magenta lines mark the borders among the four Mediterranean subregions defined by the GFCM 33/2009/2.

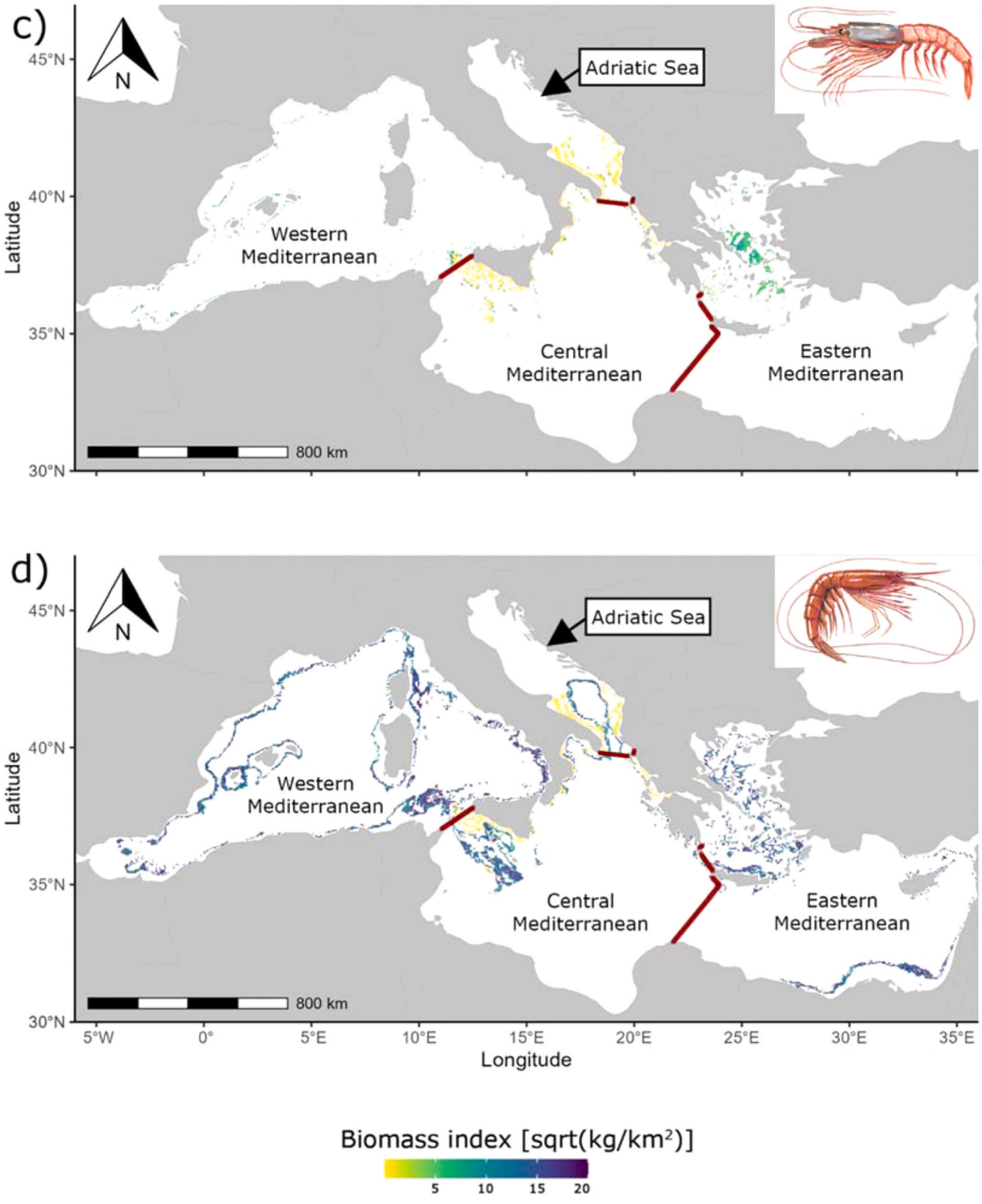


Fig. 5. (*continued*).

taken into account when the goal is to analyze them as a time series (e.g., Guijarro et al., 2019). Considering such data features, in this study we meant to reconstruct the spatial patterns of *A. antennatus* and *A. foliacea* without considering the difference that might appear in temporal scale. Nonetheless, we believe that the proposed modeling framework strengthened the usefulness of MEDITS programme by enabling a synthesis on the data, providing critical information that could guide more effective management actions.

