Environ Ecol Stat (2016) 23:155–180
DOI 10.1007/s10651-015-0333-8

Automatic classification of climate change effects on marine species distributions in 2050 using the AquaMaps model

Gianpaolo Coro[1] · **Chiara Magliozzi**[1] ·
Anton Ellenbroek[2] · **Kristin Kaschner**[3] ·
Pasquale Pagano[1]

Received: 28 April 2015 / Revised: 24 October 2015 / Published online: 8 December 2015

Abstract Habitat modifications driven by human impact and climate change may influence species distribution, particularly in aquatic environments. Niche-based models are commonly used to evaluate the availability and suitability of habitat and assess the consequences of future climate scenarios on a species range and shifting edges of its distribution. Together with knowledge on biology and ecology, niche models also allow evaluating the potential of species to react to expected changes. The availability of projections of future climate scenarios allows comparing current and future niche distributions, assessing a species' habitat suitability modification and shift, and consequently estimating potential species' reaction. In this study, differences between the distribution maps of 406 marine species, which were produced by the AquaMaps niche models on current and future (year 2050) scenarios, were estimated and evaluated. Discrepancy measurements were used to identify a discrete number of categories, which represent different responses to climate change. Clustering analysis was then used to automatically detect these categories, demonstrating their reliability com-

---

Handling Editor: Pierre Dutilleul.

Electronic supplementary material The online version of this article (doi:10.1007/s10651-015-0333-8) contains supplementary material, which is available to authorized users.

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

[1] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, via Moruzzi 1, 56124 Pisa, Italy

[2] Food and Agriculture Organization of the United Nations (FAO), Viale delle Terme di Caracalla, 00153 Rome, Italy

[3] Department of Biometry and Environmental Systems Analysis, Albert-Ludwigs University, Hauptstrasse 1, 79104 Freiburg, Germany

pared to human supervised classification. Finally, the distribution of characteristics like extinction risk (based on IUCN categories), taxonomic groups, population trends and habitat suitability change over the clustering categories was evaluated. In this assessment, direct human impact was neglected, in order to focus only on the consequences of environmental changes. Furthermore, in the comparison between two climate snapshots, the intermediate phases were assumed to be implicitly included into the model of the 2050 climate scenario.

Keywords AquaMaps · Big Data · Climate change · Clustering analysis · Ecological niche modelling · GIS · Maps comparison · OGC standards · Species distribution maps

1 Introduction

Climate change impact on marine species geographical distribution is becoming important issue in public and scientific agendas. There is broad consensus that global climate change can have serious consequences for marine organisms (Brierley and Kingsford 2009). In many locations, the coupling of anthropogenic pressure and climate change has already caused a dramatic decline of marine environments, both at population and biodiversity levels (Cheung et al. 2009). Consequently, the disruption of marine biodiversity and ecosystems is having irreversible impacts on a wide range of ecosystem services that are essential for human well-being: food provisioning, water quality regulation, leisure and recreation (Levin and Lubchenco 2008). It is therefore crucial to maintain adaptive capacities of marine ecosystems by preserving their heterogeneity, modularity and redundancy (Hughes et al. 2005).

One way to study the potential impact of climate change is to combine niche modelling techniques with predictions of future environmental conditions produced by climate change. Niche modelling allows generating predictions about potential or actual/native species distribution in geographic space by analysing the relationship between environmental conditions and observations of species occurrence (Pearson and Dawson 2003). This approach also allows projecting species habitat distributions under different climate scenarios (Dawson et al. 2011; Pearson and Dawson 2003). Consequently, niche modelling has been used to explore and predict ecosystems perturbations due to climate change (Thuiller 2004). Modelling potential shifts in geographic ranges using niche modelling represents a simplified approach, that cannot fully capture the true impact of climate change on species distribution. Species' populations have complex and dynamic geographical structures that are difficult to model. For instance, fishes communities usually do not shift rapidly at the arrival of a new climate regime and this transition phase is usually neglected by niche models. Furthermore, habitat change should be measured also in terms of abundance fluctuations other than geographic change (Genner et al. 2004). Nevertheless, analysing a large sample of estimated niche distributions having large time resolution, may allow assessing overall effects of climate change on marine ecosystems (Costa et al. 2010).

The aim of this paper is to introduce a comparative approach that relies on computational technology (e.g. Cloud computing and machine learning), and can be used in batch analyses over large biological data sets in both terrestrial and marine environments. Two niche models were used to estimate native distributions of a set of marine species, under the current environmental scenario and another one projected in 2050. Models rely on a bioclimatic envelope approach, including biogeographic information and therefore allowing to explore geographic shifts as response to climate change (Pearson and Dawson 2003; Thuiller 2004). Differences between models were estimated by automatically detecting a number of finite and discrete categories of habitat changes. Reliability of those categories were validated by comparing automatic with human supervised classification. This approach was applied on distribution data of 406 marine species, which have significant commercial and/or biological importance by FAO. As case study, several examples show habitat expansion and loss in the distribution of threatened species from several taxonomic groups according to the IUCN Red List of species (IUCN 2015).

2 Materials and methods

In this section, evidences about climate change effects on species distributions and available techniques to monitor them are reported. Additionally, the benchmark dataset and the models used are described. Finally, the proposed method is introduced.

2.1 Overview

The ability of different species to adapt to environmental variability depends on their tolerance and role in the ecosystem (Somero 2012; Sunday et al. 2012). Some species may physiologically tolerate changes through adaptation and acclimatization. Alternatively, species may migrate or shift with changing environmental conditions, resulting in local extinctions or, if neither is possible, real extinctions (Parmesan 2006). As suggested by coupled climate-population models (Cheung et al. 2011; Hare et al. 2010), these responses can result in shifting of abundances or in modified geographic distributions. Researchers have been able to detect changes in biodiversity of marine species (e.g. fishes, invertebrates and plants), often with changes in distribution ranges. For example, they have detected an increase of species richness of small-body fishes in the North Sea (Perry et al. 2005) and a general shift in the pattern of their latitudinal and depth ranges (Dulvy et al. 2008; Hiddink and Ter Hofstede 2008; Hsieh et al. 2009; Parmesan 2006). Other reported phenomena are (1) shifts of mussel-dominated communities to those dominated by fleshy algae (Wootton et al. 2008), (2) calcareous corals replaced by non-calcareous algae (Bellwood et al. 2004), (3) seagrass forests and calcareous algae being replaced by fleshy macro-algae in higher CO2 regions (Jordà et al. 2012; Koch et al. 2013). These changes may not all directly be linked to global warming, but they may also depend on other correlated factors, like basin-scale circulation and water stratification (Hyrenbach and Veit 2003; Sydeman et al. 2014). The effect of climate change is particularly evident at the poles and in the tropics. The retreat of polar ice influences food availability, organism growth and reproduction (Somero 2012; Pellissier et al. 2013; Wassmann et al. 2011). Tropical ecosystems have undergone strong declines after minor increases in temperature, and ocean acidification may hasten their deterioration (Anthony et al. 2011; Erez et al. 2011).

To address these problems, bioclimatic envelopes models correlate the observed geographical distribution of a species to climatic variables and have been used to predict habitat suitability under different climate scenarios (Araújo et al. 2005; Botkin et al. 2007; Cheung et al. 2009, 2010). Among the criticisms is the exclusion of biotic interactions, although they have been extensively used with promising results (Araujo and Guisan 2006; Berry et al. 2002; Schwartz 2012; Thuiller et al. 2005). In this context, mechanistic niche models use both biological preferences and physiological limits of a species (Chuine and Beaubien 2001), while correlative approaches automatically extract correlations between environmental characteristics and species presence (Pearson 2012). Thuiller (2004) reports methods to assess the reliability of niche models using future climate estimates.

