



An open science automatic workflow for multi-model species distribution estimation

Gianpaolo Coro¹ · Lorenzo Sana¹ · Pasquale Bove²

Received: 3 August 2023 / Accepted: 29 January 2024
© The Author(s) 2024

Abstract

Integrated Environmental Assessment systems and ecosystem models study the links between anthropogenic and climatic pressures on marine ecosystems and help understand how to manage the effects of the unsustainable exploitation of ocean resources. However, these models have long implementation times, data and model interoperability issues and require heterogeneous competencies. Therefore, they would benefit from simplification, automatization, and enhanced integrability of the underlying models. Artificial Intelligence can help overcome several limitations by speeding up the modelling of crucial functional parts, e.g. estimating the environmental conditions fostering a species' persistence and proliferation in an area (the species' ecological niche) and, consequently, its geographical distribution. This paper presents a full-automatic workflow to estimate species' distributions through statistical and machine learning models. It embeds four ecological niche models with complementary approaches, i.e. Artificial Neural Networks, Maximum Entropy, Support Vector Machines, and AquaMaps. It automatically estimates the optimal model parametrisations and decision thresholds to distinguish between suitable- and unsuitable-habitat locations and combines the models within one ensemble model. Finally, it combines several ensemble models to produce a species richness map (biodiversity index). The software is open-source, Open Science compliant, and available as a Web Processing Service-standardised cloud computing service that enhances efficiency, integrability, cross-domain reusability, and experimental reproduction and repetition. We first assess workflow stability and sensitivity and then demonstrate effectiveness by producing a biodiversity index for the Mediterranean based on ~1500 species data. Moreover, we predict the spread of the invasive *Siganus rivulatus* in the Mediterranean and its current and future overlap with the native *Salpa salpa* under different climate change scenarios.

Keywords Ecological niche modelling · Marine science · Artificial intelligence · Open Science · Alien and invasive species

1 Introduction

The unsustainable exploitation of ocean resources—with overfishing, chemical and physical pollution, and heavy maritime traffic—threatens oceans, seas, and coasts. Climate

change further exacerbates this problem [1, 2]. Digital technologies are crucial tools to understand how to manage the effects of this pressure and potentially help mitigate it. For example, Integrated Environmental Assessment systems (IEAs) and ecosystem models (EMs) allow studying the links between anthropogenic driving forces, environmental pressures, and the response of ecosystems [3–8]. However, several works on these models [9–12] have highlighted frequent data interoperability and scalability issues, adoption of heuristic and non-automatic approaches, and results produced with limited transparency.

Artificial Intelligence (AI) methods can help overcome several limitations in this context, by speeding up the modelling of crucial parts of IEAs and EMs. For example, they can help automate the discovery of natural relations between the ecosystem, environmental conditions, and anthropogenic stressors [13–16], identify the essential data for assessing the

✉ Gianpaolo Coro
gianpaolo.coro@isti.cnr.it

Lorenzo Sana
lorenzo.sana@isti.cnr.it

Pasquale Bove
pasquale.bove@igg.cnr.it

¹ Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche (ISTI-CNR), Via Moruzzi 1, 56124 Pisa, Italy

² Istituto di Geoscienze e Georisorse, Consiglio Nazionale delle Ricerche (IGG-CNR), Via Moruzzi 1, 56124 Pisa, Italy

current ecosystem status [17–19], estimate potential species distribution change over time [20–22], and predict alien and invasive species spread in an area [23–26]. Autonomous algorithms requiring minimal parametrisation are crucial to speeding up the modelling, producing more objective results, and maximising model use by communities with heterogeneous competencies [27]. Another crucial action is to improve result transparency by endowing the algorithms with Open Science features of reproducibility, repeatability, and reusability of processes and results [28]. Open Science compliance requires the models to be available under recognised standards of interoperability and integrability and the published results to be repeatable and reproducible (after changing some model-parameter values). These features also guarantee the transparency of the results to decision-making authorities and foster the consideration of the results in policy making [29–32].

Ecological niche models (ENMs) can play a crucial role in this context. ENMs estimate the set of resources and environmental conditions that foster a species' persistence and proliferation in an area (the species' *ecological niche*) [33–37]. ENMs operationally assess a species' habitat as the locations, within a study area, with suitable environmental conditions to fall within the species' *ecological niche*. Consequently, they can predict the presence/absence of the species in the area [38, 39]. If the analysed area corresponds to the known native species' areal, then the ENM will estimate the *actual* (or *native*) species' distribution. Otherwise, in areas where the species has never been observed (e.g. far from the native region), the ENM will estimate the *potential* (or *suitable*) species' distribution [40]. Mathematically, a species' ecological niche is a region in a vector space (a *hypervolume*) of the environmental variables associated with the species' subsistence. The ENM effectiveness depends on correctly identifying the most complete set of environmental variables constituting the vector space dimensions [33]. Environmental variable selection for ENMs is frequently conducted through statistical analysis [41–45] or machine learning models [46–51]. Generally, an ENM uses statistical analysis, machine learning, or expert-defined rules to estimate a function relating an ecological entity (e.g. a species, community, or ecosystem) with specific environmental conditions defined on a set of environmental variables.

Typically, the input of an ENM is a set of environmental variables represented as spatial distributions over the analysis spatial extent, e.g. coming from extensive collections of satellite or in situ probe data [52–58]. Another input is a set of species observations in the area (*presence* locations) and locations where the habitat is either known to be unsuitable (*absence* locations) or is potentially unsuitable (*background* locations) [59, 60]. As the output, the ENMs estimate a multivariate function calculating the probability that a specific location is suitable habitat for the species, given the

environmental characteristics of that location. After building the prediction function, as a possible application the ENM can use it on a study area and different climatic scenarios (e.g. future environmental conditions) to produce new spatial distributions of habitat suitability probabilities [22, 24, 35, 61–66]. ENMs can therefore identify suitable areas fostering a species presence [66–69], and be effective even with few data available [47, 48, 70, 71].

ENMs are valuable in the view of helping IEAs and EMs with functional relations automatically extracted from the data. From the point of view of IEA and EM experts, an important feature is that the ENMs are simple to use, automatic, AI-based, and endowed with Open Science features simplifying their integration with large workflows [27, 32].

This paper contributes to automatise ecological niche modelling and add Open Science features to it, a frequent requirement from IEA and EM communities of practice in international projects [72, 73]. We present a fully automatic workflow to estimate potential and actual marine species distributions based on ENMs. It integrates four statistical and machine learning models (i.e. AquaMaps, Maximum Entropy, Artificial Neural Networks, and Support Vector Machines) whose complementarity is demonstrated by this and other studies (Sect. 2.2). Eventually, it combines the models within an *ensemble* model to merge their complementary indications and produce a more stable and reliable distribution. Finally, it combines the ensemble models of several species to produce one *biodiversity index* representing the total, punctual number of species potentially present in a marine area (i.e. the grid-based species richness [74–76]). The workflow can scale to process from local-to-global-scale areas at coarse-to-detailed spatial resolutions and extensive batches of species data. One novelty of our workflow with respect to other solutions [77] is its full automatism, achieved by searching for the optimal parametrisation of each ENM, and the complete integration between the different ENMs. Moreover, the workflow automatically estimates the optimal *decision* thresholds on each ENM prediction to distinguish between suitable- and unsuitable-habitat locations. Another novelty is the availability of the workflow as an open-source and Open Science (OS) compliant software: it is available as a standardised Web service integrated with an OS-compliant cloud computing e-Infrastructure, enabling experiment repeatability, reproducibility, and cross-domain reusability. The cloud computing platform allows for processing extensive batches of species and environmental data to quickly produce a multi-species biodiversity index and future projections, e.g. under different Representative Concentration Pathways (RCP) scenarios [78].

Through specific case studies, we demonstrate the effectiveness of our workflow at producing a biodiversity index for more than 1500 Mediterranean species. Moreover, we show its potential use for (i) predicting alien and invasive

species distributions, (ii) estimating the overlap between native species and competitor invasive species distributions, (iii) estimating the potential influence of climate change on invasions, and (iv) assessing the workflow-output sensitivity to the individual ENM contributions. In the case studies, we used RCP projections as future scenarios, instead of other projections (e.g. [79]), because we could access robust data from other ENM studies [80] and included RCP-based expert-curated distributions in the evaluation [81].

Our workflow belongs to big data processing methodologies. Our case studies present typical applications involving a vast number of species and a wide study area, yet the training sets for our models can be relatively small. When applied to large species-batch processing, it is important to view the output of our workflow as the analysis of macro-patterns of species richness change rather than as detailed, cell- and species-specific answers. An erroneous species presence or absence assessment does not significantly impact macroscopic trend analyses. Other studies have shown that individual biases have limited influence when numerous models are aggregated to analyse overall trends in a study area, because model combinations produce more stable and reliable macroscopic knowledge than the individual models [82, 83]. However, it is advisable to exercise caution when interpreting the results of our workflow across multiple species, as punctually reliable indications cannot be guaranteed. On the other hand, macro-patterns can provide crucial information for ecosystem modelling and conservation, such as predicting future biodiversity changes well in advance and planning preventive actions. Our objective is to present a framework that, by default, combines models for various species to examine the aggregated model's changes over time and extract viable macro-patterns.

This paper is organised as follows: Sect. 2 describes our workflow, its Open Science-oriented Web service version, and the evaluation methodology used. Section 3, describes the workflow sensitivity analysis, the Mediterranean biodiversity index produced, and a quantitative and qualitative evaluation of the future prediction of the Mediterranean invasion by *Siganus rivulatus* and its overlap with the native *Sarpa salpa*. Finally, Sect. 4 draws the conclusions.

2 Methods

This section describes our workflow for generating ENMs, ensemble models, and a biodiversity index. Our workflow can process from local- to global-scale areas and easily manage an extensive set of species data. It is open source (“Supplementary information”) and requires minimal input, i.e. a collection of environmental variables (in the form of raster files) and species’ observation points. All options can be managed through a workflow configuration file without changing

the code. The workflow has also a Web service version with a standardised interface that complies with the Open Science directives [28] and allows for concurrently producing many species models and quickly estimating biodiversity indices.

In the following, we describe all components of our workflow, according to the schema reported in Fig. 1, i.e. data provisioning and pre-processing (Sect. 2.1), the ecological niche models integrated (Sect. 2.2), the ensemble model construction (Sect. 2.3), and the biodiversity index construction (Sect. 2.4). Moreover, we describe the Open Science-oriented Web service associated (Sect. 2.5) and the evaluation methodology used (Sect. 2.6).

2.1 Data provisioning

A user of our workflow should prepare a set of raster files of environmental variables in the ESRI-GRID (ASCII) format, one of the most frequently used formats by ecological and ecosystem models [80]. Each file should contain the distribution of one environmental variable (e.g. geophysical, oceanographic, or biological) that could be relevant for modelling the species response. The raster file should be defined on a regular grid of resolution R over the spatial extent of the area under analysis. The grid may contain pixels, i.e. raster cells, where the variable is undefined. Only pixels with fully defined variables will be utilised in the ENMs; in other words, we exclude raster cells with undefined (NA) values. All files should refer to the same spatial extent, have resolution R , and use the WGS84 coordinate system. The workflow will use these files to set the focus spatial extent and the final model resolution. It will produce output at the exact resolution of the input variables to avoid introducing re-sampling biases. The environmental variable files can come from large providers of Findable, Accessible, Interoperable, and Reusable (FAIR) data, such as Copernicus [84] and EMODnet [85], and other sources [80].

As an additional input file, the user should provide a comma-separated-values (CSV) file containing coordinate pairs of a species’ observations within the focus area (*presence* locations). This file should report one column containing the species’ scientific name (*scientific name*) and two data columns (*longitude*, *latitude*) with observation coordinate pairs (in the WGS84 coordinate system). These data could come from large providers of FAIR data of species observations, e.g. OBIS [86] and GBIF [87]. Multiple species-observation files can also be provided, which the workflow will process sequentially.