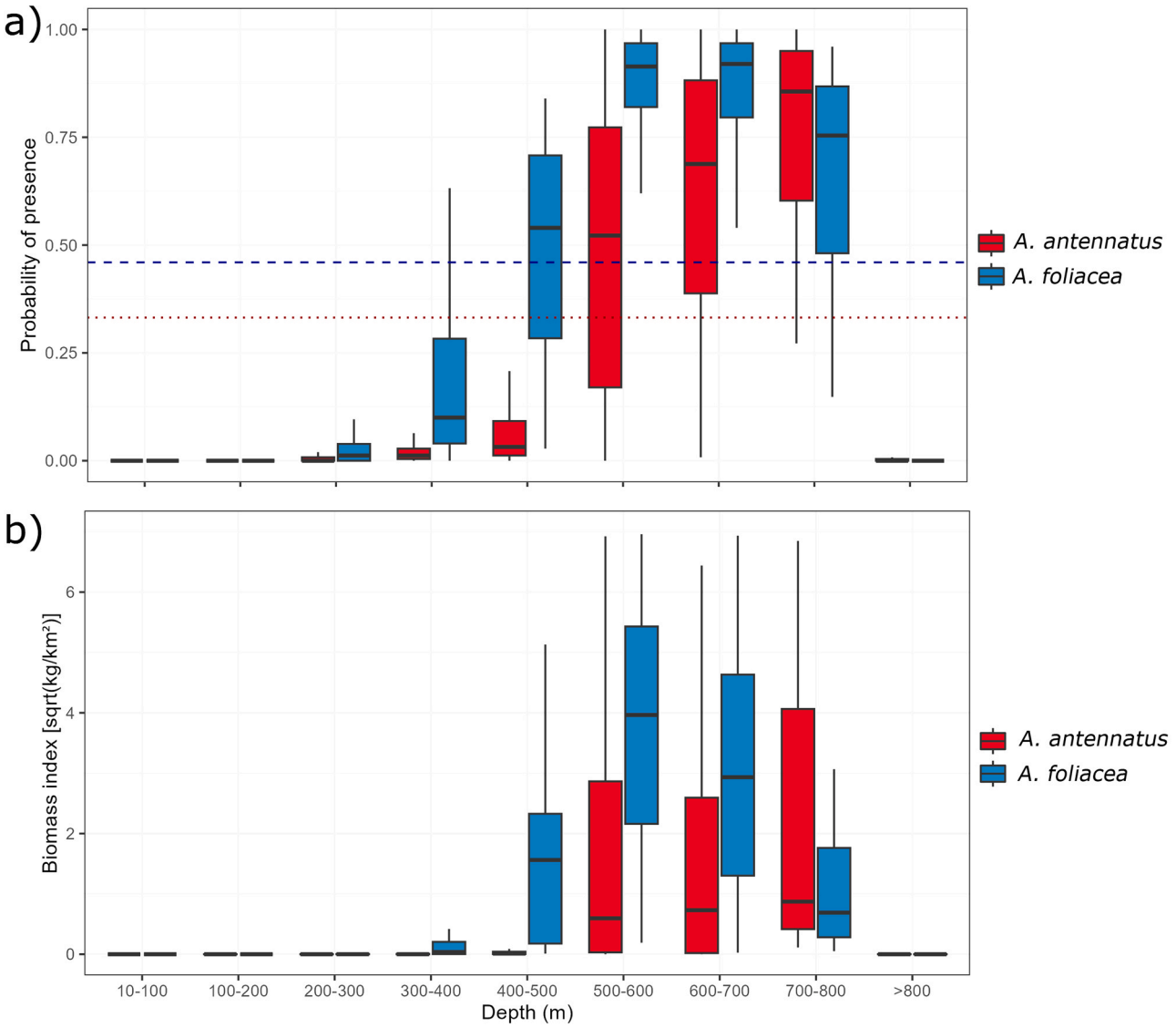


Fig. 6. Distribution of a) the classification and of b) the regression model outputs in relation to depth. The boxplots refer to the test set data only. In a) the dotted red line represents the optimal threshold (t = 0.34) for A. antennatus, while the dashed blue one (t = 0.48) refers to A. foliacea. For the sake of completeness also the pseudo-absence data identified at 1400–1500 m are shown in the plot refer to as the stratum depth> 800 m.

Indeed, the analyses performed on the models outputs of the four GSAs revealed that *A. antennatus* seemed to find its optimum – i.e., largest mean predicted values of probability of presence and of biomass index – at the deepest MEDITS depth, i.e., around 800 m. While *A. foliacea* showed an increase of values (in terms of both probability of presence and biomass index) from 10 to 500, being the 500–700 m the depth strata for the species optimum, followed by a decrease in those values in the 700–800 m range of depth. Our results further highlighted that in the 300–500 m range of depth *A. foliacea* is statistically more frequent and more abundant than *A. antennatus*. The niche overlap analyses also revealed that the degree of similarity of the habitat requirements is rather high for the occurrence of both *A. antennatus* and *A. foliacea* (mean D=0.54). Whereas the similarity is rather limited when it accounts for their biomass index (mean D=0.40). Particularly in the 300–500 m range of depth those species showed the lowest degree of similarity of habitat requirements constraining their abundance (kg/km^2), i.e., mean D= 0.38.