2.2 Benchmark models and data

The AquaMaps models (Kaschner et al. 2006) are ecological niche models that can estimate species habitat suitability under different environmental predictions. AquaMaps are presence-only models, which incorporate scientific expert knowledge into species habitats modelling, to account for the known biases and limitations of currently available marine species occurrence record data sets (Ready et al. 2010). The AquaMaps approach includes two models estimating the native ("actual") distribution of a species today (2013–2015) and in 2050: AquaMaps Native and AquaMaps Native-2050. Expert knowledge is used in modelling species-habitat relations at global scale with 0.5° resolution, relying on the following environmental variables: depth, salinity, temperature, primary production, distance from land and sea ice concentration (Corsi et al. 2000). Those models estimate habitat suitability by combining the envelopes of a species preference to the above environmental variables. They combine mechanistic assumptions and automatic procedures to estimate species presence in a certain area. The modelling approach is reliable with respect to other approaches (Ready et al. 2010), but less accurate when expert knowledge at global scale is missing. The difference between the two algorithms is that Native-2050 accounts for (1) modifications in the extensions of FAO Areas and (2) locations that the species could reach thanks to the increase of the water surface level. Furthermore, the two algorithms are based on two different climate scenarios. The environmental data in 2050 account for a global climate change scenario, described under the IPCC SRES scenario A1B (Nakicenovic and Swart 2000), of a future of rapid global economic, population and technological growth, where the average surface temperature increases, the ice concentration decreases and the salinity increases globally but decreases in some locations (Reyes 2015). In order to generate climate estimates for 2050, the AquaMaps consortium produces new environmental data using the ECHAM model (Roeckner et al. 1992) for two years: 2002 (based on observations from 1990 to 2002) and 2050 (based on estimated observations from 2040 to 2049). Each value in 2050 is then estimated as

$$
\begin{aligned}
Val.\,in\ 2050 = {} & Avg\ val.\,today \\
& + (ECHAM\ est.\ for\ 2050 - ECHAM\ est.\ for\ 2002)
\end{aligned}
$$

On the basis of these values, the AquaMaps Native-2050 algorithm estimates the native habitat suitability of the species in 2050.

For the objectives of this paper, a large repository of already produced AquaMaps distributions was used: the D4Science e-Infrastructure (Candela et al. 2014). D4Science is an e-Infrastructure that can also exploit other infrastructures, services and information systems. D4Science is currently used as support to communities in computational biology (Castelli et al. 2013), fisheries management (Castelli and Michel 2011), and oceanography (Sorvari and Brus 2012). It has links to other infrastructures such as the European Grid Infrastructure (EGI), the Global Biodiversity Information Facility (GBIF), the Ocean Biogeographic Information System (OBIS), myOcean, the World Ocean Atlas, the FAO Data repository, the Catalogue of Life and SeaDataNet. D4Science integrates these systems to exploit their respective synergies, thus providing its community of practice with a set of novel and enhanced services. D4Science also supports the creation and operation of Virtual Research Environments (VREs), i.e. web-based working environments, where groups of scientists have transparent and seamless access to a shared set of remote resources (data, tools and computing capabilities) and support collaboration across institutions, disciplines, and nations. D4Science hosts around 25,000 maps developed with the Native and Native-2050 algorithms, updated to February 2014. Differently from the maps published on the AquaMaps website[1] the maps hosted by D4Science are not manually reviewed. Nevertheless, they have been produced using datasets that incorporate expert information about the species envelopes and manual adjustments of environmental information. Thus, these maps represent a sub-optimal dataset, but they were used in the experiments reported in this paper, because this sub-optimal scenario is frequent when dealing with large data collections. The proposed modelling approach is requested to be able to describe patterns using these data, and identify systematic errors or “noisy” subsets, which require manual review after classification. Another reason for choosing the maps on D4Science is their public availability under standard formats, which makes the proposed experiment reproducible and repeatable also by other scientists. In fact, D4Science hosts the maps on a distributed network of GIS services that balances the load of requests by many users. The information in the AquaMaps distributions is published under the WMS and the WFS GIS standards (de La Beaujardière 2004). A web interface (CNR 2015) allows visualizing, superposing and inspecting the GIS datasets. Figure 1 shows an example for the bluntnose sixgill shark (*Hexanchus griseus*), for which the AquaMaps model predicts an expansion of habitat suitability range in 2050. This is an example of sub-optimal map, which was neither supervised by an expert nor contains artificial, manually-defined bounding boxes.

The analysis here presented was restricted to a subset of the species hosted on D4Science. In particular, the analysis focused on the species contained in the FAO Fact Sheets (FAO 2015). This dataset provides key information on fisheries and aquaculture-

---

[1] www.aquamaps.org.

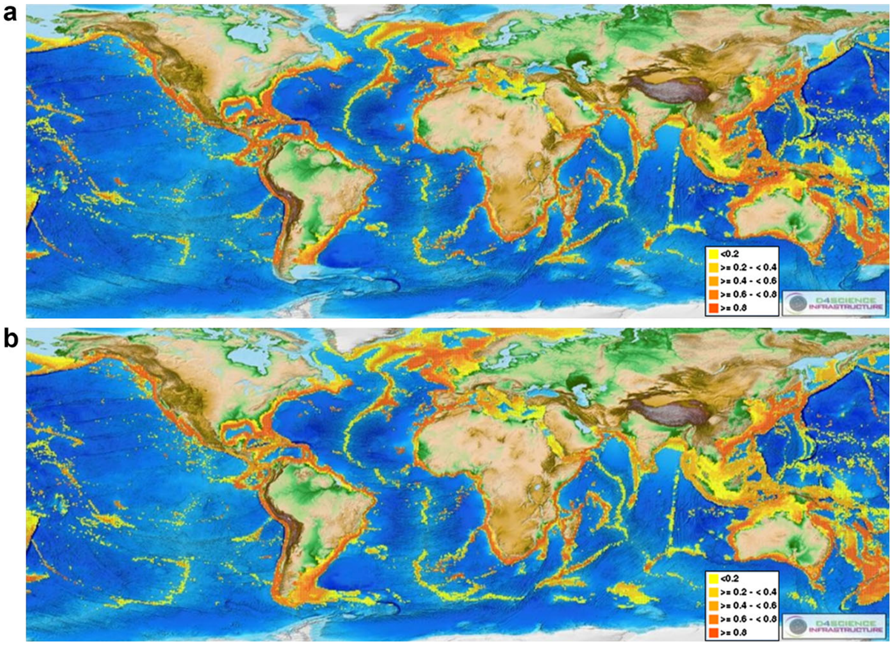

Fig. 1 AquaMaps Native (**a**) and Native-2050 (**b**) distributions of the bluntnose sixgill shark (*Hexanchus griseus*). *Darker colours* indicate higher probability areas

related subjects and is available free online[2]. The 550 aquatic species were chosen, because they represent a set of species of particular commercial, biological and human interest. D4Science already hosts the AquaMaps Native and Native-2050 distributions for 406 of them. The scientific names of these species are reported in the supplementary material of this paper[3] and the maps are available on the D4Science web portal for consultation[4].

2.3 Proposed methodology

This section describes (1) the approach used to analyse niche maps at global scale (2) the process to produce species-specific statistics of habitat change and then (3) a summarization process of these statistics that groups them into a discrete number of categories.

In order to quantify the differences between the benchmark AquaMaps distributions for today and 2050, the D4Science e-Infrastructure processing facilities were used (Coro et al. 2015). The e-Infrastructure hosts a maps comparison process relying on Cloud computing and described in Coro et al. (2014). This process performs a pixel-by-pixel comparison between two probability distribution maps at a given resolution and

---

[2] http://www.fao.org/fishery/species/search/en.

[3] Also available at this public link: http://goo.gl/ZjjOkS.

[4] https://www.d4science.org/group/biodiversitylab/geo-visualisation.