As the first processing step, our workflow associates a vector of environmental variables to the observation points (data *enrichment* operation). It extracts the variable values from each environmental variable’s raster file and associates the value of the closest raster cell to the observation’s coor-

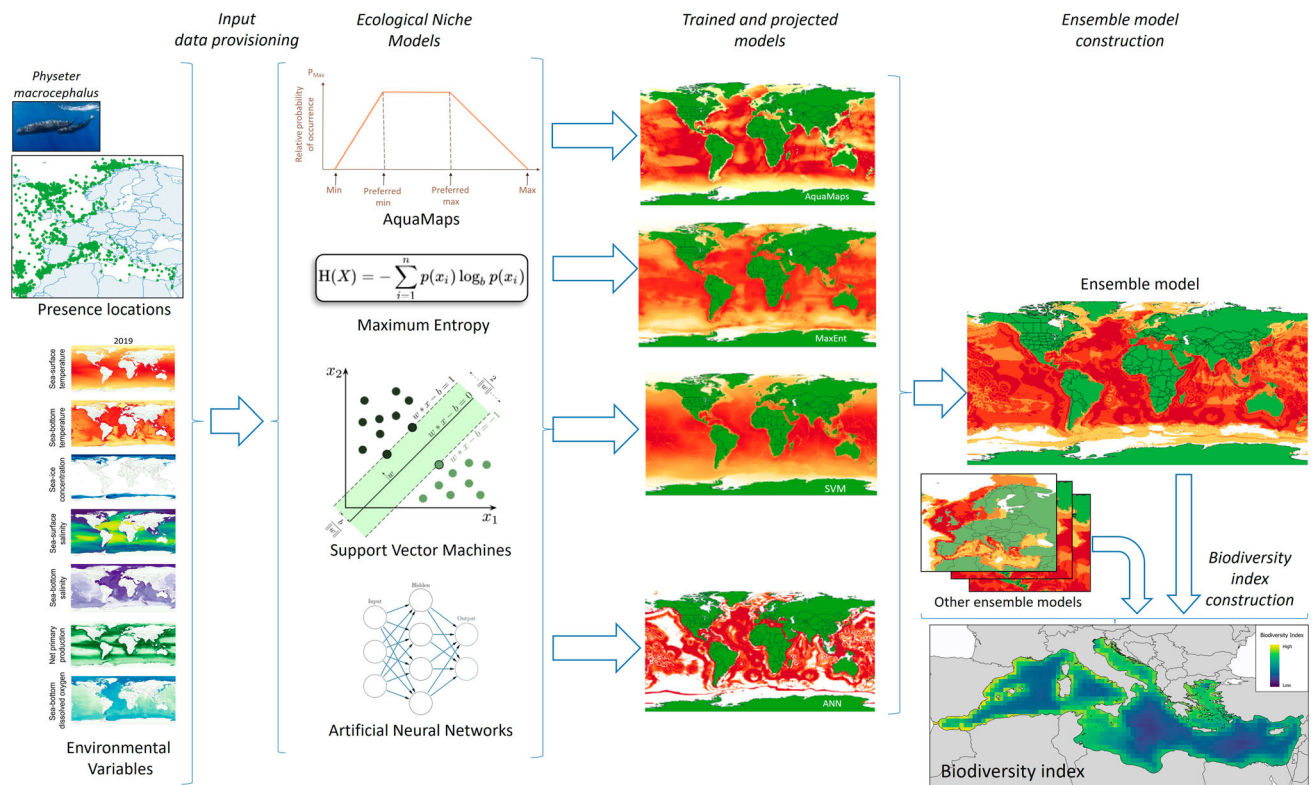


Fig. 1 General schema of our workflow

dinates. Eventually, this operation associates a vector of environmental variable values to each observation.

As a further step, the workflow proposes the *background sampling* strategy described in [88] and other works [89–93] to generate background absence data which potentially represent locations with limited suitability, or completely unsuitable, conditions for species presence. This technique consists in taking a random sample of locations from the study area's spatial extent (at resolution R), while excluding presence locations, for generating background absences during modelling. If many species observations were available from extended surveys in the native area, the background samples' locations would likely include absence locations [94]. This is the default strategy adopted by the workflow. As an alternative to background sampling and in compliance with common approaches in ecological niche modelling [60, 93, 95, 96], the workflow also allows users to externally provide reliable species absence (associated with known unsuitable habitat) or pseudo-absence locations as CSV files.

Proposing background sampling as the default strategy is particularly useful for user scenarios where species absence information is missing, e.g. for processes involving many species. Background points cannot substitute absence locations, but they can approximate negative examples, especially for nonlinear models such as Artificial Neural Networks,

MaxEnt, and Support Vector Machines (when using non-linear kernels) [97, 98]. For instance, MaxEnt can produce reasonable distributions, similar to those using reliable absences, while slightly overestimating species' presence [93]. Overestimation is reduced if the model internally enables more complex modelling of the relation between species' presence and environmental features through *hinge* features (Sect. 2.2.2) [93]. To further soften overestimation, our workflow offers a post-processing algorithm for native areas that reduces the probability of habitat suitability locations far from the species' presence locations (Sect. 2.2.5). Apart from these considerations, the reliability of background samples as absence locations can be verified through techniques evaluating how much these locations correspond to entirely different environmental conditions than those of the presence locations [80, 93, 99].

The data preparation algorithm, for each species to analyse, can be summarised as follows:

Two sets of generated environmental variable vectors ($\{f_p\}$ and $\{f_a\}$) represent multidimensional presence and absence reference environmental vectors. The ENMs will learn to distinguish these vectors by tracing hyper-volumes separating them, with some tolerance when using *background* points instead of absence locations. The additional

Algorithm 1 Data preparation and pre-processing

Read all user-provided environmental variable raster files $\{e_1, e_2, \dots, e_n\}$
 Generate a matrix-grid G_R for each variable, with the grid step representing the spatial resolution R and the matrix size corresponding to the focus-area extent
 Read all user-provided presence locations
 Generate a maximum of 10,000 random locations from the focus area not including presence locations (*background* points)
 For each presence location p :

Assign the environmental variable values to p , to obtain one vector $f_p = e_{1_p}, e_{2_p}, \dots, e_{n_p}$

For each *background* location a :

Assign the environmental variable values to a , to obtain one vector $f_a = e_{1_a}, e_{2_a}, \dots, e_{n_a}$

For each grid pixel g :

Assign the environmental variable values to g , to obtain one vector $f_g = e_{1_g}, e_{2_g}, \dots, e_{n_g}$

Save the $\{f_p\}$, $\{f_a\}$, $\{f_g\}$ vector sets as R objects.

vector set $\{f_g\}$ represents the vectors on which the trained ENMs will produce output projections.

2.2 Ecological niche models integrated

This section describes the ENMs integrated with our workflow, i.e. AquaMaps, Maximum Entropy, Artificial Neural Networks, and Support Vector Machines. These models simulate the probability that a specific location is suitable habitat for a species, given the values of a set of environmental variables associated with that location (*posterior* probability of habitat suitability). Their output, over a set of square-sized locations in an area (also outside of the native region), can be interpreted as the spatial distribution of the probability that the species could subsist in each area location (*potential* species distribution). We also describe an adjustment algorithm for estimating a species' *actual* distribution from the potential distribution when the analysed area corresponds to the species' native region.

Each ENM integrated can optionally be deactivated to use only a subset of models or substitute one model with another by other software. All ENMs can work in either training or projection mode. In training mode, they internally learn a function relating the species' habitat suitability to a set of input-provided environmental variables. In projection mode, the previously trained models project the function on the gridded environmental data associated with an area or future climatic scenarios.

One principal requirement in the design of our workflow was the necessity to restrict the number of ENMs integrated into a small core of models with complementary approaches. This requirement, also faced by other ENM frameworks

[77], comes from the fact that we aimed to reach communities of practice aside from expert ecological modellers. These communities needed (i) simplification of input/output through full automatism, (ii) integration between the ENMs, (iii) cross-community usage through Open Science features, and (iv) a computationally efficient and effective solution. To this aim, we selected AquaMaps as a representative of ENMs adopting a factorised approach in the analysis of the environmental variables (similarly to Decision Trees), because it estimates functional relations for one variable at a time and eventually combines the functions together (Sect. 2.2.1). We chose Maximum Entropy as a representative of Generalised Linear Models (GLMs), because it is equivalent to the Poisson regression GLM, which is naturally suited for species distribution modelling (as better reported in Sect. 2.2.2). We used Support Vector Machines to include linear, polynomial, and other basic nonlinear classification models (Sect. 2.2.4). Finally, we integrated Artificial Neural Networks to model up to non-convex functions and simulate very complex ecological niche functions (Sect. 2.2.3). In a previous study, we verified that these models actually bring complementary information important for improving model robustness [24], i.e. removing one of the models would result in lower prediction performance on known species locations. This feature was also confirmed by the results of the present work (Sect. 3). In the view of processing extensive batches of species data, the simplification strategy adopted by our workflow required to fix the default model optimisation strategies of the ENMs to the most commonly used ones, while excluding other alternatives [100, 101].

2.2.1 AquaMaps

AquaMaps [102] is a presence-only ENM that incorporates scientific expert knowledge to account for known biases and limitations of marine species occurrence data [103]. The models' name corresponds to two model implementations estimating the *actual* (AquaMaps-*native*) and *potential* species' habitat distributions (AquaMaps-*suitable*). The main difference between the two models is that AquaMaps-*native* restricts the distribution to the areas where the species is known to live.

AquaMaps can produce from regional- to global-scale distributions at a 0.5° resolution. The algorithm models the association between the observed locations and the environmental variables as the multiplication between mutually independent *envelope* functions, each traced on one environmental variable at a time. The envelope function is a trapezoidal function, normalised to 1, traced over the quartiles of the density of one variable values over the observations [104]. A positive slope (starting from 0) connects the 1st and 2nd quartiles, a flat region with 1-value lies between the 2nd and 3rd quartile, and a negative slope from the 3rd

quartile ends at 0 at the 4th quartile. A location whose associated environmental variable values fall in the flat regions of all envelope functions will have a habitat suitability probability equal to 1. If the values fall outside of all envelope function values, the location will have 0 habitat suitability probability associated. AquaMaps also applies mechanistic assumptions as rule-based sub-routines to revise the estimations. The default input environmental variables used by the algorithm are (i) sea-bottom and sea-surface temperature, (ii) distance from land, (iii) maximum, mean, and minimum depth, (iv) net primary production, (v) sea ice concentration, (vi) sea-bottom and sea-surface practical salinity, (vii) and sea-bottom moles of oxygen per unit of mass.

AquaMaps produces reasonably good results if compared to more complex approaches, but it often requires the bounding boxes and envelope functions to be revised by an expert [103]. One main advantage of this process is that it does not require optimisation, because expert-provided rules are embedded within the code and environmental variable quartiles are automatically extracted from the data.

AquaMaps can work with environmental variables projected for the short- and long-term future under different RCP scenarios. The AquaMaps Web site publishes expert-curated projections for 2050 and 2100 under the RCP8.5 and RCP4.5 scenarios [81, 105], which represent valuable references to evaluate other models' projections. We integrated AquaMaps as an R procedure, re-programmed from the original PHP algorithm code. We also included the expert-defined sub-routines in our re-implementation. However, to fully exploit the quality of the AquaMaps models in our case studies, we used the expert-revised distributions from the AquaMaps Web site, when available (Sect. 2.6). Our workflow users can, in fact, import the AquaMaps files to use them instead of our implementation (Sect. 2.3).

2.2.2 Maximum entropy

Maximum Entropy (MaxEnt) is a machine learning model frequently used in ecological modelling [48, 94, 106–112]. MaxEnt estimates a function $\pi(\bar{x})$ interpretable as the probability of species habitat suitability given the vector of environmental variables \bar{x} . This function has two principal constraints: (i) it has to comply with the mean values at the species presence locations, and (ii) its associated entropy function ($H = -\sum \pi(\bar{x}) \ln(\pi(\bar{x}))$) should be maximum [94, 95, 113]. MaxEnt performs a relative maximisation of the entropy function on the presence locations with respect to the entropy function on the *background* points [106]. It builds the $\pi(\bar{x})$ function to represent the complex relation existing between specific environmental variable combinations and the species habitat suitability [24, 33]. One advantage of this model is that it works well also when species presence data are only available (i.e. without absence data). However, it

is over-sensitive to biased presence and environmental data [48, 114] and might overfit small datasets [113, 115]. MaxEnt can be preferred over linear and logistic regression because it is equivalent to a Poisson regression (a GLM), a model naturally suited for modelling the probability of a number of events occurring in a fixed space such as species occurrences [116].

During the training phase, MaxEnt estimates the coefficients of a linear combination of the environmental variables, which is the core of the $\pi(\bar{x})$ function corresponding to maximum entropy [95]. These coefficient represent the variables' weights in the species' environmental preferences (named *per cent contribution*). We integrated a MaxEnt implementation by Phillips et al. [117] within our workflow. Our workflow configuration file also allows setting the species *prevalence*, i.e. the prior species occurrence probability in the area, which MaxEnt uses for modelling $\pi(\bar{x})$. This parameter is set 0.5 (uninformative) by default, assuming that no prior information is available for the species presence in the area and the model should entirely rely upon the data.

To reduce model overfitting issues, we followed the heuristic indications of other studies on MaxEnt parametrisation [94, 117–120]: we allowed the inclusion of presence points among the background points and included different types of *hinge* features in the $\pi(\bar{x})$ function to model complex species response to the environmental variables as alternatives to linear combination. The MaxEnt software we used indeed allows combining the environmental variables within $\pi(\bar{x})$ through simple-to-complex functions (*hinge* functions) to model species-environment relations. We configured the software to exhaustively test *hinge* functions among *linear* (the standard combination), *quadratic*, *product*, and *threshold* functions, and all their combinations. Eventually, the software selects the configuration producing the highest Area Under the Curve (AUC) [113]. AUC is the integral of the receiver operating characteristic (ROC) curve that reports the true-positive rate vs false-positive rate using various decision thresholds on the model output. An AUC value close to 1 indicates high-quality model training, whereas an AUC close to 0 indicates low-quality model training. Depending on the representativeness of the presence and background points, the AUC and optimal *hinge* function can change across different MaxEnt executions (Sect. 3.1), but are mostly stable with high-quality data [48].