Accordingly, the 300–500 m depth range seemed the strata more suitable for defining management actions targeted to *A. foliacea*. While deeper depth, i.e., around 800 m, seemed to be more suitable for management strategies specifically targeted to *A. antennatus*, being the latter statistically more frequent and more abundant than *A. foliacea*.

Precisely to enhance the usefulness of our outcomes in the context of fisheries management and governance, we analyzed the distribution of

Table 6
Mean predicted values of both classification and regression models for A. antennatus and A. foliacea in relation to the Mediterranean subregions identified by the GFCM over the whole basin. The values of the classification models refer to the mean probability of presence while that of the regression models to the mean biomass index, i.e. kg/km^2.

GFCM subregion	*Classification Models*		*Regression Models*	
	A. antennatus	*A. foliacea*	*A. antennatus*	*A. foliacea*
Western Mediterranean	0.37	0.30	4.08	3.12
Central Mediterranean	0.49	0.42	3.37	7.90
Adriatic Sea	0.25	0.22	2.43	4.53
Eastern Mediterranean	0.29	0.58	2.39	6.13

the extrapolated predictions in relation to the GFCM distinction of the basin (GFCM/33/2009/2, https://www.fao.org/gfcm/data/maps/gsas/en/), according to which four (macro) subregions are recognized across the whole basin, i.e., 'Western Mediterranean', 'Central Mediterranean', 'Eastern Mediterranean' and 'Adriatic Sea' subregion (Table 6). Both the species found quite suitable habitat conditions ($p = 0.46$ on average) in the 'Central Mediterranean' subregion, while rather limited suitability is shown by the 'Adriatic Sea' subregion ($p = 0.23$ on average). With regards to their biomass indices, *A. antennatus* showed the largest mean predicted value (4.08 kg/km^2) in the 'Western Mediterranean' subregion and smallest mean predicted value (2.39 kg/km^2) in the 'Eastern Mediterranean' subregion. A largely opposite trend was observed for *A. foliacea*, for which large mean predicted value of biomass index (6.13 kg/km^2) has been observed in the 'Eastern Mediterranean' subregion – with slightly larger value (7.90 kg/km^2) in the 'Central Mediterranean' one – and smallest mean value (3.12 kg/km^2) in the 'Western Mediterranean' subregion.

These results largely reflected the well-known longitudinal gradient (Cau et al., 2002; Guijarro et al., 2019) exhibited from these species, highlighting a consistency between our outcomes and observations from literature. Considering the narrow spatial distribution of our input data compared to that of the research studies above-mentioned, such consistency can be deemed rather remarkable. Mostly, it might highlight the robustness of the technical-ecological soundness of the *hyperspace* approach for model spatial extrapolation, and in turn its potential.

To this regard, it is worth noting that the four GFCM's subregions are usually used for statistical purposes in management contexts. They substantially differ from the RAR-distinction used in previous research studies (e.g. Cau et al., 2002) dealing with the two analyzed species. The RAR distinction is based on the concept of ecological homogeneity, that is clearly important, but it does not have political and/or management acknowledged meaning.

Our modeling framework and related outcomes have the potential to be applied for mapping the Essential Fish Habitat including critical spawning and nursery areas, and to meet the requirements outlined in GFCM Recommendations (GFCM/42/2018/3 and GFCM/43/2019/6), which have led to the establishment of multiannual management plans to ensure sustainable trawl fisheries targeting for *A. antennatus* and *A.* foliacea, in the Ionian Sea (GSA19, GSA20, and GSA21) and the Strait of Sicily (GSA12, GSA13, GSA14, GSA15, and GSA16), respectively. These recommendations have led to the designation of spatial/temporal restrictions, such as Fishery Restricted Areas (FRAs), that advocate for various measures. FRAs are defined as geographical areas in which some specific fishing activities are temporarily or permanently restricted in order to improve the exploitation patterns of the targeted species and promote fishery sustainability.

In a wider perspective our outcomes can inform fisheries management by identifying areas of high suitability and abundance, aiding in the implementation of effective measures. The latter would precisely align with the current necessity of implementing a holistic approach for the definition of management actions. It could indeed provide valuable information for assessing the status and the

exploitation of the species, when integrated with the data on fishery stocks. By leveraging these combined sources of information, stakeholders can make well-informed decisions to ensure the long-term conservation and responsible exploitation of precious natural resources, such as *A. antennatus* and *A. foliacea*.

5. Conclusions

In this study, a unifying modeling framework for the analysis of MEDITS data on *A. antennatus* and *A. foliacea* has been proposed based on a RF machine learning approach. Those species are widely recognized as commercially important demersal resources over the Mediterranean Sea.

The RF proved to be quite effective for modeling the occurrence and the biomass index for both the species. Interestingly, despite being based on the exact same predictive variables, the RF statistically provided different predicted values in relation to the species being modelled, suggesting an intrinsic divergence between them. Such divergence has been also assessed using the niche overlap analyses through the Schoener Index. Based on our results, the possibility of drawing more targeted management and conservation actions for both *A. antennatus* and *A. foliacea* based on MEDITS data has been revealed. This framework offers a valuable tool for the development of conservation planning and management strategies.