Table 1 Quantities calculated by the maps comparison process of Coro et al. (2014)

Quantities	Description
Mean	Mean discrepancy between the probabilities
Variance	Variance of the discrepancies
Number of errors	Number of discrepant locations
Number of comparisons	Number of compared locations
Accuracy	Number of agreed locations divided by the number of compared locations
Relative error	Geographic areas of maximum discrepancy
Maximum error	Maximum recorded discrepancy between two locations
Maximum error point	Average error divided by the maximum error
Cohens Kappa	A value representing the agreement between the two maps with respect to chance
Trend	Expansion, contraction or stationary labels indicating if the latter map is more or less extended than the former

calculates statistics (Table 1). Discrepancy between maps is estimated with respect to a threshold, which was fixed at 0.5 as also reported in Coro et al. (2014). The resolution of the AquaMaps is 0.5° latitude by 0.5° longitude, which means that global scale distributions require comparing around 170,000 points.

This process was applied to 406 benchmark species, comparing the AquaMaps Native and Native-2050 maps for each species. An R script was built, which went through the benchmark species, invoked the maps comparison process through D4Science and reported the statistics in a summary file. The script used the D4Science connection APIs (GCube 2015). Computation required 30 hours on the entire benchmark dataset and resulted in a CSV file containing 406 rows, each reporting the statistics for one species (also see supplementary material[5]).

In order to detect common properties in distribution change, a criterion to group statistics was developed. A marine biologist analysed a random sample of 50 species by hand, classifying four main types of changes in species distribution maps: (1) in one delimited area, (2) in separated but limited areas, (3) in extended continuous areas, (4) overall in many areas. This analysis suggested that it was possible to aggregate the discrepancies into a finite number of groups.

In order to automate this process and analyse all the species, clustering analysis was used. Clustering analysis is a common technique in data mining that is able to group together numeric vectors, according to a certain similarity criterion. Similarity is usually measured in terms either of density or of euclidean distance. Both the approaches were investigated, which considered the numeric values in the maps comparison output as the elements of vectors to group together. Maximum relative error was not included in the process. However, the number of comparisons and the number of errors were normalised by maximum number of comparisons. This choice helps

---

[5] Also available at this public link: http://goo.gl/ZjjOkS.

clustering analysis to detect more clusters, because using different scales in vectors dimensions would have introduced noise in the clustering algorithms. In particular, the distance-based XMeans algorithm was used (Pelleg and Moore 2000). XMeans is a variant of the popular KMeans algorithm (MacQueen 1967), which introduces several efficiency enhancements. An important difference with respect to KMeans is that the number of optimal clusters to search for is not specified *a priori*. Instead, it requires to set a minimum and a maximum number of clusters (K_{min} and K_{max}) to search for. The X–Means algorithm starts from K_{min} and adds centroids as far as K_{max} is reached. At each step, the K–Means algorithm is run, which finds the best assignment of the vectors to the indicated number of clusters. K–Means indicates a score for this assignment, based on the distortion measure, i.e. the average squared distance of the points to their clusters centroids. The output of X–Means is the result of the K–Means that has given the best score, and consequently the best number of clusters. X–Means also adds efficiency enhancements to K–Means, using *kd*-trees (Bentley 1975) and *blacklisting* to support the processing. Furthermore, at each step of the computation, the location of the centroids of the additional clusters is decided using the Bayesian Information Criterion (BIC) (Schwarz 1978). For further details see Pelleg and Moore (2000). Apart from X-Means, also the DBScan density-based clustering algorithm was used (Ester et al. 1996), but no setting for this algorithm resulted in <100 clusters, thus its results will be not reported.

Clustering analysis was applied to the produced maps comparison statistics using the implementations running on the D4Science services (Coro et al. 2013, 2015), which host these procedures as-a- Service. The X-Means algorithm was executed searching for a number of clusters between 1 and 50. Although the interval was large, the algorithm ended in only *seven* optimal clusters. Thus, the algorithm found an optimal separation of the vectors according to their relative euclidean distance and a semantic interpretation was given to each of them. When this assignment was completed, a biologist interpreted and evaluated the results. The result of clustering analysis is reported in Table 2, along with the values of the cluster centroids.

Cluster 6 and 7 identify the same trend but separate different distribution scales. Clusters 1 and 3 identify the trend originally captured by the biologist for geographically-limited changes, but they refine the separation by distinguishing between global-scale species with poor changes and species with moderate habitat variations. Cluster 1 contains most of the species in the benchmark dataset. Cluster 2, instead, identifies the species with the largest variation due to climate change.

3 Results

This section reports the result of the proposed methodology. The reliability of the clustering aggregation is demonstrated, and possible correlations between the distribution of taxonomic groups/IUCN categories and the clusters are highlighted. Furthermore, the relation between IUCN assessed short-term population trends and long-term trends of expected habitat change are investigated, and examples of correspondence between automatic classification and ecological studies are reported. Finally, the role of expert review on the output of the proposed method is clarified.

Table 2 Interpretation of the output of clustering analysis, with indication of the values associated to the clusters centroids

Clusters	Mean	Variance	Norm. n. of errors	Norm. n. of comparisons	Norm. accuracy	Max. error	Cohens' Kappa	No. of sp. in the cluster	Interpretation
Cluster1	0.77	0.026	0.0028	0.045	0.91	0.99	0.69	159	From Small to Large scale Good overall agreement with localized moderate probability changes (40 % in 1 FAO Area)
Cluster2	0.84	0.020	0.0056	0.019	0.66	1	0.29	47	From Small to Medium scale Strong (>=40 %) overall probability changes in all the FAO Areas
Cluster3	0.62	0.0081	0.012	0.66	0.98	0.8525	0.82	32	Global scale distr. Conservative with poor localized probability changes (from 10 to 20 % in 1 or 2 FAO Areas)
Cluster4	0.67	0.0096	0.0015	0.058	0.95	0.901392	0.75	79	Small and Medium scale Moderate,uniform,over-all probability changes (from 20 to 40 % in allthe FAO areas)
Cluster5	0.56	0.00019	0.00017	0.056	0.99	0.627308	0.83	52	Small and Medium scale. Overall, uniform,poor probability changes (from 10 to 20 % in all the FAO areas)
Cluster6	0	0	0	0.42	1	0	0.90	6	Medium scale Conservative, without changes
Cluster7	0	0	0	0.035	1	0	0.83	31	Small scale Conservative, without changes

Table 3 Comparison between the assignments of the clusters to the species by a biologist and by clustering analysis

Biologist	Clustering						
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	26	2	0	0	3	0	0
Cluster 2	0	22	0	1	0	0	0
Cluster 3	0	0	16	0	0	0	0
Cluster 4	2	0	0	23	1	0	0
Cluster 5	1	0	0	5	23	0	0
Cluster 6	0	0	0	0	0	5	0
Cluster 7	0	0	0	0	0	0	22

The agreement matrix between the two evaluators should be read per column: for example, column one reports that the biologist agreed on 26 species for cluster 1, but assigned two species to cluster 4 and one species to cluster 5

3.1 Reliability of classification results

In order to assess the reliability of the interpretation of the clusters, the biologist who carried out the preliminary analysis, manually assessed the classification of 152 randomly selected species. In the assessment, the biologist assigned one of the interpretations of the X–Means clusters to each species, by looking at its current and future (year 2050) distribution map. Table 3 reports the confusion matrix, i.e. the assignments made by the biologist to the clusters on the basis of their descriptions.

The confusion matrix was used to estimate the agreement between the biologist and X–Means as the percentage of matching assignments (absolute percentage of agreement). Furthermore, the Cohen's Kappa (Cohen et al. 1960) was calculated, which estimates the agreement between two evaluators compared to purely random assignments. The resulting values and the the Kappa coefficient (Table 4) both indicate that the agreement is excellent. This gives more reliability to the semantic interpretations reported in Table 2, and consequently to the categories identified in the maps variations.