2.2.3 Artificial neural networks

Artificial Neural Networks (ANNs) are machine learning models constituted by interconnected representations of biological neurons [121]. ANNs are extensively used ecological modelling [122–126], because they allow for modelling complex, nonlinear functions [127]. ANNs can also perform classification by discretising their output values over differ-

ent classes [128]. In Feed-Forward Neural Networks [129] (used in our workflow), the digital neurons are organised into “layers”. The first layer receives and processes the input vector directly; the last layer produces the output vector; and intermediate layers (“hidden layers”) process the in-between information. One layer is fully connected only to the next layer through weighted edges, i.e. a neuron in one layer is connected to all neurons of the next layer. An ANN can be trained on known data acting as examples. A learning algorithm (e.g. “backpropagation” [130]) adjusts the edge weights to produce the expected output on the training data. To assess prediction accuracy, a trained ANN can be used to produce estimates on known input data not included in the training set (*test* data). The optimal number of hidden layers and neurons can be found by testing different topologies [131], e.g. by adding neurons and layers as far as the error on the test set decreases (“growing” strategy [128]). One drawback of ANNs is that they do not provide the analytical form of the simulated function combining the input variables. Unlike traditional mathematical models that yield an explicit equation describing the relationship between input and output variables, ANNs operate as complex, interconnected systems of neurons, and the mapping between inputs and outputs is hardly expressible in a concise mathematical form. The lack of a readily interpretable analytical expression can make it more difficult for researchers and practitioners to gain a direct, intuitive understanding of the underlying relationships encoded within the ANN. Moreover, the ANN performance can be sensible to the network weight initialisation.

We integrated ANNs through the *neuralnet* R package [132]. The ANN-based ENM has one input per environmental feature and one output neuron producing a number between 0 (unsuitability) and 1 (high suitability). If externally provided absence locations were available, the ANN will use them as a reference for unsuitability; otherwise, it will approximate unsuitable environmental conditions using the background points described in Sect. 2.1 (similarly to ANN-based approaches in other domains [98, 133]).

As the default strategy to automatically select the optimal ANN topology, our workflow splits the training set into 10 parts. It iteratively trains the ANN with 9 parts and tests it with the remaining part (tenfold *cross-validation*). Our workflow uses a “growing” strategy to identify the optimal number of *hidden* neurons and layers achieving the highest average cross-validation accuracy. As the default configuration, our workflow tests between one and two *hidden* layers and from 10 to 200 neurons in each layer. This configuration results from our previous works estimating and testing the range of layers and neurons typically required to process species observation data from OBIS and GBIF without overfitting the model [22, 47, 48, 134]. The validity of this configuration was also tested on the case studies of the present paper

(Sect. 2.6). This setup is unsurprising because a two-hidden-layer ANN can simulate any multivariate nonlinear function and complex classification regions in the input space [128, 135, 136]. Nevertheless, the number of layers and neurons can easily be changed from the workflow configuration file to test more complex architectures.

We used cross-validation, instead of other techniques, due to the high variability of the number of species occurrence records our workflow could encounter during an extensive species-batch process. The number of occurrence records depends on many factors, such as the scientific surveys’ sampling frequency and extent, the species’ commonness, and the population change over time. Therefore, the number of presence points is normally low unless a species is widespread, frequently observed, and frequently targeted by scientific surveys. As we observed in our Mediterranean Sea case study, the number of occurrence records across a set of many species follows a log-normal-like distribution, with most species having few occurrences associated (10–500) and fewer species having many occurrences associated (over 1000–10,000). Since our aim was to provide a solution for processing large batches of species data, we adopted cross-validation as a strategy commonly used by other ENMs that could work for both data-poor and data-rich cases [93].

2.2.4 Support vector machines

Support Vector Machines (SVMs, [137]) are a machine learning method frequently used to build binary classifiers, also in ecological modelling [24, 138–142]. The method projects the input data onto a higher-dimensional feature space through a *kernel* function. Then, it searches for a linear separation of this space into two regions. The training algorithm searches for the optimal separation hyperplane maximising the distance (*margin*) of the training vectors of different classes from the hyperplane. The closest training vectors to the margin are named *support vectors*. The SVM training process searches for the optimal hyperplane and the largest *margin* through a fast optimisation algorithm. The requirement to sharply separate the two classes can be relaxed—to avoid overfitting—by allowing some classification error through *cost* weights [143]. After training, the distance of a vector from the separation hyperplane can be used to simulate a probability function. An output value of 1 corresponds to a vector confidently falling in the region above the hyperplane (interpreted as suitable habitat in our ENM). Conversely, a 0 output value corresponds to a vector well below the hyperplane (unsuitable habitat). During the projection phase, an SVM receives an environmental variable vector, transforms it through the kernel function, and calculates its belonging class. Then, it assigns a score between 0 and 1 according to the normalised distance from the hyperplane within the belonging class.

SVMs require an accurate parametrisation to optimise vector separability. We integrated the SVM implementation of the *e1071* R package [144]. This software can manage four kernel function types (*linear*, *polynomial*, *radial*, and *sigmoid*) with different parametrisations. Moreover, it allows adjusting penalties (*costs*) for misclassifications during training to reduce data-overfitting risk. Our workflow tests all supported kernel functions with multiple parametrisations. Specifically, it tests the performance of linear, three- and four-grade polynomials, and the radial and sigmoid functions with their *gamma* parameter ranging between 10^{-3} and 10^2 . Moreover, it tests all *cost* values between 10^{-3} and 10^2 . During training, the workflow conducts a tenfold cross-validation for each configuration and eventually selects the optimal parametrisation (i.e. the optimal kernel function and parameters and costs). The workflow configuration file includes a section that allows the kernel functions to be selected and the parameter ranges to be changed for testing different parametrisations.

2.2.5 Native distribution adjustment

The integrated ENMs estimate the potential species' habitat suitability on a projection grid. However, within the spatial extent of the native area, some species might only inhabit a subset of all suitable-habitat locations. This could be attributed to factors such as geographical obstacles or environmental hindrances preventing access to certain areas. The exact knowledge of these hindrances would require specific analyses of the species' behaviour and native area, which is usually unavailable to ENMs. The *native* distribution indeed corresponds to the *actual* distribution in the native bounding box only if complete knowledge about the environmental conditions for the species' subsistence is available and correctly captured by ENMs [103]. Knowledge gaps, often present in practical applications, can create a significant discrepancy between the species' native and actual distributions [60]. Therefore, an enhancement of the native distribution estimation is required to include the effects of other unknown variables regulating the species' presence in the area. Such refinement can approximately be obtained by analysing the species observations' distribution instead of searching for additional environmental variables. Suppose, for example, a reasonable number of species observations is available. In such case, the observations' spatial distribution implicitly indicates if the species is spread across the territory or localised in specific regions [145]. For example, if a species were localised in a coastal area, its observations would likely present accumulation regions close to the coast and fewer, less dense points far from the coasts. Therefore, an analysis of the distribution of minimum distances between the observations could indicate if we can expect species observations very far from the available observations. Our

workflow approximates the distribution of the mutual distances between the observations as a log-normal distribution to create a decay weighting function for the species' presence in the *native* area. The log-normal shape of this distribution derives from our previous heuristic analyses of the OBIS data [24, 145]. This shape is not an ecological property of species, but depends on the partial, sampled species information contained in large observation-data collections [60, 146, 147]. Distances below the upper confidence limit of this distribution can be considered *plausible* for observing the species. Conversely, distances higher than the upper confidence limit can be classified as *too far* (i.e. less plausible). The log-normal distribution can be used for these locations as a multiplicative decay function for the ENM probability function. This way, the locations far from areas with a high observation density are assigned a lower habitat suitability probability. In other studies, we have demonstrated that this technique can effectively simulate geographical reachability from habitat suitability [24, 44]. Our workflow thus produces a new weighted ENM distribution by multiplying the *too far* ENM output values by the decay function value. This new ENM distribution is an approximation of the native species distribution. The workflow user can activate this *native-distribution adjustment* of the potential habitat distribution through the workflow configuration file.

The algorithm can be summarised as follows:

Algorithm 2 Native distribution adjustment

```

For each presence location  $p_i$ :
    Calculate the distance  $d(p_i, p_j)$  (with  $i \neq j$ ) from all other
    presence locations
    Record the minimum distance  $d_{min}(p_i)$ 
    Approximate the  $d_{min}(\cdot)$  distribution with a log-normal distribution
     $L(d_{min}(\cdot))$  (with height normalised to 1)
    Calculate the upper confidence limit of the log-normal distribution
     $d_{min-upper}$ 
    Scan the  $G_R$  grid of an ENM ( $m$ ), and for each pixel  $g$ :
        Calculate the minimum distance from a presence location
         $d_{min}(g)$ 
        If  $d_{min}(g) > d_{min-upper} \rightarrow m'(g) = L(d_{min}(g)) \cdot m(g)$ , with
         $m(g)$  being the ENM value in  $g$ ; substitute the ENM value
         $m(g)$  with  $m'(g)$ 
    Update all  $m$  models.

```

This algorithm thus adjusts the *less plausible* ENM values through a log-normal decay function and eventually overwrites the previous ENM output files with the new values. Since the algorithm does not process environmental variables, we added it as an optional adjustment for native distributions instead of including it as an additional ENM.

2.3 Ensemble model construction

Our workflow executes all ENMs concurrently and generates one spatial distribution of habitat suitability for each, in the ESRI-GRID (ASCII) format. The workflow also produces one metadata file for each model, indicating the optimal model variables and the optimal decision threshold for dichotomic classification (suitable/unsuitable habitat). This threshold, which is likely different for each ENM, is the cut-off value maximising the prediction accuracy of the optimal model. It allows assessing the R -squared cells (*grid pixels*) corresponding to suitable (1) and unsuitable (0) habitat for each ENM. Specifically, the workflow selects the numeric threshold, over the optimal model, that maximises the separation of the training presence and background locations into suitable and unsuitable locations, respectively. For this task, we use a similar strategy to ROC curve and AUC calculation [93], i.e. we vary a numeric threshold over the outputs of the optimal model on the training set and eventually select the value corresponding to the highest prediction accuracy. The workflow uses this strategy for all models. We preferred it to alternative thresholds—e.g. omission-sensitivity balance, equal sensitivity–specificity, and others [148]—which would have required specific prior knowledge on the models’ performance for each species data and area, normally unavailable for large sets of species.

Based on this binarisation, as a further computational step, our workflow generates an *ensemble* distribution that takes advantage of the complementary properties between the models, consequent to their likely different functional forms and training processes [24]. For this task, our workflow sums the corresponding pixels’ binary values. This operation generates an ensemble model ready for a consensus-based model like the biodiversity index model described in the next section. This approach is compliant with general consensus-based classifiers [149–153] and simulates different experts assessing species presence cell by cell. The underlying assumption is that each model produces independent assessments of the species’ presence because they use different rationales, resulting in independent probability distributions. Other works on weighted-averaging machine learning have indeed demonstrated that such distributions—especially those of nonlinear models—are hardly comparable and often have forms not associable with known probability distributions [24, 82, 126, 154, 155]. Averaging their values might bias the results towards the sharpest distributions, whose high values do not necessarily indicate high model confidence and robust training. Overall, we use the sum of the binarised models as the sum of mutually independent probabilities. On the one hand, this strategy loses definition in assessing the ensemble probability over the area. On the other hand, it prevents the production of biased results due to the combination of too different probability distributions.

One critical point of this ensemble strategy is that it assigns equal weight to all models, with the rationale that suitability could be due even to one distribution estimating a high probability. This choice depends on the fact that the default configuration of our workflow mainly addresses the production of a species richness index based on a large set of species. It is therefore conceived for feeding other processes that aim to discover macro-patterns of species richness change over time and space. Discovering macro patterns is in fact more affordable than producing accurate cell-wise assessments [83, 156]. Reliable patterns can indeed emerge from the statistical aggregation of many model outputs, even if these contain individual biases. The biases blur when several models are aggregated and can fade away when overall trend discovery is the target of the analysis. In this view, assigning the same weight to every model means assuming we miss prior information about the optimal model(s) for each species, which is a common condition for large species batches. Therefore, also considering that the involved models likely bring complementary information (Sect. 2.2), we assigned the same weight to every model as the default workflow configuration. Nevertheless, our tool allows the modification of the models’ weights in the ensemble when prior information on the models’ performance is available.

The algorithm to build the ensemble model can be summarised as follows:

Algorithm 3 Ensemble model construction

```

Train the individual ENMs → detect the optimal model variables
and decision thresholds  $\{t_1, t_2, t_3, t_4\}$ 
(Optional) Add native distribution adjustment
Import the ENM results and metadata
Scan the  $G_R$  grid, and for each pixel  $g$ :

    Retrieve the associated environmental variable vector  $\{f_g\}$ 
    For each ENM  $m$ :
        Calculate the model binary habitat-suitability assessment
         $d_m(g) = \{1 \text{ if } m(f_g) \geq t_m; 0 \text{ otherwise}\}$ 
        Sum the binary assessments to calculate the ensemble-model
        value for pixel  $g$ :  $E(g) = \sum_1^4 d_m(g)$ 

Save all  $E(g)$  values to an ESRI-GRID (ASCII) file.