We further defined a novel approach for spatially extrapolating our models at Mediterranean scale. Using our *hyperspace* approach, we were able to identify the sites across the basin that resulted consistent with those upon which our models have been trained and validated. This substantially reduces the necessity of a proper interpretation of "what is beyond a predicted value", offering a more straightforward method for spatially-extending a model while enhancing its reliability, and in turn its usefulness for management. Our outcomes highlight the need to value these modeling approaches in areas not or poorly covered by the EU Data Collection Framework to achieve the ecosystem-based approach for natural resources management.

The lack of a common approach for data sharing and for data analysis could always limit the feasibility of adopting a holistic and effective management approach. This study could represent an important effort in such direction, providing a modeling framework for the analysis of MEDITS data and a technically sound approach for models spatial extrapolation. It represents a step forward for a more comprehensive ecological understanding and effective assessment and management of fisheries resources.

CRediT authorship contribution statement

Fabio Fiorentino: Data curation, Writing – review & editing. **Diego Panzeri:** Conceptualization, Data curation, Writing – original draft. **Chryssi Mytilineou5:** Data curation, Writing – review & editing. **Roberto Carlucci:** Data curation, Writing – review & editing. **Elena Catucci:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Giulia Cipriano:** Data curation, Writing – review & editing. **Federico Quattrocchi:** Data curation, Writing – review & editing. **Stefanos Kavadas:** Supervision, Writing – review & editing. **Irida Maina:** Data curation, Formal analysis, Writing – review & editing. **Germana Garofalo:** Data curation, Writing – original draft, Writing – review & editing. **Gianpiero Cossarini:** Data curation, Software. **Tommaso Russo:** Conceptualization, Writing – original draft, Writing – review & editing. **Simone Libralato:** Conceptualization, Writing – original draft, Writing – review & editing. **Sergio Vitale:** Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We wish to sincerely thank Professor Michele Scardi for his precious feedback. His expertise and insights are always invaluable.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2024.107257.

Data Availability

Data will be made available on request.

References

Aarts, G., Fieberg, J., Matthiopoulos, J., 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. Methods Ecol. Evol. 3, 177–187. https://doi.org/10.1111/j.2041-210X.2011.00141.x.

Aeberhard, W.H., Mills Flemming, J., Nielsen, A., 2018. Review of state-space models for fisheries science. Annu. Rev. Stat. Its Appl. 5, 215–235. https://doi.org/10.1146/annurev-statistics-031017-100427.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118. https://doi.org/10.1016/S0304-3800(02)00205-3.

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423. https://doi.org/10.1111/j.1365-2664.2006.01136.x.

Bertrand, J.A., de Sola, L.G., Papaconstantinou, C., Relini, G., Souplet, A., 2002. The general specifications of the MEDITS surveys. Sci. Mar. 66, 9–17.

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

Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., 2012. 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.

Breiman, L., 2001a. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2001b. Statistical modeling: the two cultures. Two Cultures 33.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Group 37, 237–251.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.-J., Randin, C., Zimmermann, N.E., Graham, C. H., Guisan, A., 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. Glob. Ecol. Biogeogr. 21, 481–497. https://doi.org/10.1111/j.1466-8238.2011.00698.x.

Carbonell, A., Llompart, P.J., Gaza, M., Mir, A., Aparicio-González, A., Álvarez-Barastegui, D., Balbín, R., Cartes, J.E., 2017. Long-term climatic influences on the physiological condition of the red shrimp Aristeus antennatus in the Western Mediterranean Sea. Clim. Res. 72, 111–127.

Cardillo, M., L. Warren, D., 2016. Analysing patterns of spatial and niche overlap among species at multiple resolutions. Glob. Ecol. Biogeogr. 25, 951–963. https://doi.org/10.1111/geb.12455.

Cartes, J.E., Fanelli, E., Kapiris, K., Bayhan, Y.K., Ligas, A., López-Pérez, C., Murenu, M., Papiol, V., Rumolo, P., Scarcella, G., 2014. Spatial variability in the trophic ecology and biology of the deep-sea shrimp Aristaeomorpha foliacea in the Mediterranean Sea. Deep Sea Res. Part I: Oceanogr. Res. Pap. 87, 1–13. https://doi.org/10.1016/j.dsr.2014.01.006.

Catucci, E., Buonocore, E., Franzese, P.P., Scardi, M., 2022. Assessing the natural capital value of Posidonia oceanica meadows in the Italian seas by integrating Habitat Suitability and Environmental Accounting Models. ICES J. Mar. Sci. 80, 739–750. https://doi.org/10.1093/icesjms/fsac034.