3.2 Statistical analysis

This section presents a statistical analysis of the distribution of different taxonomic groups and of the IUCN categories over the identified clusters. Figure 2a shows the distribution of the taxonomic groups involved in the benchmark dataset. Species of fish are the most represented, followed by marine mammals, crustaceans, molluscs, reptiles, echinoderms and corals. Figure 2b shows the distribution of the IUCN categories over the entire dataset. The NE (Non-Evaluated) category was introduced for the scope of this paper, for species not included in the IUCN Red List at the moment. Table 5 reports the number of species in each category. Apart of the NE category, most of the species fall in the "Least concern" (LC) category and many are "Data Deficient" (DD). In order to enhance the maps comparison process, a species "trend"

Table 4 Agreement between a biologist and clustering analysis in terms of Cohen's Kappa (Cohen et al. 1960), absolute percentage of agreement and according to two reference interpretations of the Kappa (Fleiss 1971; Landis and Koch 1977)

Kappa	0.88
Absolute perc. of agreement	90.13 %
Intepretation acc. to Fleiss	Excellent
Interpretation acc to Landis-Koch	Almost perfect

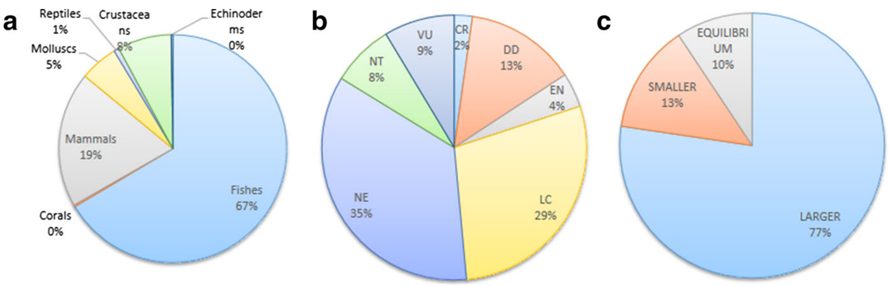

Fig. 2 Distributions of **a** taxonomic groups, **b** IUCN Red List categories (*LC* least concern, *DD* data deficient, *NE* non-evaluated, *CR* critically endangered, *EN* endangered, *VU* vulnerable, *NT* near threatened), **c** "trends" in habitat change over the entire benchmark dataset

Table 5 Report of the number of benchmark species falling in each IUCN Category

IUCN category	Number of species
Non-evaluated (NE)	143
Least concern (LC)	116
Data deficient (DD)	55
Critically endangered (CR)	9
Endangered (EN)	17
Vulnerable (VU)	35
Near threatened (NT)	31

was defined using categories as described in Coro et al. (2014), renaming them as LARGER, SMALLER and EQUILIBRIUM. The definition in Coro et al. (2014) for LARGER indicates an overall increase of suitable areas in 2050, with possible localized unsuitability. SMALLER habitats have overall fewer suitable locations in 2050, which does not exclude limited local expansions. EQUILIBRIUM indicates habitat sizes that do not significantly change. Figure 2c shows that most of the species in the benchmark dataset will have larger habitat suitability in 2050, this will be further discussed in the next sections.

Figure 3a shows details of the distribution of the IUCN categories over the clusters. Species without evaluation (NE) are statistically significant and are present in all the clusters, except for cluster 6 (medium scale conservative). LC species are found in all the clusters. Vulnerable, endangered and near threatened species are mostly present

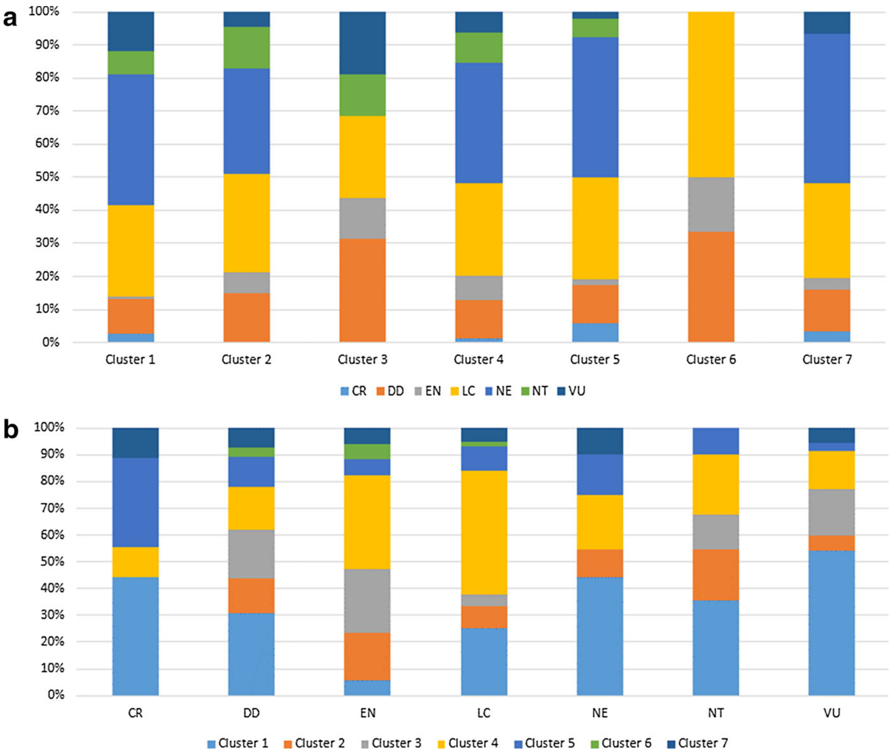

Fig. 3 **a** Distributions of the IUCN Red List categories of the benchmark species over the automatically detected clusters. **b** Distribution of the automatically detected clusters over the IUCN Red List categories

in cluster 3 (global scale with few changes), which indicates that their overall suitable habitat change by 2050 will be small. Several critically endangered species are present in clusters 5 and 7 split (small and medium scale with poor or no changes), which indicates slight changes in the habitat locations of these species. This means that their current distribution is likely to remain unchanged. Figure 3b shows that very few clusters (e.g. 3 and 6) seem to correlate with the IUCN classifications. Cluster 2 (strong habitat variations) is represented in almost all categories.

Figure 4 shows that the taxonomic groups are not balanced over the clusters. This is due to the abundance of fish species in the benchmark dataset. On the other hand, cluster 3 contains only mammals and fishes. Cluster 4 (small and medium scale with moderate changes) is the one containing the only echinoderm of the benchmark dataset, i.e. the Chilean sea urchin (*Loxechinus albus*), which is indeed sensitive to moderate habitat changes. Molluscs and crustaceans are well-present in clusters 7, 5 and 1, indicating their good overall resistance to climate change. On the other hand, cluster 2 also contains *Illex argentinus* (a cephalopod) and *Acetes japonicus* (a shrimp), whose suitable habitat increases. The only coral species in the benchmark dataset is *Corallium rubrum*, which falls in cluster 4. According to the proposed analysis, the suitable habitat of this species will change moderately, with an overall shrinking of its range (Table 6).