```

As the result, this algorithm produces one spatial distribution (in the ESRI-GRID format, at resolution R) reporting the pixel-by-pixel total number of models that indicated suitable habitat in the focus area (*ensemble* model).

The ensemble model construction algorithm can also integrate new models or alternative versions of the integrated ENMs. This modular design allows for easily extending the number and types of models to combine. During the data import phase, the algorithm can read the ESRI-GRID files of other ENMs’ outputs, with plain text metadata associated, to include these distributions in the ensemble. We also

provide conversion tools (“Supplementary information”) to transform plain CSV files (with *longitude*, *latitude*, *probability* columns) containing species distribution information into integrable data. This procedure allowed us, for example, to integrate the official, expert-curated AquaMaps distributions in some evaluation cases instead of the automatically estimated ones (Sect. 2.6).

The ensemble model can be used to estimate an *overlap* index between multiple ensemble models. This index estimates the number of different, punctual, overlapping species habitats. Our workflow performs this operation by transforming the species ensemble habitat distributions into binary distributions (suitable/unsuitable habitats) and then summing them pixel by pixel. One valid heuristic to transform the ensemble distribution into a binary distribution considers the agreement between the component models [24]: if at least three models out of four indicate suitable environmental conditions for the species in a grid pixel, this location can overall be classified as suitable habitat. As a generalisation, since in our workflow the number individual models can be changed, the minimum agreement between the ENMs for assessing potential distribution is “number of models – 1”. In the case of only one model used, the minimum would be 1. This threshold can be reduced to set more relaxed habitat suitability assessment conditions. For example, a one-model threshold would allow identifying locations with minimal potential habitat suitability.

2.4 Biodiversity index construction

An overall cross-species *overlap* index is obtained by summing the binary ensemble model values. This index measures the number of different species potentially living in each pixel of the analysed area, i.e. it can be used as a proxy for a *biodiversity* index (or *species richness* index, alternatively). Although this index does not consider the species’ mutual interactions, it can be demonstrated that this type of approach can produce reliable biodiversity information for long time frames (e.g. over one year) [33, 35].

The biodiversity index construction algorithm can be summarised as follows:

As a result, this algorithm produces one spatial distribution (in the ESRI-GRID format, at resolution R) reporting the pixel-by-pixel total number of species’ ensemble models indicating suitable locations. This is the final result of our workflow.

2.5 Open science-oriented web service

We developed our workflow as an open-source R software suite (“Supplementary information”). The software availability as a suite (internally using almost R-base packages only) instead of a CRAN package increases its integrabil-

Algorithm 4 Biodiversity index construction

For each species s under analysis:

Retrieve the ensemble distribution E_s

Scan the G_R grid, and for each pixel g :

For each species s :

Calculate the *ensemble* model binary habitat-suitability assessment $d_s(g) = \{1 \text{ if } E_s(g) \geq |m| - 1; 0 \text{ otherwise}\}$, with $|m|$ being the number of models

Sum the binary assessments to estimate the *biodiversity index* for pixel g : $I(g) = \sum_s d_s(g)$

Save all $I(g)$ values to an ESRI-GRID (ASCII) file.

ity, maintenance, and long-term compatibility with multiple versions of R. This choice also made it easy to publish the workflow as a Web service supporting secure cloud- and parallel-processing and Open Science-oriented features [28]. To this aim, we integrated it with the DataMiner cloud computing platform of the D4Science e-Infrastructure [49, 157–159], which published the process under the Web Processing Service standard of the Open Geospatial Consortium (OGC-WPS [160]) (“Supplementary information”). This standard allows for direct integration of the process with widely used geospatial data processing software supporting it (e.g. QGIS and ArcGIS) [32]. DataMiner automatically produced a graphic user interface based on the software input/output definitions. It also tracks the model parameters, input and output data at each execution (computational *provenance*) in a user’s private data space as XML documents following the Prov-O ontological specifications [161]. Provenance tracking is crucial for computational repeatability, reproducibility, and experimental history tracking [162, 163]. Through D4Science, the users can also share the computational provenance, compare and merge different results, and collaborate during the experimentation [164]. All these features helped us meet Open Science requirements of software reusability across different application domains and enhanced the reproducibility and repeatability of the experiments [28].

The workflow Web interface requires uploading a ZIP-compressed file containing a set of raster files, in the ESRI-GRID (ASCII) format, each representing the distribution of an environmental variable associated with the species’ ecological niche in a focus area. These files can come from open repositories of geospatial data (e.g. Copernicus, EMODnet, or others [80]). The files should all be at the same spatial resolution. As an additional input, our workflow requires providing a list of observations for one species as a CSV file (with *scientific name*, *longitude*, *latitude* columns) that it will enrich with environmental variable values to train the models. All files should be uploaded on the D4Science platform-integrated distributed storage sys-

tem [159]. The workflow can be executed either through a WPS-HTTP (POST/GET) call [158, 160] or the online Web interface. As the output, the workflow produces one ZIP-compressed file containing (i) all ENM distributions as ESRI-GRID (ASCII) files, (ii) their metadata as plain text files, (iii) the trained models as binary R files, and (iv) the ensemble model as an ESRI-GRID (ASCII) file. The biodiversity index distribution can be obtained as an additional ESRI-GRID (ASCII) file by running the workflow script to assemble several ensemble models. The script should be executed with only the biodiversity index construction algorithm activated in the workflow configuration file.

The open-source R software allows customising all workflow variables through the configuration file, e.g. all variables' ranges used during model optimisation, the number of background points to sample or the alternative absence locations to use, and the minimum ensemble-agreement threshold for the biodiversity index ("Supplementary information").

Hosting our workflow on the DataMiner allowed distributing the executions for multiple species on a cloud of 15 machines equipped with Ubuntu 18.04.5 LTS x86 64 operating system, 16 virtual cores, 32 GB of Random Access Memory, and 100 GB of disc for each machine. Each machine managed up to 4 executions simultaneously (i.e. $15 \times 4 = 60$ concurrent executions). This integration allows processing the data of many species concurrently because the DataMiner automatically distributes the single-species requests across the machines while balancing the computational load. Eventually, the workflow can assemble the cloud computation results within one biodiversity index. This way, it took ~ 5 hours to produce a biodiversity index for 1508 European marine species with full-automatic model optimisation (Sect. 3.2), instead of the ~ 7 days required by sequential processing.

2.6 Evaluation methodology

We evaluated our workflow with four different case studies. First, we evaluated the change in the individual and ensemble models' results across repeated executions (Sect. 3.1). We selected all species for which the AquaMaps [105] and FishBase [165] open repositories presented observations and maps in European marine basins (1508 species, including fishes and non-fishes) with local-to-widespread distributions. We estimated the stability of the ENMs, ensembles, and biodiversity index on this large species set, in terms of the sensitivity of their convergence to the same solution after model initialisation. The ENMs parametrisation, internal topologies, and results indeed depend on (i) the model initialisation (e.g. ANN weight initialisation, background point selection, etc.), (ii) the quality of the observation and environmental data, and (iii) the quantity and density of the species observations. Our first case study assessed how much these factors

influence the individual ENMs and whether the ensemble and biodiversity index models mitigate variability issues. To assess model stability, we repeated 10 executions and evaluated the per cent number of species distributions remaining almost stable, i.e. producing the same binary assessments for at least 60% of the grid pixels.

As a second case study, we produced a biodiversity index for the Mediterranean Sea based on the 1508 selected European species to demonstrate the capacity of our workflow to process large sets of species data through cloud computing (Sect. 3.2). We integrated expert-reviewed AquaMaps [81] for this assessment to improve output reliability and demonstrate the integrability of externally provided model outputs.

As a third case study, we studied the invasion of the Mediterranean Sea by *Siganus rivulatus*, a Lessepsian species currently invading the basin (Sect. 3.3). The distribution of this species often overlaps with the one of the native *Sarpa salpa*. The two species can coexist but *S. rivulatus* tends to consume the habitat resources with consequent risks for *S. salpa* survival [166]. To estimate habitat distributions in the Mediterranean, we used a native-adjusted model for *S. salpa* and a potential distribution model for *S. rivulatus*. The training-set locations were taken from their respective native areas. We studied the accuracy of our individual and ensemble ENMs at predicting the current observation records of *S. rivulatus* reported in OBIS in the last ten years (153 observations), and the potential change of habitat distribution in 2050 and 2100 under the RCP4.5 and RCP8.5 scenarios [78]. We also calculated the per cent overlap (as the fraction of shared high-suitability locations) between the *S. salpa* and *S. rivulatus* distributions to estimate overlap change over time and whether climate change will favour it. In this case study, we did not use AquaMaps because expert-curated data were unavailable for *S. rivulatus* across the RCP scenarios.

As a final case study, we analysed the correspondence between expert studies and the estimated ensemble distributions of *S. rivulatus* to assess the overall reliability of the identified high habitat suitability locations (Sect. 3.4).

All ENMs used the same set of environmental variables at a 0.5° resolution over the Mediterranean Sea [80], i.e.:

1. Sea-bottom and sea-surface temperature ($^\circ\text{C}$)
2. Distance from land (km)
3. Maximum, mean, minimum depth (m)
4. Net primary production ($\text{mgC m}^{-3} \text{day}^{-1}$)
5. Sea ice concentration (0–1 fraction)
6. Sea-bottom and sea-surface practical salinity (PSU)
7. Sea-bottom moles of oxygen per unit of mass ($\mu\text{mol/kg}$)

We used their publicly available projections in 2050 and 2100 [80] under the RCP4.5 and RCP8.5 future scenarios for case study 3. These variables are those also used by the official expert-revised AquaMaps distributions, which

the AquaMaps Consortium considers as containing sufficient information for general habitat suitability assessments [167].

3 Results

3.1 Sensitivity analysis

Table 1 reports the average per cent number of species distributions remaining almost stable (*matching* percentage) across several executions of (i) the integrated ENMs, (ii) the ensemble model, and (iii) the biodiversity index. This table assesses our workflow's sensitivity to model initialisation and background point selection: 66% ANN distributions remained stable, whereas the other ENMs presented much higher stability (between 90 and 100%). The lower ANN stability was probably due to its over-sensitivity to using background points as a proxy for habitat unsuitability [33, 60, 168]. Since AquaMaps is independent of initialisation and has no model parameters to optimise (Sect. 2.2.1), it reached a 100% matching percentage.

It is worth noting that although the ensemble model was obtained from ENMs with 66-to-100% matching percentages, it reached a 96% matching percentage. This result demonstrates the capacity of the ensemble model to compensate for instability by leveraging model complementary aspects, in agreement with other studies on ENM combinations [169–171].

The biodiversity index further improved this stability, reaching a 100% matching, i.e. it was independent of ENM initialisation and background point selection.

3.2 Mediterranean biodiversity index

The biodiversity index calculated from the ensemble distributions of 1508 marine species represented a species richness overview of the Mediterranean Sea (Fig. 2). Coasts presented a higher index than the basin's centre, in agreement with other studies [172]. Since the biodiversity index depends on species

distribution models rather than on species-observation density, this result is unlikely subject to observation-sampling biases.

The highest index values were in the western Mediterranean and decreased eastwards, as also highlighted by other studies [173, 174]. This gradient likely depends on the extensive range of physicochemical water conditions suitable for many organisms in the western region and the influx of Atlantic species [174]. In the rest of the basin, higher index values were mainly present in the Adriatic Sea, the Strait of Sicily, and the Aegean Sea, which agreed with other assessments [174–176]. The Strait of Sicily is indeed a crucial biodiversity hot spot because of its border location between the Mediterranean eastern and western sides [176]. As for the Adriatic, although it has areas with freshwater presence having less biodiversity richness [174], it is indeed an overall biodiversity hot spot [177, 178]. Finally, the Eastern Mediterranean showed lower levels of species richness in our map, still in agreement with other studies [174, 179].

3.3 Quantitative evaluation of species invasion prediction

The ensemble model distribution reached a high accuracy across all scenarios (80%) at predicting the presence locations of *Siganus rivulatus* (Fig. 3). This result indicates that the ensemble model valuably reused the complementary output information of the individual ENMs. Indeed, SVMs and ANNs underestimated the presence locations (40% and 45% accuracy, respectively), whereas MaxEnt (80% accuracy) strongly contributed to improving the ensemble model accuracy. The gap between the models' accuracy persisted across the RCP scenarios. The ensemble model covered increasing pixels across the years and RCP scenarios (with a slight reduction in 2050 under RCP8.5). This observation suggests that climate change in the worst scenario (RCP8.5) will likely favour the invasion. The ensemble model could not predict some pixels in the western Mediterranean. Rather, it predicted that the habitat will remain unsuitable far from the coasts in this area.

The ensemble models also predicted that the overlap between the ensemble distributions of *S. rivulatus* and *S. salpa* will progressively increase from 2019 to 2100 (Fig. 4). In the RCP8.5 scenario, there would be an overlap increment of 6.7% in ~30 years (from 70 to 75.3%) and of 25.6% in ~80 years (from 70 to 78.9%) with respect to 2019. Instead, in the RCP4.5 scenario, there would be an increment of 7.6% in ~30 years (from 70 to 74.4%) and 12.7% in ~80 years (from 70 to 87.9%). This result confirms the prediction of a long-term fostering of the invasion by climate change, and comparable effects in the shorter term although greenhouse gas emission mitigation [35, 180].