Catucci, E., Scardi, M., 2020a. A Machine Learning approach to the assessment of the vulnerability of Posidonia oceanica meadows. Ecol. Indic. 108, 105744. https://doi.org/10.1016/j.ecolind.2019.105744.

Catucci, E., Scardi, M., 2020b. Modeling Posidonia oceanica shoot density and rhizome primary production. Sci. Rep. 10, 16978. https://doi.org/10.1038/s41598-020-73722-9.

Catucci, E., Scardi, M., 2022. Fractal dimension of Posidonia oceanica meadows for the assessment of their ecological condition. Estuar., Coast. Shelf Sci. 274, 107925. https://doi.org/10.1016/j.ecss.2022.107925.

Cau, A., Carbonell, A., Follesa, M.C., Mannini, A., Relini, L.O., Politou, C.Y., Ragonese, S., Rinelli, P., 2002. MEDITS-based information on the deep water red shrimps Aristaeomorpha foliacea and Aristeus antennatus (Crustacea: Decapoda: Aristeidae). Sci. Mar. 66, 103–124.

Chan, T.Y., 1998. Shrimps and prawns. FAO species identification guide for fishery purposes. Living Mar. Resour. West. Cent. Pac. 2, 851–966.

Cohen, J., 1960. Kappa: Coefficient of concordance. Educ. Psych. Meas. 20, 37–46.

Coll, M., Steenbeek, J., Pennino, M.G., Buszowski, J., Kaschner, K., Lotze, H.K., Rousseau, Y., Tittensor, D.P., Walters, C., Watson, R.A., Christensen, V., 2020. Advancing Global Ecological Modeling Capabilities to Simulate Future Trajectories of Change in Marine Ecosystems. Front. Mar. Sci. 7.

Cossarini, G., Feudale, L., Teruzzi, A., Bolzon, G., Coidessa, G., Solidoro, C., Di Biagio, V., Amadio, C., Lazzari, P., Brosich, A., Salon, S., 2021. High-Resolution Reanalysis of the Mediterranean Sea Biogeochemistry (1999–2019). Front. Mar. Sci. 8.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Found. Trends Comput. Graph. Vis. 7, 81–227. https://doi.org/10.1561/0600000035.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random Forests. In: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning. Springer US, Boston, MA, pp. 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random Forests for Classification in Ecology. Ecology 88, 2783–2792.

D'Onghia, G., Maiorano, P., Capezzuto, F., Carlucci, R., Battista, D., Giove, A., Sion, L., Tursi, A., 2009. Further evidences of deep-sea recruitment of Aristeus antennatus (Crustacea: Decapoda) and its role in the population renewal on the exploited bottoms of the Mediterranean. Fish. Res. 95, 236–245.

Deval, M.C., 2019. Population dynamics and biological patterns of commercial crustacean species in the Antalya Bay, Eastern Mediterranean Sea: III. The giant red shrimp Aristaeomorpha foliacea Risso, 1827. Turk. J. Fish. Aquat. Sci. 20, 311–323.

Duarte, C.M., Amthor, J.S., Maranger, R.J., Pace, M.L., Pastor, J.J., Running, S.W., 2003. The limits to models in ecology, in: Eds.). Models in Ecosystem Science. Princeton University Press, pp. 437–451.

Dunham, A.E., 2013. 12. REALIZED NICHE OVERLAP, RESOURCE ABUNDANCE, AND INTENSITY OF INTERSPECIFIC COMPETITION. in: 12. REALIZED NICHE OVERLAP, RESOURCE ABUNDANCE, AND INTENSITY OF INTERSPECIFIC COMPETITION. Harvard University Press, pp. 261–280. https://doi.org/10.4159/harvard.9780674183384.c15.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M., E. Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Escudier, R., Clementi, E., Cipollone, A., Pistoia, J., Drudi, M., Grandi, A., Lyubartsev, V., Lecci, R., Aydogdu, A., Delrosso, D., Omar, M., Masina, S., Coppini, G., Pinardi, N., 2021. A High Resolution Reanalysis for the Mediterranean Sea. Front. Earth Sci. 9, 702285. https://doi.org/10.3389/feart.2021.702285.

Escudier, R., Clementi, E., Omar, M., Cipollone, A., Pistoia, J., Aydogdu, A., Drudi, M., Grandi, A., Lyubartsev, V., Lecci, R., 2020. Mediterranean Sea Physical Reanalysis (CMEMS MED-Currents)(version 1)[Data Set]. Copernicus Monitoring Environment Marine Service (CMEMS).

Fabbrizzi, E., Scardi, M., Ballesteros, E., Benedetti-Cecchi, L., Cebrian, E., Ceccherelli, G., De Leo, F., Deidun, A., Guarnieri, G., Falace, A., Fraissinet, S., Giommi, C., Mačić, V., Mangialajo, L., Mannino, A.M., Piazzi, L., Ramdani, M., Rilov, G., Rindi, L., Rizzo, L., Sarà, G., Souissi, J.B., Taskin, E., Fraschetti, S., 2020. Modeling Macroalgal Forest Distribution at Mediterranean Scale: Present Status, Drivers of Changes and Insights for Conservation and Management. Front. Mar. Sci. 7. https://doi.org/10.3389/fmars.2020.00020.