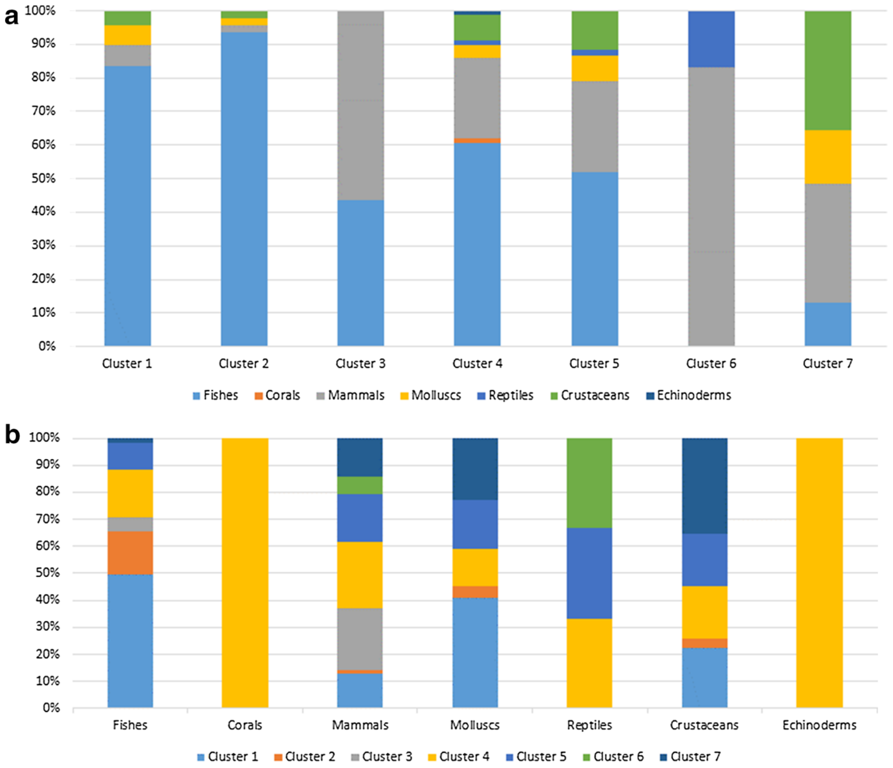

Fig. 4 **a** Distributions of the taxonomic groups in the benchmark species over the automatically detected clusters. **b** Distribution of the automatically detected clusters over taxonomic groups

3.3 Habitat suitability loss and increase

The distribution of the "trend" parameter values in Fig. 5a shows that resilient species (cluster 6 and 7) present EQUILIBRIUM trends. Most of the species that loose habitat fall in cluster 5 and this indicates that expected loss is slight, but can involve local unsuitability areas. Figure 5b shows the distribution of the trends over the IUCN categories. Figure 2c reports that suitable habitat for most species is likely to increase. This agrees with bioclimatic theories that predict overall expanding niches and increasing abundance trends with climate change (Genner et al. 2004). Table 6 lists species whose habitats shrink in 2050 according to AquaMaps. These species are strongly influenced by changes in temperature, salinity and other abiotic factors at macro-geographic scale.

As further comparison, the general population trends indicated by IUCN in the Red List were compared to the clustering analysis. These represent IUCN-assessed short-term population trends of abundance and distribution, and thus they were compared to the long-term behaviours indicated by clustering analysis. Population trends are available in the IUCN Red List, although only 96 of the 406 benchmark species were assessed at the time this evaluation was made. Possible IUCN assessment values are "increasing", "stable" and "decreasing". Figure 6a reports distribution over the clus-

Table 6 Species reducing their overall habitat suitability in 2050 according to AquaMaps

Species scient. name	Mean error	Error variance	Norm. n. of errors	Norm. n. of comparisons	Norm. accuracy	Max. error	Cohens' Kappa	Cluster ID	IUCN category	Trend	Tax. group
Acipenser oxyrinchus oxyrinchus	0.61	~0	~0	0.006	0.974	0.76	0.768	5	NT	SMALLER	fish
Acipenser sturio	0.5	~0	~0	0.016	1	0.5	0.873	5	CR	SMALLER	fish
Alopias pelagicus	0.75	0.02	0.004	0.027	0.851	1	0.611	1	VU	SMALLER	fish
Alopias vulpinus	0.66	0.01	0.008	0.577	0.987	0.98	0.826	3	VU	SMALLER	fish
Anguilla anguilla	0.54	~0	0.001	0.106	0.991	0.59	0.778	5	CR	SMALLER	fish
Arctica islandica	0.57	~0	~0	0.025	0.996	0.69	0.778	5	NE	SMALLER	mollusc
Carcharias taurus	0.65	0.01	0.007	0.057	0.875	1	0.623	4	VU	SMALLER	fish
Corallium rubrum	0.66	0.01	~0	0.003	0.855	0.88	0.457	4	EN	SMALLER	coral
Crangon crangon	0.52	~0	~0	0.006	0.997	0.52	0.776	5	NE	SMALLER	crustacean
Dentex angolensis	0.55	~0	~0	0.003	0.981	0.57	0.888	5	NT	SMALLER	fish
Dentex canariensis	0.52	~0	~0	0.003	0.992	0.52	0.839	5	LC	SMALLER	fish
Dentex congoensis	0.55	~0	~0	0.002	0.996	0.55	0.742	5	LC	SMALLER	fish
Engraulis ringens	0.54	~0	~0	0.002	0.972	0.6	0.836	5	LC	SMALLER	fish
Farfantepenaeus aztecus	0.65	0.01	0.001	0.028	0.96	0.86	0.619	4	NE	SMALLER	crustacean
Farfantepenaeus duorarum	0.62	0.01	~0	0.024	0.98	0.96	0.601	4	NE	SMALLER	crustacean
Feresa attenuata	0.55	~0	~0	0.451	0.999	0.72	0.93	3	DD	SMALLER	marine mammal
Ginglymostoma cirratum	0.6	0.01	0.002	0.019	0.913	0.91	0.562	4	DD	SMALLER	fish
Heterodontus japonicus	0.81	0.02	0	0.002	0.925	0.99	0.692	1	LC	SMALLER	fish
Heterodontus zebra	0.75	0.02	0.002	0.009	0.768	1	0.438	2	LC	SMALLER	fish

Table 6 continued

Species scient. name	Mean error	Error variance	Norm. n. of errors	Norm. n. of comparisons	Norm. accuracy	Max. error	Cohens' Kappa	Cluster ID	IUCN category	Trend	Tax. group
Hydrurga leptonyx	0.55	~0	~0	0.376	1	0.64	0.871	5	LC	SMALLER	marine mammal
Istiophorus platypterus	0.67	0.01	0.033	0.3	0.89	1	0.526	4	LC	SMALLER	fish
Lagenorhynchus cruciger	0.81	~0	~0	0.284	1	0.81	0.882	4	LC	SMALLER	marine mammal
Lamna nasus	0.6	0.01	0.001	0.325	0.998	0.9	0.869	4	DD	SMALLER	fish
Leptonychotes weddellii	0.64	0.01	0.001	0.102	0.986	0.97	0.688	4	LC	SMALLER	marine mammal
Lissodelphis borealis	0.62	0.01	~0	0.056	0.999	0.78	0.821	4	LC	SMALLER	marine mammal
Lissodelphis peronii	0.54	~0	~0	0.266	0.999	0.66	0.895	5	DD	SMALLER	marine mammal
Litopenaeus setiferus	0.63	0.02	~0	0.015	0.973	0.95	0.716	1	NE	SMALLER	crustacean
Lobodon carcinophaga	0.57	~0	~0	0.353	1	0.64	0.933	5	LC	SMALLER	marine mammal
Loligo pealeii	0.7	0.02	0.002	0.013	0.861	1	0.575	1	NE	SMALLER	mollusc
Merluccius capensis	0.78	0.02	~0	0.002	0.987	0.86	0.915	4	NE	SMALLER	fish
Mesoplodon grayi	0.58	0.01	~0	0.33	0.999	0.85	0.909	4	DD	SMALLER	marine mammal
Mesoplodon mirus	0.52	~0	~0	0.163	0.999	0.56	0.912	5	DD	SMALLER	marine mammal
Mugil cephalus	0.54	~0	0.003	0.088	0.971	0.64	0.648	5	LC	SMALLER	fish
Odontaspis ferox	0.69	0.02	0.002	0.039	0.941	0.98	0.76	1	VU	SMALLER	fish
Ommatophoca rossii	0.58	~0	~0	0.1	0.996	0.71	0.639	5	LC	SMALLER	marine mammal
Pandalus borealis	0.68	0.02	~0	0.028	0.992	0.97	0.703	1	NE	SMALLER	crustacean
Pandalus montagui	0.53	~0	~0	0.01	0.999	0.53	0.848	5	NE	SMALLER	crustacean
Penaeus semisulcatus	0.75	~0	~0	0.052	0.999	0.78	0.887	5	NE	SMALLER	crustacean