Table 1 Average number of matching species distributions after 10 consecutive workflow executions on 1508 European species data, across the integrated ecological niche models, their ensemble model, and the biodiversity index

	Matching distributions (%)
Artificial neural networks	66
Support vector machines	90
Maximum entropy	94
AquaMaps	100
Ensemble model	96
Biodiversity index	100

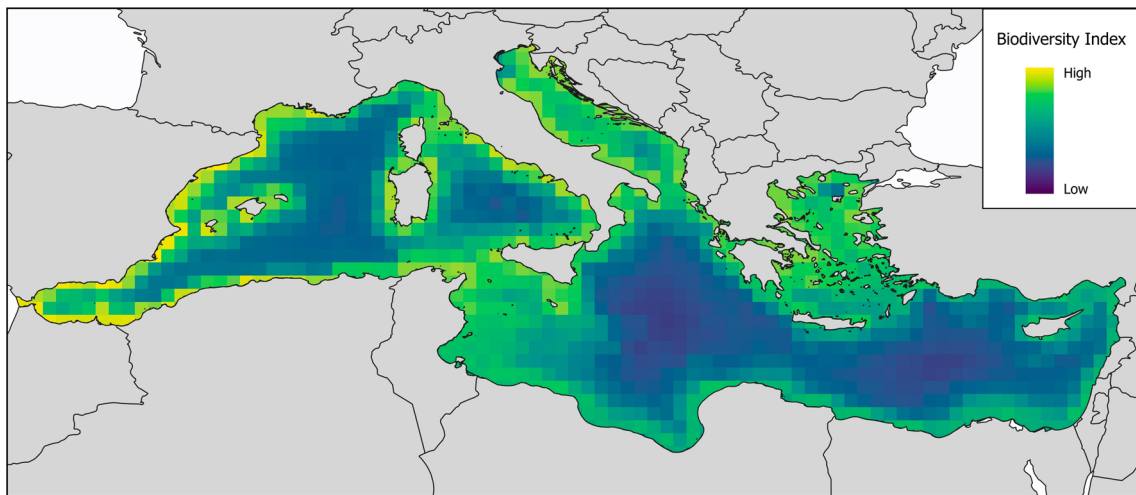


Fig. 2 Biodiversity index (species richness) at half-degree resolution, produced by our workflow after processing 1508 Mediterranean species data

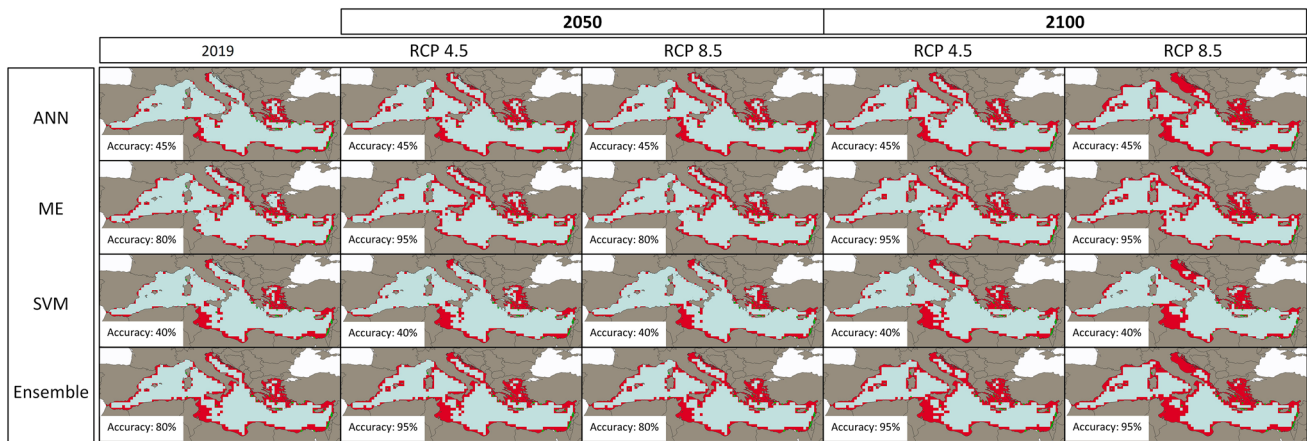


Fig. 3 Accuracy comparison between three individual ecological niche models integrated with our workflow—i.e. Artificial Neural Networks (ANN), Maximum Entropy (ME), and Support Vector Machines (SVM), and their Ensemble model—in the prediction of the *Siganus*

rivulatus distribution in the Mediterranean Sea in 2019. Projections for 2050 and 2100 are reported for the RCP4.5 and RCP8.5 scenarios. Small green dots report the *S. rivulatus* observations from OBIS

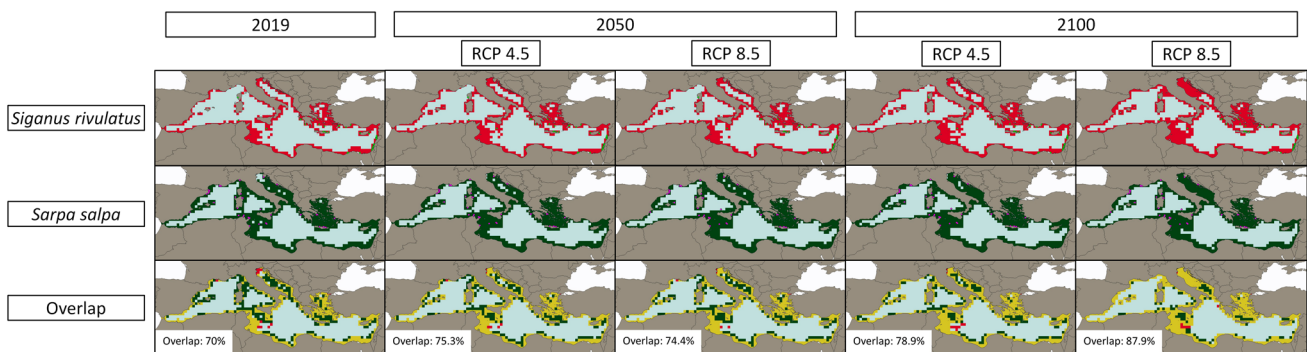


Fig. 4 Per cent overlap between the estimated distributions of *Siganus rivulatus* and *Sarpa salpa* in 2019, 2050 and 2100. Future projections are reported for the RCP4.5 and RCP8.5 scenarios. Small green dots report the *S. rivulatus* observations from OBIS, and purple dots those of *S. salpa*

3.4 Qualitative evaluation of species invasion prediction

Our ensemble model identified high-suitability areas for *S. rivulatus* confirmed by expert studies on the eastern Mediterranean Sea, e.g. off the coasts of Turkey [181], Egypt [182], Cyprus [183, 184], Crete [184], and Israel [185, 186]. Stable species presence has been reported in other locations also predicted by our model, e.g. off the Albanian [187] and eastern Greek coasts [188]. In the central Mediterranean Sea, the northernmost observation from expert studies is near the Bobara island (south-eastern Adriatic) [189], and the westernmost is near the Pelagie Islands (Strait of Sicily) [190]. Our ensemble model also predicted habitat suitability in these locations (Fig. 3). Additionally, in the Strait of Sicily it foresaw a particular increase of habitat suitability over the years.

S. rivulatus has been rarely reported in the western Mediterranean Sea. Presence off the French coasts could only be inferred through eDNA analysis in ports [191] and has been unofficially reported by fishers [192]. One observation off the western Corsican coast was just indirectly inferred from a picture [193]. These considerations, and the low observation frequency, might confirm a low habitat suitability in this area in agreement with our model's suggestions.

Our model indicated low habitat suitability also far from the northern Tunisian coasts. However, two expert observations are available in this area from a transect report [188]. Nevertheless, also *Siganus luridus* might be present in the area [194], which can be mistaken for *S. rivulatus* given their similarity. The *S. luridus* distribution often overlaps with that of *S. salpa* [193], and the species outcompetes *S. rivulatus* when present in the same area [195]. In a similar case, in Malta, the *S. rivulatus* observations were indeed re-classified as *S. luridus* observations after expert verification [196].

4 Conclusions

We have presented an automatic workflow to produce potential and actual species distributions over a marine area, through four integrated ENMs with complementary aspects. The workflow combines the ENMs to produce one overall ensemble model, which is more stable and accurate than the individual models. The ensemble model has a lower sensitivity to initialisation and background point selection and a higher predictive accuracy than the individual ENMs.

The workflow allowed for predicting the invasion of the Mediterranean Sea by an invasive Lessepsian species (*S. rivulatus*) and its current and future distribution overlap with a native competitor species (*S. salpa*). The ensemble model was particularly reliable at predicting known presence locations of the invasive species in the Mediterranean with a large agreement with expert studies. The invasion assessment was

also projected in future (2050 and 2100) medium and high greenhouse gas emission scenarios (RCP4.5 and RCP8.5). The projections highlighted that climate change will likely foster the invasion of the basin by *S. rivulatus* and increase its distribution overlap with *S. salpa*, especially in the RCP8.5 scenario. Therefore, climate change would increase the risk of *S. salpa* habitat loss and fisheries change in the Mediterranean.

We have also demonstrated the possibility of easily producing a biodiversity index for many species through cloud computing, which was (i) independent of the individual ENMs' initialisation, (ii) more stable than the species' ensemble models, and (iii) in agreement with expert studies.

Our workflow is general enough to process the data of other areas, species, and scenarios than those presented in the case studies. Moreover, it can integrate the outputs of additional ENMs. Its Open Science compliance makes it easily integrable with GIS software and improves communication transparency towards result stakeholders (e.g. ecological and ecosystem modellers and policymakers) [9–11, 29–31]. This compliance, combined with the use of general models, allows for reusing the workflow in other domains than marine science. For example, it might be used for terrestrial species (using AquaMaps as a pure envelope function estimator or disabling it, in this case) [19, 66, 111]. We also plan to extend the workflow with new general ENMs, while distinguishing data-poor and data-rich scenarios to optimise model effectiveness and stability.

One essential aim of our workflow is to extend the use of ENMs to heterogeneous communities of practice, even with low expertise in ecological niche modelling. Open Science compliance and full automatism support this goal. Drawbacks are the default integration of a small (but representative) number of ENMs using complementary approaches and the use of pre-defined model optimisation strategies as the default configuration. However, our workflow is quickly integrable with other software using alternative ENMs and evaluation strategies. Additionally, our workflow customisation allows for producing single-species models with high accuracy because expert modellers can easily modify all parameters and processing steps. Specifically, the workflow configuration allows for modifying many model parameters (such as the ANN layers, the MaxEnt parameters, and the SVM kernels), and the modular open-source architecture of the software allows for quickly changing specific scripts that implement precise workflow steps. For example, model combination weights in species richness estimation can easily be changed to assign more importance to specific models.

Our workflow should be contextualised within big data processing approaches. In our case studies, the number of species and the study area were extensive but the models' training sets were relatively small. The results should be

considered as the bases of more complex analyses. They indicated macro-patterns of species richness and habitat change rather than cell- and species-wise detailed answers. Although an erroneous or missing indication of one species' presence in a cell is not essential in macroscopic trend analyses [83, 170, 171], caution should be used when interpreting the results of our workflow over many species because punctually reliable indications cannot be guaranteed.

The full automatism makes our workflow suitable for supporting IEAs in the automatic discovery of the mutual relations between the different driving forces stressing an ecosystem [3, 5, 197]. This is a crucial focus also of modern designs for Digital Twins of the Ocean (DTOs), which aim to produce digital representations of oceanic ecosystems that use real-time and historical data to assess and forecast ecosystem status [27]. These models are attracting scientific interest but require new automatic solutions (possibly AI-based) for modelling and discovering ecosystem functions and assessing habitat suitability. Our workflow meets these scopes, and we plan to propose its use in the DTO context.

Supplementary information

The source code and all experiments' input and output are available on the GitHub at <https://github.com/cybprojects65/EcologicalNicheModellingWithR>

The software was tested with R 4.2.3 on the MS Windows and Linux Operating Systems.

The GitHub repository contains all evaluation experiments' input and output and the conversion tools used. It also includes the list of 1508 species used in our biodiversity index for the Mediterranean Sea (https://github.com/cybprojects65/EcologicalNicheModellingWithR/blob/main/List_of_1508_EU_species_from_AquaMaps_and_FishBase.txt)

The Web service interface and WPS access point is available on the D4Science e-Infrastructure (<https://services.d4science.org/>).

Subscription to the (free-to-use) BiodiversityLab Virtual Research Environment is required to properly size the computational resources to the users' request load (<https://services.d4science.org/group/d4science-services-gateway/explore>).

After subscription, the ENM Workflow Web interface will be freely accessible at https://services.d4science.org/group/biodiversitylab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducers.ECOLOGICAL_NICHE_MODELLER

Example input datasets are available at URL: data.d4science.net/cGHy and URL: data.d4science.net/MQTR, as indicated in the Web interface.