Foley, M.M., Halpern, B.S., Micheli, F., Armsby, M.H., Caldwell, M.R., Crain, C.M., Prahler, E., Rohr, N., Sivas, D., Beck, M.W., 2010. Guiding ecological principles for marine spatial planning. Mar. Policy 34, 955–966.

Garofalo, G., Fortibuoni, T., Gristina, M., Sinopoli, M., Fiorentino, F., 2011. Persistence and co-occurrence of demersal nurseries in the Strait of Sicily (central Mediterranean): Implications for fishery management. J. Sea Res. 66, 29–38. https://doi.org/10.1016/j.seares.2011.04.008.

Geary, W.L., Bode, M., Doherty, T.S., Fulton, E.A., Nimmo, D.G., Tulloch, A.I.T., Tulloch, V.J.D., Ritchie, E.G., 2020. A guide to ecosystem models and their environmental applications. Nat. Ecol. Evol. 4, 1459–1471. https://doi.org/10.1038/s41559-020-01298-8.

Giannoulaki, M., Belluscio, A., Colloca, F., Fraschetti, S., Scardi, M., Smith, C., Panayotidis, P., Valavanis, V., Spedicato, M.T. 2013 Mediterranean Sensitive Habitats. DG MARE Specific Contract SI2.600741, Final Report.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognit. Lett. 27, 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.

Guijarro, B., Bitetto, I., D'Onghia, G., Follesa, M.C., Kapiris, K., Mannini, A., Marković, O., Micallef, R., Ragonese, S., Skarvelis, K., Cau, A., 2019. Spatial and temporal patterns in the Mediterranean populations of Aristaeomorpha foliacea and Aristeus antennatus (Crustacea: Decapoda: Aristeidae) based on the MEDITS surveys. Sci. Mar. 83, 57. https://doi.org/10.3989/scimar.05012.04A.

Guillaumot, C., Moreau, C., Danis, B., Saucède, T., 2020. Extrapolation in species distribution modelling. Application to Southern Ocean marine species. Prog. Oceanogr. 188, 102438. https://doi.org/10.1016/j.pocean.2020.102438.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139028271.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

Hazen, E.L., Scales, K.L., Maxwell, S.M., Briscoe, D.K., Welch, H., Bograd, S.J., Bailey, H., Benson, S.R., Eguchi, T., Dewar, H., Kohin, S., Costa, D.P., Crowder, L.B., Lewison, R. L., 2018. A dynamic ocean management tool to reduce bycatch and support sustainable fisheries. Sci. Adv. 4, eaar3001. https://doi.org/10.1126/sciadv.aar3001.

Hijmans, R.J., 2018. raster: Geographic data analysis and modeling. R package version 2, 8.

Hirzel, A.H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. J. Appl. Ecol. 45, 1372–1381. https://doi.org/10.1111/j.1365-2664.2008.01524.x.

Howell, D., Schueller, A.M., Bentley, J.W., Buchheister, A., Chagaris, D., Cieri, M., Drew, K., Lundy, M.G., Pedreschi, D., Reid, D.G., Townsend, H., 2021. Combining Ecosystem and Single-Species Modeling to Provide Ecosystem-Based Fisheries Management Advice Within Current Management Systems. Front. Mar. Sci. 7.

Hurlbert, S.H., 1978. The Measurement of Niche Overlap and Some Relatives. Ecology 59, 67–77. https://doi.org/10.2307/1936632.

Kapiris, K., Thessalou-Legaki, M., 2011. Feeding ecology of the deep-water blue–red shrimp Aristeus antennatus (Decapoda: Aristeidae) in the Greek Ionian Sea (E. Mediterranean). J. Sea Res. 65, 151–160. https://doi.org/10.1016/j.seares.2010.09.005.

Keyl, F., Wolff, M., 2008. Environmental variability and fisheries: what can models do? Rev. Fish. Biol. Fish. 18, 273–299. https://doi.org/10.1007/s11160-007-9075-5.

Komai, T., Komatsu, H., Fujita, T., 2009. Deep-sea shrimps and lobsters (Crustacea: Decapoda) from northern Japan, collected during the project "Research on Deep-sea Fauna and Pollutants off Pacific Coast of Northern Japan. Natl. Mus. Nat. Sci. Monogr. 39, 495–580.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics 159–174.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. N. 2, 18–22.

Lindstrom, E., Gunn, J., Fischer, A., McCurdy, A., Glover, L.K., Members, T.T., 2012. A Framework for Ocean Observing. Eur. Space Agency. https://doi.org/10.5270/OceanObs09-FOO.