Table 6 continued

Species scient. name	Mean error	Error variance	Norm. n. of errors	Norm. n. of comparisons	Norm. accuracy	Max. error	Cohens' Kappa	Cluster ID	IUCN category	Trend	Tax. group
Peponocephala electra	0.61	0.02	~0	0.438	1	0.8	0.909	3	LC	SMALLER	marine mammal
Phocoena sinus	0.9	~0	~0	~0	0.971	0.9	0.943	4	CR	SMALLER	marine mammal
Platichthys flesus	0.54	~0	~0	0.021	0.999	0.54	0.921	5	LC	SMALLER	fish
Psettodes erumei	0.7	0.02	0.001	0.016	0.909	1	0.66	1	NE	SMALLER	fish
Rastrelliger kanagurta	0.7	0.02	0.003	0.038	0.934	1	0.738	1	DD	SMALLER	fish
Rhincodon typus	0.66	0.01	0.028	0.416	0.933	0.98	0.726	3	VU	SMALLER	fish
Ruditapes philippinarum	0.55	~0	~0	0.066	0.998	0.57	0.896	5	NE	SMALLER	mollusc
Scylla serrata	0.63	0.01	~0	0.035	0.998	0.77	0.818	4	LC	SMALLER	fish
Sicyonia brevirostris	0.61	0.01	0.001	0.039	0.985	0.95	0.694	4	NE	SMALLER	crustacean
Spisula solidissima	0.69	0.02	~0	0.009	0.957	0.98	0.849	1	NE	SMALLER	mollusc
Squatina africana	0.55	~0	~0	0.005	0.997	0.55	0.572	5	DD	SMALLER	fish
Stegostoma fasciatum	0.72	0.02	0.002	0.023	0.924	1	0.746	1	VU	SMALLER	fish
Trachurus trachurus	0.76	0.02	~0	0.02	0.993	1	0.862	1	NE	SMALLER	fish
Trichiurus lepturus	0.69	0.02	0.006	0.071	0.914	1	0.678	1	NE	SMALLER	fish
Turbo cornutus	0.52	~0	~0	0.003	0.998	0.52	0.739	5	NE	SMALLER	mollusc
Xiphopenaeus kroyeri	0.5	~0	~0	0.029	1	0.5	0.938	5	NE	SMALLER	crustacean

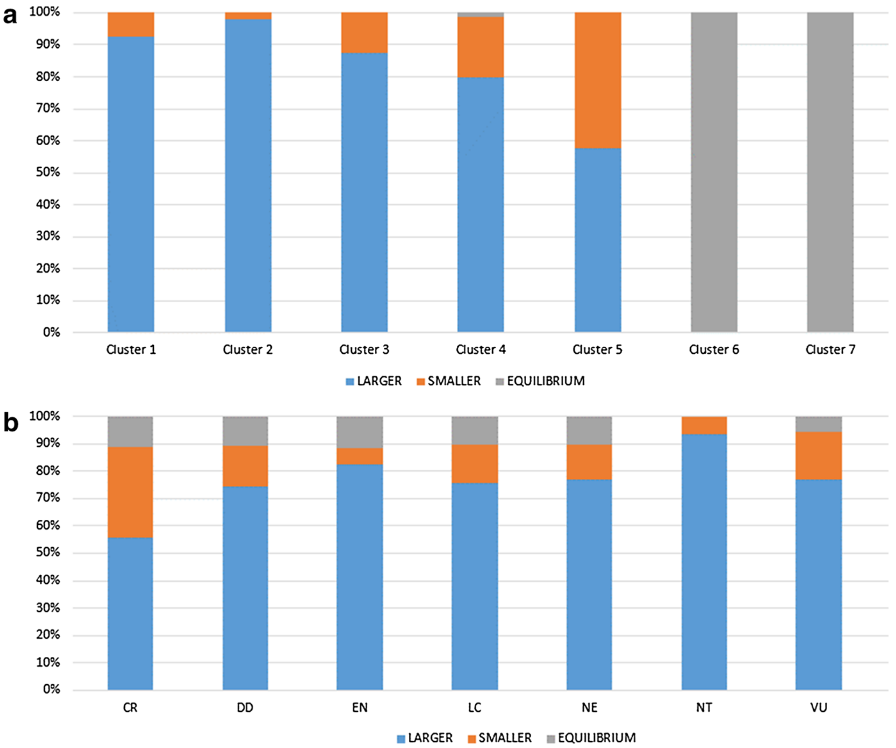

Fig. 5 **a** Distributions of habitat suitability changes ("trends") between today and 2050 over the automatically detected clusters. **b** Distribution of the automatically detected clusters over habitat suitability trends

ters and Fig. 6b shows abundance in the clusters. Cluster 6 contains only one species whose population trend has been assessed by IUCN, and whose habitat will increase according to AquaMaps. This cluster has therefore a poor statistical meaning. Cluster 7 contains many stable populations, this cluster contains most of the EQUILIBRIUM trends (Fig. 5a). Many decreasing trends fall in clusters 3, 4 and 5, which agrees with the distribution of long-term trends in Fig. 5a, although the SMALLER trends are more abundant in cluster 5. It is also possible to compare the IUCN short-term trend with the maps comparison long-term predictions: 12 of 13 species (92.3 %) having increasing IUCN population trend also result in LARGER trends, thus there is high agreement between short- and long-term increasing trends. SMALLER trends coincide with IUCN decreasing trends only for 9 of 52 species (17.3 %), whereas three decreasing trends are classified as EQUILIBRIUM trends (5.8 %). Thus, much of the decreasing trends are classified as LARGER long-term trends, but they fall in the clusters that contain most of the SMALLER trends. This could indicate a correlation between short-term decreasing trends and long-term SMALLER trends. EQUILIBRIUM trends coincide with stable trends for 3 of 31 species (9.7 %), whereas they correspond to SMALLER trends for two species (6.4 %). Thus, most of the stable trends correspond to LARGER trends, but several of them fall in cluster 7, which could indicate a stable equilibrium for these populations.

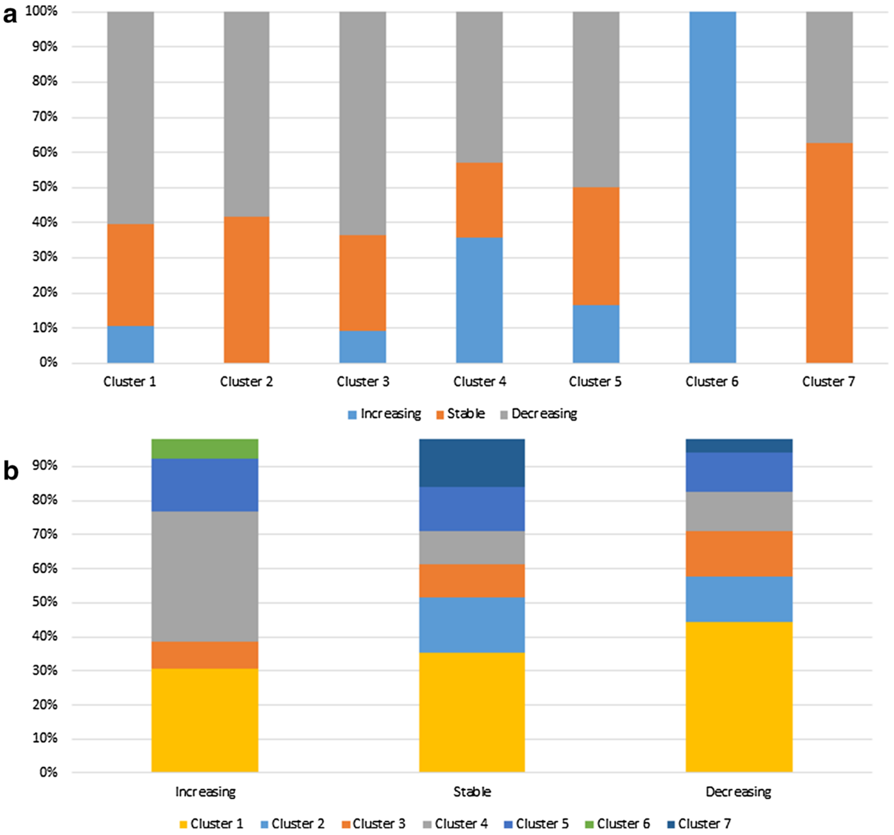

Fig. 6 **a** Distributions of population trends indicated in the IUCN Red List for the benchmark species over the automatically detected clusters. **b** Distribution of the automatically detected clusters over IUCN population trends