No fee is required to use the service.

Acknowledgements This work has been conducted in the context of the EcoScope European Project (H2020-EU.3.2.3.1—Grant Agreement No. 101000302) and the ITINERIS PNRR Italian project (project code No. IR0000032—ESFRI Environment).

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement.

Declarations

Competing Interest The authors declare no competing interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Froese, R., Winker, H., Coro, G., Demirel, N., Tsikliras, A.C., Dimarchopoulou, D., Scarcella, G., Quaas, M., Matz-Lück, N.: Status and rebuilding of European fisheries. *Mar. Policy* **93**, 159–170 (2018)
2. Espinosa, F., Bazairi, H.: *Coastal Habitat Conservation*, pp. 1–16. Elsevier, Amsterdam pp (2023)
3. Antunes, P., Santos, R.: Integrated environmental management of the oceans. *Ecol. Econ.* **31**, 215–226 (1999)
4. Christensen, V., Walters, C.J.: Ecopath with ecosim: methods, capabilities and limitations. *Ecol. Model.* **172**(2–4), 109–139 (2004)
5. Kristensen, P.: The DPSIR framework, European topic centre on water. *Eur. Environ. Agency* pp. 1–10 (2004)
6. Coll, M., Bundy, A., Shannon, L.J.: Ecosystem modelling using the ecopath with ecosim approach. In: *Computers in Fisheries Research*, pp. 225–291. https://doi.org/10.1007/978-1-4020-8636-6_8(2009)
7. Colléter, M., Valls, A., Guitton, J., Gascuel, D., Pauly, D., Christensen, V.: Global overview of the applications of the ecopath with ecosim modeling approach using the ecobase models repository. *Ecol. Model.* **302**, 42–53 (2015)
8. Heymans, J.J., Coll, M., Link, J.S., Mackinson, S., Steenbeek, J., Walters, C., Christensen, V.: Best practice in ecopath with ecosim food-web models for ecosystem-based management. *Ecol. Model.* **331**, 173–184 (2016)
9. Gari, S.R., Newton, A., Icely, J.D.: A review of the application and evolution of the DPSIR framework with an emphasis on coastal social–ecological systems. *Ocean Coast Manag* **103**, 63–77 (2015)
10. Taconet, P., Chassot, E., Guitton, J., Fiorellato, F., Anello, E., Barde, J.: Data toolbox for fisheries: the case of tuna fisheries (2016). Accessible online at https://www.iotc.org/sites/default/files/documents/2018/04/IOTC-2016-WPDCS-12-2_-TUNA_DATA_TOOLBOX.pdf

11. James, M., Mendo, T., Jones, E.L., Orr, K., McKnight, A., Thompson, J.: AIS data to inform small scale fisheries management and marine spatial planning. *Mar. Policy* **91**, 113–121 (2018)
12. Kearney, K.A.: *ecopath_matlab*: a matlab-based implementation of the ecopath food web algorithm. *J Open Source Softw* **2**(9), 64 (2017)
13. Wu, P.P.Y., Mengersen, K., McMahon, K., Kendrick, G.A., Chartrand, K., York, P.H., Rasheed, M.A., Caley, M.J.: Timing anthropogenic stressors to mitigate their impact on marine ecosystem resilience. *Nat. Commun.* **8**(1), 1263 (2017)
14. Hu, J.H., Tsai, W.P., Cheng, S.T., Chang, F.J.: Explore the relationship between fish community and environmental factors by machine learning techniques. *Environ. Res.* **184**, 109262 (2020)
15. Park, S., Sin, Y.: Artificial neural network (ANN) modeling analysis of algal blooms in an estuary with episodic and anthropogenic freshwater inputs. *Appl. Sci.* **11**(15), 6921 (2021)
16. Niu, L., Xiao, L.: Ecological environment management system based on artificial intelligence and complex numerical optimization. *Microprocess. Microsyst.* **80**, 103627 (2021)
17. Satir, O., Berberoglu, S., Donmez, C.: Mapping regional forest fire probability using artificial neural network model in a mediterranean forest ecosystem. *Geomat. Nat. Haz. Risk* **7**(5), 1645–1658 (2016)
18. Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T.: Importance of spatial predictor variable selection in machine learning applications—moving from data reproduction to spatial prediction. *Ecol. Model.* **411**, 108815 (2019)
19. Coro, G., Trumpy, E.: Predicting geographical suitability of geothermal power plants. *J. Clean. Prod.* **267**, 121874 (2020)
20. Ghareghan, F., Ghanbarian, G., Pourghasemi, H.R., Safaeian, R.: Prediction of habitat suitability of *Morina persica* l. species using artificial intelligence techniques. *Ecol. Ind.* **112**, 106096 (2020)
21. Zhang, Z., Mammola, S., Liang, Z., Capinha, C., Wei, Q., Wu, Y., Zhou, J., Wang, C.: Future climate change will severely reduce habitat suitability of the critically endangered Chinese giant salamander. *Freshw. Biol.* **65**(5), 971–980 (2020)
22. Coro, G., Bove, P., Ellenbroek, A.: Habitat distribution change of commercial species in the adriatic sea during the covid-19 pandemic. *Eco. Inf.* **69**, 101675 (2022)
23. Paine, D.R., Sheppard, A.W., Cook, D.C., De Barro, P.J., Worner, S.P., Thomas, M.B.: Global threat to agriculture from invasive species. *Proc. Natl. Acad. Sci.* **113**(27), 7575–7579 (2016)
24. Coro, G., Vilas, L.G., Magliozzi, C., Ellenbroek, A., Scarponi, P., Pagano, P.: Forecasting the ongoing invasion of *Lagocephalus sceleratus* in the mediterranean sea. *Ecol. Model.* **371**, 37–49 (2018)
25. Morisette, J.T., Reaser, J.K., Cook, G.L., Irvine, K.M., Roy, H.E.: Right place. right time. right tool: guidance for using target analysis to increase the likelihood of invasive species detection. *Biol. Invasions* **22**(1), 67–74 (2020)
26. Martinez, B., Reaser, J.K., Dehgan, A., Zamft, B., Baisch, D., McCormick, C., Giordano, A.J., Aicher, R., Selbe, S.: Technology innovation: advancing capacities for the early detection of and rapid response to invasive species. *Biol. Invasions* **22**(1), 75–100 (2020)
27. Campana, E.F., Ciappi, E., Coro, G.: The role of technology and digital innovation in sustainability and decarbonization of the blue economy. *Bull. Geophys. Oceanogr.* **3**, 123–130 (2021)
28. Hey, A.J., Tansley, S., Tolle, K.M., et al.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1. Microsoft Research Redmond, Washington (2009)
29. Jennings, S., Lee, J.: Defining fishing grounds with vessel monitoring system data. *ICES J. Mar. Sci.* **69**, 51–63 (2012)
30. Dunn, D.C., Jablonicky, C., Crespo, G.O., McCauley, D.J., Kroodsmas, D.A., Boerder, K., Gjerde, K.M., Halpin, P.N.: Empowering high seas governance with satellite vessel tracking data. *Fish Fish.* **19**, 729–739 (2018)
31. Song, A.M., Johnsen, J.P., Morrison, T.H.: Reconstructing governability: how fisheries are made governable. *Fish Fish.* **19**, 377–389 (2018)
32. Coro, G.: Open science and artificial intelligence supporting blue growth. *Environ. Eng. Manag. J. (EEMJ)* **19**(10), 1719–1729 (2020)
33. Pearson, R.G.: Species’ distribution modeling for conservation educators and practitioners. *Synth. Am. Mus. Nat. Hist.* **50**, 54–89 (2007)
34. Jones, M.C., Dye, S.R., Pinnegar, J.K., Warren, R., Cheung, W.W.: Modelling commercial fish distributions: prediction and assessment using different approaches. *Ecol. Model.* **225**, 133–145 (2012)
35. Coro, G., Magliozzi, C., Ellenbroek, A., Kaschner, K., Pagano, P.: Automatic classification of climate change effects on marine species distributions in 2050 using the AquaMaps model. *Environ. Ecol. Stat.* **23**, 155–180 (2016)
36. Weber, M.M., Stevens, R.D., Diniz-Filho, J.A.F., Grelle, C.E.V.: Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. *Ecography* **40**(7), 817–828 (2017)
37. Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., Joly, A.: Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Comput. Biol.* **17**(4), e1008856 (2021)
38. Hirzel, A.H., Le Lay, G.: Habitat suitability modelling and niche theory. *J. Appl. Ecol.* **45**(5), 1372–1381 (2008)
39. Elith, J., Leathwick, J.R.: Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **40**(1), 677–697 (2009). <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
40. Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., Peterson, A.T., Soberón, J., Pearson, R.G.: *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton University Press, Princeton (2011)
41. Sánchez-Tapia, A., de Siqueira, M.F., Lima, R.O., Barros, F.S.M., Gall, G.M., Gadelha, L.M., da Silva, L.A.E., Osthoff, C.: In: *Latin American High Performance Computing Conference*, Springer, Berlin, pp. 218–232 (2017)
42. Guo, Q., Liu, Y.: ModEco: an integrated software package for ecological niche modeling. *Ecography* **33**(4), 637–642 (2010)
43. Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M., Anderson, R.P.: ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models. *Methods Ecol. Evol.* **5**(11), 1198–1205 (2014)
44. Magliozzi, C., Coro, G., Grabowski, R.C., Packman, A.I., Krause, S.: A multiscale statistical method to identify potential areas of hyporheic exchange for river restoration planning. *Environ. Model. Softw.* **111**, 311–323 (2019)
45. Schnase, J.L., Carroll, M.L., Gill, R.L., Tamkin, G.S., Li, J., Strong, S.L., Maxwell, T.P., Aronne, M.E., Spradlin, C.S.: Toward a Monte Carlo approach to selecting climate variables in maxent. *PLoS ONE* **16**(3), e0237208 (2021)
46. Warren, D.L., Seifert, S.N.: Ecological niche modeling in maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21**(2), 335–342 (2011)
47. Coro, G., Pagano, P., Ellenbroek, A.: *International Conference on Adaptive and Natural Computing Algorithms*, pp. 346–355. Springer, Berlin (2013)
48. Coro, G., Magliozzi, C., Ellenbroek, A., Pagano, P.: Improving data quality to build a robust distribution model for *Architeuthis dux*. *Ecol. Model.* **305**, 29–39 (2015)

49. Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L.: Parallelizing the execution of native data mining algorithms for computational biology. *Concurr. Comput.: Pract. Exp.* **27**(17), 4630–4644 (2015)
50. Zeng, Y., Low, B.W., Yeo, D.C.: Novel methods to select environmental variables in maxent: a case study using invasive crayfish. *Ecol. Model.* **341**, 5–13 (2016)
51. Bargain, A., Marchese, F., Savini, A., Taviani, M., Fabri, M.C.: Santa Maria di Leuca province (Mediterranean Sea): identification of suitable mounds for cold–water coral settlement using geomorphometric proxies and maxent methods. *Front. Mar. Sci.* **4**, 338 (2017)
52. Durand, M., Fu, L.L., Lettenmaier, D.P., Alsdorf, D.E., Rodriguez, E., Esteban-Fernandez, D.: The surface water and ocean topography mission: observing terrestrial surface water and oceanic submesoscale eddies. *Proc. IEEE* **98**(5), 766–779 (2010)
53. Werdell, P.J., Bailey, S.W.: An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sens. Environ.* **98**(1), 122–140 (2005)
54. Alvera-Azcárate, A., Barth, A., Rixen, M., Beckers, J.M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the adriatic sea surface temperature. *Ocean Model.* **9**(4), 325–346 (2005)
55. Peterson, A.T.: Predicting species' geographic distributions based on ecological niche modeling. *The Condor* **103**(3), 599–605 (2001). <https://doi.org/10.1093/condor/103.3.599>
56. Huang, Y.P., Kao, L.J., Sandnes, F.E.: Efficient mining of salinity and temperature association rules from Argo data. *Expert Syst. Appl.* **35**(1–2), 59–68 (2008)
57. Ravdas, M., Zacharioudaki, A., Korres, G.: Implementation and validation of a new operational wave forecasting system of the Mediterranean monitoring and forecasting centre in the framework of the copernicus marine environment monitoring service. *Nat. Hazard.* **18**(10), 2675–2695 (2018)
58. Scarponi, P., Coro, G., Pagano, P.: A collection of Aquamaps native layers in NetCDF format. *Data Brief* **17**, 292–296 (2018)
59. Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J.: A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**(1), 43–57 (2011). <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
60. Coro, G., Magliozzi, C., Berghe, E.V., Bailly, N., Ellenbroek, A., Pagano, P.: Estimating absence locations of marine species from data of scientific surveys in obis. *Ecol. Model.* **323**, 61–76 (2016)
61. Peterson, T., Papeş, M., Eaton, M.: Transferability and model evaluation in ecological niche modeling: a comparison of GARP and maxent. *Ecography* **30**(4), 550–560 (2007)
62. Ben Rais Lasram, F., Guilhaumon, F., Albouy, C., Somot, S., Thuiller, W., Mouillot, D.: The Mediterranean sea as a 'cul-de-sac' for endemic fishes facing climate change. *Glob. Change Biol.* **16**(12), 3233–3245 (2010)
63. Friedlaender, A.S., Johnston, D.W., Fraser, W.R., Burns, J., Costa, D.P., et al.: Ecological niche modeling of sympatric krill predators around Marguerite Bay, Western Antarctic Peninsula. *Deep Sea Res. Part II* **58**(13–16), 1729–1740 (2011)
64. Ashraf, U., Peterson, A.T., Chaudhry, M.N., Ashraf, I., Saqib, Z., Rashid Ahmad, S., Ali, H.: Ecological niche model comparison under different climate scenarios: a case study of *Olea* spp. in Asia. *Ecosphere* **8**(5), e01825 (2017)
65. Chala, D., Roos, C., Svenning, J.C., Zinner, D.: Species-specific effects of climate change on the distribution of suitable baboon habitats-ecological niche modeling of current and last glacial maximum conditions. *J. Hum. Evol.* **132**, 215–226 (2019)
66. Coro, G.: A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate. *Ecol. Model.* **431**, 109187 (2020)
67. Peterson, A.T.: Predicting the geography of species' invasions via ecological niche modeling. *Q. Rev. Biol.* **78**(4), 419–433 (2003)
68. Menchetti, M., Guéguen, M., Talavera, G.: Spatio-temporal ecological niche modelling of multigenerational insect migrations. *Proc. R. Soc. B* **286**(1910), 20191583 (2019)
69. Araujo, M.B., Naimi, B.: Spread of SARS CoV-2 coronavirus likely to be constrained by climate (2020, preprint). <https://doi.org/10.1101/2020.03.12.20034728>
70. de Siqueira, M.F., Durigan, G., de Marco Júnior, P., Peterson, A.T.: Something from nothing: using landscape similarity and ecological niche modeling to find rare plant species. *J. Nat. Conserv.* **17**(1), 25–32 (2009)
71. Chunco, A.J., Phimmachak, S., Sivongxay, N., Stuart, B.L.: Predicting environmental suitability for a rare and threatened species (*Lao Newt, Laotriton laoensis*) using validated species distribution models. *PLoS ONE* **8**(3), e59853 (2013)
72. Blue Cloud Consortium. The Blue Cloud European project—e-Infrastructures (2023). Available at <https://blue-cloud.org/e-infrastructures>
73. EcoScope. The EcoScope European Project (2023). Available online at <https://ecoscopium.eu/>
74. Koleff, P., Gaston, K.J., Lennon, J.J.: Measuring beta diversity for presence-absence data. *J. Anim. Ecol.* **72**(3), 367–382 (2003)
75. Kreft, H., Jetz, W.: Global patterns and determinants of vascular plant diversity. *Proc. Natl. Acad. Sci.* **104**(14), 5925–5930 (2007)
76. Ferrier, S., Guisan, A.: Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* **43**(3), 393–404 (2006)
77. Sillero, N., Campos, J.C., Arenas-Castro, S., Barbosa, A.M.: A curated list of R packages for ecological niche modelling. *Ecol. Model.* **476**, 110242 (2023)
78. Van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.F., et al.: The representative concentration pathways: an overview. *Clim. Change* **109**, 5–31 (2011)
79. Nazarenko, L.S., Tausnev, N., Russell, G.L., Rind, D., Miller, R.L., Schmidt, G.A., Bauer, S.E., Kelley, M., Ruedy, R., Ackerman, A.S., et al.: Future climate change under SSP emission scenarios with GISS-E2.1. *J. Adv. Model. Earth Syst.* **14**(7), e2021MS002871 (2022)
80. Coro, G., Bove, P., Kesner-Reyes, K.: Global-scale parameters for ecological models. *Sci. Data* **10**(1), 7 (2023)
81. AquaMaps.org. The AquaMaps Web site (2022). Available at www.aquamaps.org
82. Coro, G., Bove, P., Ellenbroek, A.: Habitat distribution change of commercial species in the Adriatic Sea during the COVID-19 pandemic. *Ecol. Inf.* **69**, 101675 (2022). <https://doi.org/10.1016/j.ecoinf.2022.101675>
83. Queiroz, N., Humphries, N.E., Couto, A., Vedor, M., Da Costa, I., Sequeira, A.M., Mucientes, G., Santos, A.M., Abascal, F.J., Abercrombie, D.L., et al.: Reply to: Caution over the use of ecological big data for conservation. *Nature* **595**(7866), E20–E28 (2021)
84. Copernicus. Copernicus-Marine environment monitoring service (2020)
85. EMODNET. EMODnet Bathymetry (2020). Available at <https://emodnet.ec.europa.eu/en/bathymetry>
86. Grassele, J.F.: The ocean biogeographic information system (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. *Oceanography* **13**, 5–7 (2000)
87. GBIF Home Page. GBIF.org (2023). Available at <https://www.gbif.org>
88. Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**(3–4), 231–259 (2006)
89. Ferrier, S., Watson, G.: An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the

- Distribution of Biological Diversity. Environment Australia, Canberra (1997)
90. Ferrier, S., Watson, G., Pearce, J., Drielsma, M.: Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. I. Species-level modelling. *Biodiv. Conserv.* **11**, 2275–2307 (2002)
 91. Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T.: New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* **19**(9), 497–503 (2004)
 92. Aspinall, P.J.: The future of ethnicity classifications. *J. Ethnic Migr. Stud.* **35**(9), 1417–1435 (2009)
 93. Elith, J., Kearney, M., Phillips, S.: The art of modelling range-shifting species. *Methods Ecol. Evol.* **1**(4), 330–342 (2010)
 94. Phillips, S.J., Dudík, M.: Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography* **31**(2), 161–175 (2008)
 95. Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J.: A statistical explanation of maxent for ecologists. *Divers. Distrib.* **17**(1), 43–57 (2011)
 96. Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C.G., Sousa-Guedes, D., Martínez-Freiría, F., Real, R., Barbosa, A.M.: Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. *Ecol. Model.* **456**, 109671 (2021)
 97. Patil, V.B., Sonnad, S., Lalita, Y.: User identification using HTK. *Int. J. Eng. Res.* (2014). <https://doi.org/10.17577/IJERTV3IS060721>
 98. Coro, G., Bardelli, S., Cuttano, A., Scaramuzzo, R.T., Ciantelli, M.: A self-training automatic infant-cry detector. *Neural Comput. Appl.* **35**(11), 8543–8559 (2023)
 99. MacLeod, C.D.: Habitat representativeness score (HRS): a novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. *J. Mar. Biol. Assoc. U.K.* **90**(7), 1269–1277 (2010)
 100. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**(8), 913–929 (2017)
 101. Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., Huijbregts, M.A.: Assessing the reliability of species distribution projections in climate change research. *Divers. Distrib.* **27**(6), 1035–1050 (2021)
 102. Kaschner, K., Watson, R., Trites, A., Pauly, D.: Mapping worldwide distributions of marine mammal species using a relative environmental suitability (RES) model. *Mar. Ecol. Prog. Ser.* **316**, 285–310 (2006)
 103. Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R.: Predicting the distributions of marine organisms at the global scale. *Ecol. Model.* **221**(3), 467–478 (2010). <https://doi.org/10.1016/j.ecolmodel.2009.10.025>
 104. Booth, T.H., Nix, H.A., Busby, J.R., Hutchinson, M.F.: Bioclim: the first species distribution modelling package, its early applications and relevance to most current maxent studies. *Divers. Distrib.* **20**(1), 1–9 (2014)
 105. Reyes, K.: AquaMaps: Algorithm and Data Sources for Aquatic Organisms (2015). Available at http://www.aquamaps.org/main/FB_Book_KReyes_AquaMaps_JG.pdf
 106. Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**(3–4), 231–259 (2006)
 107. Phillips, S.J., Dudík, M., Schapire, R.E.: In: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, p. 83 (2004)
 108. Baldwin, R.A.: Use of maximum entropy modeling in wildlife research. *Entropy* **11**(4), 854–866 (2009)
 109. Capezzuto, F., Sion, L., Ancona, F., Carlucci, R., Carluccio, A., Cornacchia, L., Maiorano, P., Ricci, P., Tursi, A., D’Onghia, G.: Cold-water coral habitats and canyons as Essential Fish Habitats in the southern Adriatic and northern Ionian Sea (central Mediterranean). *Ecol. Quest.* **29**(3), 9–23 (2018)
 110. Angeletti, L., Prampolini, M., Fogliani, F., Grande, V., Taviani, M.: Seafloor Geomorphology as Benthic Habitat, pp. 811–824. Elsevier, Amsterdam (2020)
 111. Coro, G., Bove, P.: A high-resolution global-scale model for COVID-19 infection rate. *ACM Trans. Spat. Algorithms Syst. (TSAS)* **8**(3), 1–24 (2022)
 112. Coro, G., Tassetti, A.N., Armelloni, E.N., Pulcinella, J., Ferrà, C., Sprovieri, M., Trincardi, F., Scarcella, G.: COVID-19 lockdowns reveal the resilience of adriatic sea fisheries to forced fishing effort reduction. *Sci. Rep.* **12**(1), 1–14 (2022)
 113. Merow, C., Smith, M.J., Silander, J.A., Jr.: A practical guide to maxent for modeling species’ distributions: what it does, and why inputs and settings matter. *Ecography* **36**(10), 1058–1069 (2013)
 114. Elith, J., Graham, C.H.: Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32**(1), 66–77 (2009)
 115. Wang, L., Kerr, L.A., Record, N.R., Bridger, E., Tupper, B., Mills, K.E., Armstrong, E.M., Pershing, A.J.: Modeling marine pelagic fish species spatiotemporal distributions utilizing a maximum entropy approach. *Fish. Oceanogr.* **27**(6), 571–586 (2018)
 116. Renner, I.W., Warton, D.I.: Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69**(1), 274–281 (2013)
 117. Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E.: Opening the black box: an open-source release of maxent. *Ecography* **40**(7), 887–893 (2017)
 118. Zaniwsky, A.E., Lehmann, A., Overton, J.M.: Predicting species spatial distributions using presence-only data: a case study of native new Zealand ferns. *Ecol. Model.* **157**(2–3), 261–280 (2002)
 119. Dudík, M., Phillips, S., Schapire, R.E.: Correcting sample selection bias in maximum entropy density estimation. *Adv. Neural. Inf. Process. Syst.* **18**, 323–330 (2005)
 120. Hengl, T., Sierdsema, H., Radović, A., Dilo, A.: Spatial prediction of species’ distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecol. Model.* **220**(24), 3499–3511 (2009)
 121. Minsky, M.: Steps toward artificial intelligence. *Comput. Thought* **406**, 450 (1963)
 122. Lek, S., Guégan, J.F.: Artificial neural networks as a tool in ecological modelling, an introduction. *Ecol. Model.* **120**(2), 65–73 (1999)
 123. Olden, J.D., Joy, M.K., Death, R.G.: An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol. Model.* **178**(3), 389–397 (2004)
 124. Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S.: Application of neural networks to modelling non-linear relationships in ecology. *Ecol. Model.* **90**(1), 39–52 (1996)
 125. Gevrey, M., Dimopoulos, I., Lek, S.: Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* **160**(3), 249–264 (2003)
 126. Coro, G., Pagano, P., Ellenbroek, A.: Combining simulated expert knowledge with neural networks to produce ecological niche models for *Latimeria chalumnae*. *Ecol. Model.* **268**, 55–63 (2013)
 127. Patterson, D.W.: Artificial Neural Networks: Theory and Applications. Prentice Hall PTR, Hoboken (1998)
 128. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)