Link, J.S., Huse, G., Gaichas, S., Marshak, A.R., 2020. Changing how we approach fisheries: A first attempt at an operational framework for ecosystem approaches to fisheries management. Fish Fish 21, 393–434. https://doi.org/10.1111/faf.12438.

Louppe, G., 2014. Understanding Random Forests: From Theory to Practice. arXiv: 1407.7502 [stat].

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, 26. Curran Associates, Inc, pp. 431–439.

Masnadi, F., Criscoli, A., Lanteri, L., Mannini, A., Osio, G.C., Sartor, P., Sbrana, M., Ligas, A., 2018. Effects of environmental and anthropogenic drivers on the spatial distribution of deep-sea shrimps in the Ligurian and Tyrrhenian Seas (NW Mediterranean). Hydrobiologia 816, 165–178. https://doi.org/10.1007/s10750-018-3581-4.

Massutí, E., Monserrat, S., Oliver, P., Moranta, J., López-Jurado, J.L., Marcos, M., Hidalgo, M., Guijarro, B., Carbonell, A., Pereda, P., 2008. The influence of oceanographic scenarios on the population dynamics of demersal resources in the western Mediterranean: Hypothesis for hake and red shrimp off Balearic Islands. J. Mar. Syst., Wrap. IDEA Proj. 71, 421–438. https://doi.org/10.1016/j.jmarsys.2007.01.009.

Maynou, F., 2008. Environmental causes of the fluctuations of red shrimp (Aristeus antennatus) landings in the Catalan Sea. J. Mar. Syst., Wrap. IDEA Proj. 71, 294–302. https://doi.org/10.1016/j.jmarsys.2006.09.008.

Melo-Merino, S.M., Reyes-Bonilla, H., Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. Ecol. Model. 415. https://doi.org/10.1016/j.ecolmodel.2019.108837.

Merow, C., Smith, M.J., Edwards Jr, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281. https://doi.org/10.1111/ecog.00845.

Miloslavich, P., Bax, N.J., Simmons, S.E., Klein, E., Appeltans, W., Aburto-Oropeza, O., Andersen Garcia, M., Batten, S.D., Benedetti-Cecchi, L., Checkley, D.M., Chiba, S., Duffy, J.E., Dunn, D.C., Fischer, A., Gunn, J., Kudela, R., Marsac, F., Muller-Karger, F.E., Obura, D., Shin, Y.-J., 2018. Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. Glob. Change Biol. 24, 2416–2433. https://doi.org/10.1111/gcb.14108.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective, Adaptive computation and machine learning series. MIT Press, Cambridge, MA.

Mytilineou, C., Politou, C., Papaconstantinou, C., Kavadas, S., D onghia, G., Sion, L., 2005. Deep-water fish fauna in the Eastern Ionian Sea. Belg. J. Zool. 135, 229.

Orsi Relini, L., Mannini, A., Relini, G., 2013. Updating knowledge on growth, population dynamics, and ecology of the blue and red shrimp, Aristeus antennatus (Risso, 1816), on the basis of the study of its instars. Mar. Ecol. 34. https://doi.org/10.1111/j.1439-0485.2012.00528.x.

Owens, H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C.M., Myers, C.E., Peterson, A.T., 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. Ecol. Model. 263, 10–18. https://doi.org/10.1016/j.ecolmodel.2013.04.011.

Panzeri, D., Bitetto, I., Carlucci, R., Cipriano, G., Cossarini, G., D'Andrea, L., Masnadi, F., Querin, S., Reale, M., Russo, T., 2021. Developing spatial distribution models for demersal species by the integration of trawl surveys data and relevant ocean variables. J. OPERATIONAL Oceanogr. 14, s114–s123.

Panzeri, D., Russo, T., Arneri, E., Carlucci, R., Cossarini, G., Isajlović, I., Krstulović Šifner, S., Manfredi, C., Masnadi, F., Reale, M., Scarcella, G., Solidoro, C., Spedicato, M.T., Vrgoč, N., Zupa, W., Libralato, S., 2023. Identifying priority areas for spatial management of mixed fisheries using ensemble of multi-species distribution models. Fish. Fish. N./a. https://doi.org/10.1111/faf.12802.

Papaconstantinou, C., Kapiris, K., 2001. Distribution and population structure of the red shrimp (Aristeus antennatus) on an unexploited fishing ground in the Greek Ionian Sea. Aquat. Living Resour. 14, 303–312.

Pau, S., Wolkovich, E.M., Cook, B.I., Davies, T.J., Kraft, N.J.B., Bolmgren, K., Betancourt, J.L., Cleland, E.E., 2011. Predicting phenology by integrating ecology, evolution and climate science. Glob. Change Biol. 17, 3633–3643. https://doi.org/10.1111/j.1365-2486.2011.02515.x.

Pennino, M., Paradinas, I., Illian, J., Muñoz, F., Bellido, J., López-Quílez, A., Conesa, D., 2019. Accounting for preferential sampling in species distribution models. Ecol. Evol. 9, 653–663. https://doi.org/10.1002/ece3.4789.