3.4 Correspondence between automatic classification and literature studies

In this section, a sample of species from the benchmark dataset is taken to show correspondence between automatic classification and biological/ecological studies. This section does not aim to formally confirm the habitat change indicated by the clustering analysis. Rather, it gives hints about how much this analysis goes in the direction indicated by experts. The AquaMaps distributions used in the reported experiment are sub-optimal, thus only overall changes in the map trends and clusters indications have been considered in the following analysis.

Cluster 2 (strong habitat variations) contains endangered and nearly threatened species, for instance *Actinopterygii* and *Elasmobranchii* Classes. Some species in these classes could benefit from larger habitats as reported by positive trend variation in the AquaMaps distribution in 2050. One example of agreement with this prediction can be found for sturgeons (*Actinopterygii* Class), species having unique life history characteristics and very sensitive to environmental changes. Studies have reported that these fishes are mainly threatened by anthropogenic pressures (e.g. infrastructures and fishing), although researches have also shown severe consequences on their physiological traits due to changes in abiotic conditions (Knights 2003; Sardella et al. 2008). Sturgeons are distributed over several clusters, which means that their habitats can present heterogeneous modifications due to climate change. As reported by Lassalle and Rochard (2009), factors such as temperature and precipitations can be relevant for diadromous species, which have shown shifting ranges north-eastwards depending on these factors. For example, *Anguilla japonica* (*Actinopterygii* Class) might be influenced by this kind of climate variability (Tzeng et al. 2012). In particular, this species has complex life history traits and its recruitment fluctuation may be driven by non-linear combinations of climatic factors, such as sea water temperature, salinity and precipitations. The zebra bullhead shark (*Heterodontus zebra*) is the only species in cluster 2 having lower habitat suitability in 2050. Its distributions are shown in Fig. 7. AquaMaps estimates that this species will likely disappear in Indonesia, mostly due to a combination of changes in temperature and salinity. Indeed, the sensitivity of this species to climate change has recently been demonstrated, with poleward shifts and increase of species abundance in Tasmanian and Bahamas waters (BEST Commission 2003). Climate change, in fact, affects the species overall biogeographic affinity, negatively influencing its breeding practices and habitat selection (Barratt and Cavanagh 2015).

The overall expansion of suitable habitat detected by the analysis in Sect. 3.3, can be supported by bioclimatic theories correlating temperature increase with food availability (Genner et al. 2004). In particular, many benchmark species have temperature-dependent feeding behaviour. Therefore, the expansion range detected in the analysis can be related to a poleward increase of primary productivity and

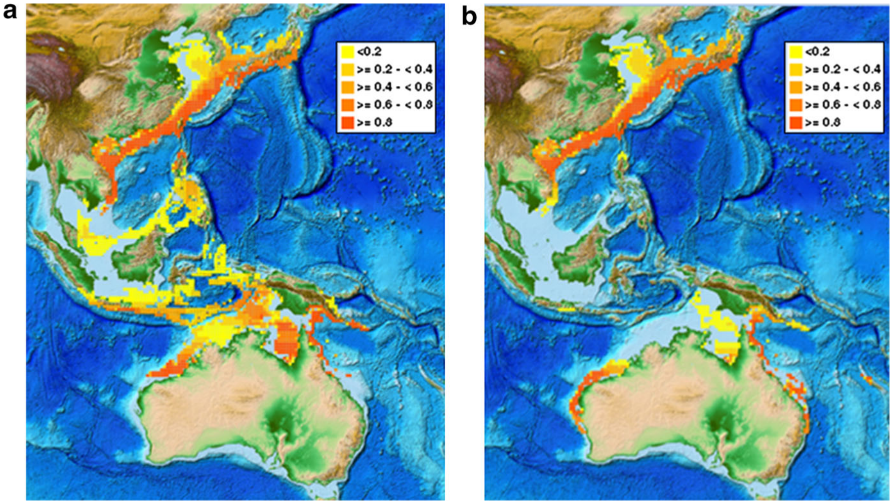

Fig. 7 AquaMaps Native (**a**) and Native-2050 (**b**) distributions of the zebra bullhead shark (*Heterodontus zebra*). *Darker colours* indicate higher probability areas. A strong reduction of habitat suitability is estimated for 2050

consequently to food availability (Arrigo et al. 2008; Brierley and Kingsford 2009). Several empirical and theoretical studies have identified demographic shifts of marine fishes and invertebrates from North Sea (Perry et al. 2005) and Mid-Atlantic Bight (Nye et al. 2009) to higher latitudes and/or to deep seas (Mueter and Litzow 2008). One example is the blue whiting (*Micromesistius poutassou*): according to the AquaMaps distribution change, the range of this species could increase in 2050. This prediction agrees with evidence of distribution shifts westwards and distribution shrinking when temperature decreases (Heino et al. 2008). A more complex example is that of anadromous species, which show heterogeneous patterns in their distribution over the clusters. Anadromous fishes can be very sensitive to climatic variations, because they may occupy different habitats during their life cycles. For example, these species move to find suitable salinity and temperature gradients for reproducing, e.g. near estuaries (Sardella et al. 2008). In some cases, the heterogeneity in the clusters agrees with biological studies: according to the AquaMaps distributions, the green sturgeon (*Acipenser medirostris*) presents a demographic shift to northern latitudes, whereas the atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*) shows marked local unsuitability in 2050 in the Equatorial and North Temperate zones. These anadromous species have different sensitivity and physiological tolerance to abiotic factors, even if they belong to the same Genus. In particular, variations in salinity gradient within the estuaries may have consequences on the species physiological traits, including their thermal tolerance (Lassalle and Rochard 2009; Sardella et al. 2008).

Confirmation of predicted shifts to warmer locations can be found for two cartilaginous fishes: *Orectolobus ornatus* and *Prionace glauca*. Ecological studies have confirmed that these species have moderate vulnerability to abiotic changes (Chin et al. 2010). According to the reported analysis, these will likely have larger habitat suitability in 2050 with local expansions. Indeed, habitat shift for mobile reef sharks, such as *Orectolobus ornatus*, is often correlated with reef habitats degradation and food availability, which are (indirectly) taken into account by the AquaMaps model.

Evidence of agreement between the clustering analysis and biological studies can be found also for those species whose habitat shrinks in 2050. For example, *Alopias vulpinus* belongs to cluster 3 (global scale distributed with small localized changes) and thus will likely have its habitat decreased in localized zones (particularly in the Equatorial zone). Indeed, recent observations in United Kingdom waters and in the North Sea have positively correlated decreasing presence of *A. vulpinus* with increasing sea water temperature (Cao et al. 2011). A similar case is the surf clam (*Spisula solidissima*), in cluster 1 (localized moderate probability changes), which presents localized-moderate reduction of its estimated habitat. For this species, empirical studies near the U.S. coasts have highlighted that thermal stress may contribute to their decline in the Mid-Atlantic Bight (Marzec et al. 2010).