129. Bebis, G., Georgiopoulos, M.: Feed-forward neural networks. *IEEE Potent.* **13**(4), 27–31 (1994)
130. Rumelhart, D., Hinton, G., Williams, R.: Steps toward artificial intelligence. *Nature* **6088**(323), 533–536 (1986)
131. Özesmi, S.L., Tan, C.O., Özesmi, U.: Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Model.* **195**(1), 83–93 (2006)
132. Fritsch, S., Guenther, F., Guenther, M.F.: Package ‘neuralnet’. *The Comprehensive R Archive Network* (2016)
133. Coro, G., Massoli, F.V., Origlia, A., Cutugno, F.: Psycho-acoustics inspired automatic speech recognition. *Comput. Electr. Eng.* **93**, 107238 (2021)
134. Coro, G., Fortunati, L., Pagano, P.: In: 2013 MTS/IEEE OCEANS-Bergen, IEEE, pp. 1–5 (2013)
135. Paluzo-Hidalgo, E., Gonzalez-Diaz, R., Gutiérrez-Naranjo, M.A.: Two-hidden-layer feed-forward networks are universal approximators: a constructive approach. *Neural Netw.* **131**, 29–36 (2020)
136. Shen, Z., Yang, H., Zhang, S.: Neural network approximation: three hidden layers are enough. *Neural Netw.* **141**, 160–173 (2021)
137. Boser, B.E., Guyon, I.M., Vapnik, V.N.: In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, pp. 144–152 (1992)
138. Brown, M., Gunn, S.R., Lewis, H.G.: Support vector machines for optimal classification and spectral unmixing. *Ecol. Model.* **120**(2), 167–179 (1999)
139. Schölkopf, B., Burges, C.J., Smola, A.J.: *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge (1999)
140. Guo, Q., Kelly, M., Graham, C.H.: Support vector machines for predicting distribution of sudden oak death in California. *Ecol. Model.* **182**(1), 75–90 (2005)
141. Drake, J.M., Randin, C., Guisan, A.: Modelling ecological niches with support vector machines. *J. Appl. Ecol.* **43**(3), 424–432 (2006)
142. Vapnik, V.: *The Nature of Statistical Learning Theory*, vol. 1. Springer, Berlin (2013)
143. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge (2000)
144. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., Leisch, M.F.: The e1071 package. *Misc Functions of Department of Statistics (e1071)*, TU Wien pp. 297–304 (2006)
145. Coro, G., Webb, T.J., Appeltans, W., Bailly, N., Cattrijsse, A., Pagano, P.: Classifying degrees of species commonness: North Sea fish as a case study. *Ecol. Model.* **312**, 272–280 (2015)
146. Hill, J., Hamer, K., Lace, L., Banham, W.: Effects of selective logging on tropical forest butterflies on Buru, Indonesia. *J. Appl. Ecol.* **32**, 754–760 (1995)
147. Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V., Worm, B.: Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**(7310), 1098–1101 (2010)
148. Phillips, S.J., Miroslav, D., E, S.R.: Maxent software for modeling species niches and distributions (version 3.4.1) (2019). http://biodiversityinformatics.amnh.org/open_source/maxent/
149. Fritsch, J., Hilliker, J., Ross, J., Vislocky, R.: Model consensus. *Weather Forecast.* **15**(5), 571–582 (2000)
150. Czarnowski, I., Jędrzejowicz, P.: In: 2011 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, pp. 936–941 (2011)
151. Ala'raj, M., Abbod, M.F.: A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Syst. Appl.* **64**, 36–55 (2016)
152. Alhashmi, A.A., Darem, A.A.: Consensus-based ensemble model for Arabic cyberbullying detection. *Comput. Syst. Sci. Eng.* **41**(1), 241–254 (2022)
153. Alzubi, O.A., Alzubi, J.A.A., Tedmori, S., Rashaideh, H., Almomani, O.: Consensus-based combining method for classifier ensembles. *Int. Arab J. Inf. Technol.* **15**(1), 76–86 (2018)
154. Card, D., Zhang, M., Smith, N.A.: In: Proceedings of the conference on fairness, accountability, and transparency, pp. 369–378 (2019)
155. Austin, P.C., Harrell, F.E., Jr., Lee, D.S., Steyerberg, E.W.: Empirical analyses and simulations showed that different machine and statistical learning methods had differing performance for predicting blood pressure. *Sci. Rep.* **12**(1), 9312 (2022)
156. Weber, D., Hintermann, U., Zangger, A.: Scale and trends in species richness: considerations for monitoring biological diversity for political purposes. *Glob. Ecol. Biogeogr.* **13**(2), 97–104 (2004)
157. Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F.: Species distribution modeling in the cloud. *Concurr. Comput.: Pract. Exp.* **28**(4), 1056–1079 (2016)
158. Coro, G., Panichi, G., Scarponi, P., Pagano, P.: Cloud computing in a distributed e-infrastructure using the web processing service standard. *Concurr. Comput.: Pract. Exp.* **29**(18), e4219 (2017)
159. Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Pagano, P., Panichi, G., et al.: Enacting open science by D4Science. *Futur. Gener. Comput. Syst.* **101**, 555–563 (2019)
160. Schut, P., Whiteside, A.: OpenGIS Web Processing Service. OGC project document <http://www.opengeospatial.org/standards/wps> (2007)
161. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. *W3C Recommendation vol. 30* (2013)
162. Koop, D., Santos, E., Mates, P., Vo, H.T., Bonnet, P., Bauer, B., Surer, B., Troyer, M., Williams, D.N., Tohline, J.E., et al.: A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Comput. Sci.* **4**, 648–657 (2011)
163. Freire, J., Bonnet, P., Shasha, D.: In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 593–596 (2012)
164. Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Dell'Amico, A., Frosini, L., Lelii, L., Lettere, M., Mangiacrapa, F., et al.: Virtual research environments co-creation: The D4Science Experience. *Concurrency and Computation: Practice and Experience* p. e6925 (2022)
165. Froese, R.: Fishbase. *Oceanogr. Lit. Rev.* **3**(43), 321 (1996)
166. Marras, S., Cucco, A., Antognarelli, F., Azzurro, E., Milazzo, M., Bariche, M., Butenschön, M., Kay, S., Di Bitetto, M., Quattrocchi, G., et al.: Predicting future thermal habitat suitability of competing native and invasive fish species: from metabolic scope to oceanographic modelling. *Conserv. Physiol.* **3**(1), cou059 (2015)
167. Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R.: Predicting the distributions of marine organisms at the global scale. *Ecol. Model.* **221**(3), 467–478 (2010)
168. Lobo, J.M., Jiménez-Valverde, A., Hortal, J.: The uncertain nature of absences and their importance in species distribution modelling. *Ecography* **33**(1), 103–114 (2010)
169. Carvalho, B.M., Rangel, E.F., Ready, P.D., Vale, M.M.: Ecological niche modelling predicts southward expansion of *lutzomyia* (*Nyssomyia*) *flaviscutellata* (Diptera: Psychodidae: Phlebotominae), vector of leishmania (*leishmania*) *amazonensis* in South America, under climate change. *PLoS ONE* **10**(11), e0143282 (2015)
170. Hannemann, H., Willis, K.J., Macias-Fauria, M.: The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. *Glob. Ecol. Biogeogr.* **25**(1), 26–35 (2016)

171. Iannella, M., Cerasoli, F., D'Alessandro, P., Console, G., Biondi, M.: Coupling GIS spatial analysis and ensemble niche modelling to investigate climate change-related threats to the Sicilian pond turtle *Emys trinacris*, an endangered species from the Mediterranean. *PeerJ* **6**, e4969 (2018)
172. Danovaro, R., Company, J.B., Corinaldesi, C., D'Onghia, G., Galil, B., Gambi, C., Gooday, A.J., Lampadariou, N., Luna, G.M., Morigi, C., et al.: Deep-sea biodiversity in the Mediterranean sea: the known, the unknown, and the unknowable. *PLoS ONE* **5**(8), e11832 (2010)
173. Jimenez, C., Andreou, V., Evriviadou, M., Munkes, B., Hadjioannou, L., Petrou, A., Abu Alhaja, R.: Epibenthic communities associated with unintentional artificial reefs (modern shipwrecks) under contrasting regimes of nutrients in the Levantine Sea (Cyprus and Lebanon). *PLoS ONE* **12**(8), e0182486 (2017)
174. Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Bianchi, C.N., Corbera, J., Dailianis, T., et al.: The biodiversity of the Mediterranean sea: estimates, patterns, and threats. *PLoS ONE* **5**(8), e11842 (2010)
175. Di Lorenzo, M., Sinerchia, M., Colloca, F.: The north sector of the Strait of Sicily: a priority area for conservation in the Mediterranean Sea. *Hydrobiologia* **821**, 235–253 (2018)
176. Spanò, N., Domenico, E.D.: In: Fuerst-Bjelis, B. (ed.) *Mediterranean Identities*. IntechOpen, Rijeka, chap. 6, pp. 129–148. <https://doi.org/10.5772/intechopen.68942> (2017)
177. Duliü, J., Soldo, A., Jardas, I.: Adriatic fish biodiversity and review of bibliography related to Croatian small-scale coastal fisheries. In: *Adriatic Sea Small-Scale Fisheries*, pp. 103–125 (2005)
178. Bastari, A., Micheli, F., Ferretti, F., Pusceddu, A., Cerrano, C.: Large marine protected areas (LMPAS) in the Mediterranean Sea: the opportunity of the adriatic sea. *Mar. Policy* **68**, 165–177 (2016)
179. Abdul Malak, D., Livingstone, S., Pollard, D., Polidoro, B., Cuttelod, A., Bariche, M., Bilecenoglu, M., KE, C., Collette, B., Francour, P., Goren, M., Kara, H., Massuti, E., Papaconstantinou, C., Tunesi, L.: *Overview of the Conservation Status of the Marine Fishes of the Mediterranean Sea*, IUCN (2011)
180. Coro, G., Pagano, P., Ellenbroek, A.: Detecting patterns of climate change in long-term forecasts of marine environmental parameters. *Int. J. Digit. Earth* **13**(5), 567–585 (2020)
181. Öndes, F., Ünal, V.: the dominance of non-indigenous species in the catch composition of small-scale fisheries: a case study from the Kaş–Kekova special environmental protection area, Türkiye, Eastern Mediterranean. *Acta Ichthyol. Piscat.* **53**, 27–35 (2023)
182. Bakhoum, S.: Fish assemblages in surf zone of the Egyptian Mediterranean coast off Alexandria. *Turk. J. Fish. Aquat. Sci.* **19**(4), 351–362 (2018)
183. Kleitou, P., Moutopoulos, D.K., Giovos, I., Kletou, D., Savva, I., Cai, L.L., Hall-Spencer, J.M., Charitou, A., Elia, M., Katselis, G., et al.: Conflicting interests and growing importance of non-indigenous species in commercial and recreational fisheries of the Mediterranean Sea. *Fish. Manage. Ecol.* **29**(2), 169–182 (2022)
184. Santana-Garcon, J., Bennett, S., Marbà, N., Vergés, A., Arthur, R., Alcoverro, T.: Tropicalization shifts herbivore pressure from seagrass to rocky reef communities. *Proc. R. Soc. B* **290**(1990), 20221744 (2023)
185. Steinitz, W.: Beiträge zur kenntnis der küstenfauna palästinas. *Publicazioni della Stazione Zoologica di Napoli* **13**(1), 311–353 (1927)
186. Diamant, A.: Red-med immigration: a fish parasitology perspective, with special reference to the myxosporea. In: Golani, D., Appelbaum-Golani, B. (eds.) pp. 85–97 (2010)
187. Kamberi, E., Beqiri, K., Luli, K., Bakiu, R.: Tracking changes in fish diversity in the South-Eastern Adriatic Sea (Albania) based on local ecological knowledge. *Croat. J. Fish.* **80**(1), 17–25 (2022)
188. Azzurro, E., Ballerini, T., Antoniadou, C., Aversa, G.D., Souissi, J.B., Blašković, A., Cappanera, V., Chiappi, M., Cinti, M.F., Colloca, F., et al.: Climatefish: a collaborative database to track the abundance of selected coastal fish species as candidate indicators of climate change in the Mediterranean Sea. *Front. Mar. Sci.* **9**, 910887 (2022)
189. Dulčić, J., Pallaoro, A.: First record of the marbled spinefoot *Siganus rivulatus* (Pisces: Siganidae) in the Adriatic Sea. *J. Mar. Biol. Assoc. U.K.* **84**(5), 1087–1088 (2004)
190. Stamouli, C., AKEL, E.K., Azzurro, E., Bakiu, R., Bas, A., Bitar, G., Boyaci, Y., Cakalli, M., Corsini-Foka, M., Crocetta, F., et al.: New Mediterranean biodiversity records (December 2017). *Mediter. Mar. Sci.* **18**(3), 534–556 (2017)
191. Aglieri, G., Quattrocchi, F., Mariani, S., Baillie, C., Spatafora, D., Di Franco, A., Turco, G., Tolone, M., Di Gerlando, R., Milazzo, M.: Fish eDNA detections in ports mirror fishing fleet activities and highlight the spread of non-indigenous species in the Mediterranean Sea. *Mar. Pollut. Bull.* **189**, 114792 (2023)
192. Iglesias, S.P., Bergot, P., Breton, P., Brunelle, S., Camusat, M., Causse, R., Charbonnel, É., Chevaldonne, P., Cordier, Y., Cosquer, P., et al.: French ichthyological records for 2018. *Cybiuim: Revue Internationale d'Ichtyologie* **44**(4), 285–307 (2020)
193. Otero, M., Cebrian, E., Francour, P., Galil, B., Savini, D.: Monitoring marine invasive species in mediterranean marine protected areas (MPAS): a strategy and practical guide for managers. IUCN, Malaga (2013)
194. Daniel, B., Piro, S., Charbonnel, E., Francour, P., Letourneur, Y.: Lessepsian rabbitfish *Siganus luridus* reached the French Mediterranean coasts. *Cybiuim* **33**(2), 163–164 (2009)
195. Bardamaskos, G., Tsiamis, K., Panayotidis, P., Megalofonou, P.: New records and range expansion of alien fish and macroalgae in Greek waters (south-east Ionian sea). *Mar. Biodivers. Rec.* **2**, e124 (2009)
196. Schembri, P., Deidun, A., Falzon, M.: One siganus or two? On the occurrence of *Siganus luridus* and *Siganus rivulatus* in the maltese islands. *Mar. Biodivers. Rec.* (2012). <https://doi.org/10.1017/S175526721200053X>
197. Coro, G., Ellenbroek, A., Pagano, P.: An open science approach to infer fishing activity pressure on stocks and biodiversity from vessel tracking data. *Eco. Inform.* **64**, 101384 (2021)

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