Podda, C., Palmas, F., Cabiddu, S., Pesci, P., Sabatini, A., 2020. Exploring relationships between the distribution of giant red shrimp Aristaeomorpha foliacea (Risso, 1827) and environmental factors in the Central-Western Mediterranean Sea. Adv. Oceanogr. Limnol. 11. https://doi.org/10.4081/aiol.2020.9471.

Politou, C.-Y., Kavadas, S., Mytilineou, C., Tursi, A.R., Carlucci, R., Lembo, G., 2001. Fisheries resources in the deep-waters of the Eastern Mediterranean (Greek Ionian Sea). J. Northwest Atl. Fish. Sci. 31, 35–46.

R Development Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨https://www.R-project.org/⟩.

Revelle, W.R., 2017. psych: Procedures for personality and psychological research.

Rinelli, P., Bianchini, M.L., Casciaro, L., Giove, A., Mannini, A., Politou, C.-Y., Profeta, A., Ragonese, S., Sabatini, A., 2013. Occurrence and abundance of the deep-water red shrimps Aristeus antennatus (Risso, 1816) and Aristaeomorpha foliacea (Risso, 1827) in the central eastern Mediterranean Sea. Cah. Biol. Mar. 54, 335–347.

Robinson, L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P., Richardson, A.J., 2011. Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities: Marine species distribution models. Glob. Ecol. Biogeogr. 20, 789–802. https://doi.org/10.1111/j.1466-8238.2010.00636.x.

Russo, T., Catucci, E., Franceschini, S., Labanchi, L., Libralato, S., Sabatella, E.C., Sabatella, R.F., Parisi, A., Fiorentino, F., 2022. Defend as You Can, React Quickly: The Effects of the COVID-19 Shock on a Large Fishery of the Mediterranean Sea. Front. Mar. Sci. 9. https://doi.org/10.3389/fmars.2022.824857.

Sardà, F., Cartes, J.E., 1997. Morphological features and ecological aspects of early juvenile specimens of the aristeid shrimp Aristeus antennatus (Risso, 1816). Mar. Freshw. Res. 48, 73–77.

Sardà, F., Company, J.B., 2012. The deep-sea recruitment of Aristeus antennatus (Risso, 1816) (Crustacea: Decapoda) in the Mediterranean Sea. J. Mar. Syst. 105–108, 145–151. https://doi.org/10.1016/j.jmarsys.2012.07.006.

Sardà, F., D'Onghia, G., Politou, C.Y., Company, J.B., Maiorano, P., Kapiris, K., 2004. Deep-sea distribution, biological and ecological aspects of *Aristeus antennatus* (Risso, 1816) in the western and central Mediterranean Sea. Sci. Mar. 68, 117–127. https://doi.org/10.3989/scimar.2004.68s3117.

Schoener, T.W., 1974. Some Methods for Calculating Competition Coefficients from Resource-Utilization Spectra. Am. Nat. 108, 332–340. https://doi.org/10.1086/282911.

Scornet, E., 2017. Tuning parameters in random forests. ESAIM: Proc. Surv. 60, 144–162.

Shatnawi, R., 2017. The application of ROC analysis in threshold identification, data imbalance and metrics selection for software fault prediction. Innov. Syst. Softw. Eng. 13, 201–217. https://doi.org/10.1007/s11334-017-0295-0.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2015. Package 'ROCR.'. Vis. Perform. scoring Classif. 1–14.

Spedicato, M.T., Massutí, E., Mérigot, B., Tserpes, G., Jadaud, A., Relini, G., 2019. The MEDITS trawl survey specifications in an ecosystem approach to fishery management. Sci. Mar. 83, 9–20.

Stokland, J.N., Halvorsen, R., Støa, B., 2011. Species distribution modelling—Effect of design and sample size of pseudo-absence observations. Ecol. Model. 222, 1800–1809. https://doi.org/10.1016/j.ecolmodel.2011.02.025.

Teruzzi, A., Di Biagio, V., Feudale, L., Bolzon, G., Lazzari, P., Salon, S., 2021. Data from: mediterranean Sea Biogeochemical Reanalysis (CMEMS MED-Biogeochemistry, MedBFM3 system)(Version 1)[Data set]. Copernicus Monitoring Environment Marine Service (CMEMS), doi 10.

Townsend, H., Harvey, C.J., deReynier, Y., Davis, D., Zador, S.G., Gaichas, S., Weijerman, M., Hazen, E.L., Kaplan, I.C., 2019. Progress on Implementing Ecosystem-Based Fisheries Management in the United States Through the Use of Ecosystem Models and Analysis. Front. Mar. Sci. 6.

Wager, S., 2016. Asymptotic Theory for Random Forests.

Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecol. 9, 8. https://doi.org/10.1186/1472-6785-9-8.