Regarding the marine mammals in Table 6, the predicted habitat shrink can be confirmed by major vulnerabilities related to food resources. It is very difficult to correlate changes in the distribution of this group as direct effect of climatic changes. Indeed, marine mammals live in a wide range of latitude encompassing large ranges of water temperatures. The AquaMaps distributions of the baleen whale (*Balaenoptera musculus*), which falls in cluster 3, shows habitat changes in the polar regions and in the Indian and Pacific oceans. These are due to changes in the temperature and ice concentration parameters, but especially in the primary production parameter of the AquaMaps 2050 scenario. Indeed, the estimated trend and the cluster classification agree with indications by biological studies (Lawler et al. 2007). For another marine mammal, *Pusa hispida*, belonging to cluster 2 (strong overall change), AquaMaps 2050 shows a shift towards higher latitudes and unsuitability in North Alaska and in the Bering Sea. This agrees with the fact that sea ice represents their reference for breeding and hunting, thus sea ice loss affects their life history events and shrinks their distribution (Carlens et al. 2006). As final example, the predicted habitat loss of the Gray's beaked whale (*Mesoplodon grayi*), in cluster 4 (moderated overall changes), is confirmed by biological studies to depend on temperature increase. Beaked whales are known to be threatened by anthropogenic activity but also to be sensitive to sea water temperature variation (Thompson et al. 2013). MacLeod (2009) highlighted that these marine mammals have upper and lower temperature limits to their tolerance range and consequently are confined to waters of intermediate temperature. Hence, species' ranges will unlikely increase overall with sea water temperature, as reported by AquaMaps 2050 and cluster 4.

3.5 Detection of anomalies

The experiment described in this paper used sub-optimal datasets. Nevertheless, patterns highlighted by the proposed method would probably be valid also if the data had been manually reviewed. For example, the fact that most species will likely have their habitat suitability increased is a generally valid property and is also in agreement with other studies. Furthermore, examples of species have been shown, for which the forecast by the reported analysis is compliant with their observed attitude towards climate change. On the other hand, the role of the proposed process is also to inform scientists about possible anomalies in the data. Anomalies can be systematic errors or noise introduced by missing manual review of the maps after their production. For example, this is the case of species having narrow distribution (delimited by a less than $40^\circ \times 40^\circ$ square area), on which the AquaMaps experts usually debate. The problem is that AquaMaps usually estimates larger delimiting geographic bounding boxes in 2050 for these species. Thus, it is not clear if a larger habitat suitability in 2050 is simply due to a larger bounding box or to a concrete increase of habitat suitability. On the other hand, there is a number of these small-range species that instead present a habitat reduction trend. For example, *Phocoena sinus*, *Crangon crangon*, *Dentex congoensis* and *Heterodontus japonicus* have geographical bounding boxes with square side lower than 40°, but will likely have their habitat suitability reduced in 2050. Thus, it was not possible to *a priori* delete all narrow bounding box species from the analysis. On the other hand, the information included in the output dataset allows filtering out these species after the processing: for example, small scale distributions can be identified based on their belonging cluster definition and on the number of point-to-point map comparisons.

4 Conclusions

In this paper, differences between current and future 2050 habitat suitability estimations have been analysed. In the reported experiment, clustering analysis was used to detect a discrete number of categories identifying local, regional and global habitat distribution changes. The reliability of this result was evaluated using human assessment. A computational and data e-Infrastructure (D4Science) was used, which hosts generated models as GIS maps, to perform a large number of maps comparisons and to produce statistics. The distributions of the IUCN categories and of different taxonomic groups over the habitat distribution change categories were explored. No relevant correlation between the detected categories and the IUCN classifications was found, whereas few correlations were detected for taxonomic groups. On the other hand, agreement was found between short-term population trends indicated by IUCN and long-term estimated habitat trends indicated by the clustering analysis. This agreement is lower for species whose habitat will shrink or not change according to IUCN, but one possible explanation of this is that climate change is only one of the factors, such as over-exploitation, habitat destruction, pollution, invasive species etc., that may affect the population and distribution of endangered species. Furthermore, a general expansion trend of habitat suitability in 2050 was observed, with few remarkable exceptions. In particular, a general habitat expansion towards the poles was highlighted. Additionally, specific cases were shown where the overall habitat suitability increase was accompanied by unsuitability in limited areas.

The proposed analysis relies on the AquaMaps models, and uses powerful technologies like Cloud computing, machine learning and use of data descriptive standards to make biological assessments. Using the approach presented in this paper, biologists can further refine their models and manually revise groups of species distributions. Generally speaking, the described work is close to Big Data analysis approaches (Assunçaoa et al. 2013; Campbell 2008), since around 100 GB of data were processed using Cloud computing. As in Big Data analysis, data mining techniques were applied to extract information from noisy data. Furthermore, unsupervised data mining techniques were used, because a complete preliminary manual review would have been unpractical on the used large benchmark dataset. In the case of the presented experiment, noise (and possible systematic errors) was represented by missing human review of data and properties were extracted that would have been valid also without noise. Furthermore, different categories of data were detected, to help experts later focus on certain subsets.

One of the advantages of the presented method is also that it fits with modern requirements of Science, mostly advised by novel Science paradigms (Hey et al. 2009; Waldrop 2008): (1) it facilitates human review of data, (2) it is repeatable and reproducible, (3) it can be directly applied to other data.

Acknowledgments The reported work has been partially supported by the i-Marine project (FP7 of the European Commission, INFRASTRUCTURES-2011-2, Contract No. 283644) and by the Giovanisi project of the Presidency of the Tuscan Regional Government.

References

Anthony K, Maynard JA, Diaz-Pulido G, Mumby PJ, Marshall PA, Cao L, Hoegh-Guldberg O (2011) Ocean acidification and warming will lower coral reef resilience. Glob Change Biol 17(5):1798–1808

Araujo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. J Biogeogr 33(10):1677–1688

Araújo MB, Pearson RG, Thuiller W, Erhard M (2005) Validation of species-climate impact models under climate change. Glob Change Biol 11(9):1504–1513

Arrigo KR, van Dijken GL, Bushinsky S (2008) Primary production in the Southern Ocean, 1997–2006. J Geophys Res Oceans (1978–2012) 113(C8):15587–15600

Assunçaoa MD, Calheirosb RN, Bianchia S, Nettoa MA, Buyyab R (2013) Big Data computing and clouds: challenges, solutions, and future directions. arXiv:1312.4722

Barratt P, Cavanagh RD (2015) *Heterodontus zebra* in the IUCN red list of threatened species. Version 2014.3, www.iucnredlist.org

Bellwood D, Hughes T, Folke C, Nyström M (2004) Confronting the coral reef crisis. Nature 429(6994):827–833

Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18(9):509–517

Berry P, Dawson T, Harrison P, Pearson R (2002) Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. Glob Ecol Biogeogr 11(6):453–462

BEST Commission (2003) The national invasive species strategy for the Bahamas. BEST, Nassau, The Bahamas 40

Botkin DB, Saxe H, Araujo MB, Betts R, Bradshaw RH, Cedhagen T, Chesson P, Dawson TP, Etterson JR, Faith DP, Ferrier S, Guisan A, Skjoldborg Hansen A, Hilbert DW, Loehle C, Margules C, New M, Sobel MJ, Stockwell DRB (2007) Forecasting the effects of global warming on biodiversity. Bioscience 57(3):227–236

Brierley AS, Kingsford MJ (2009) Impacts of climate change on marine organisms and ecosystems. Curr Biol 19(14):R602–R614

Campbell P (2008) Editorial on special issue on big data: community cleverness required. Nature 455(7209):1

Candela L, Castelli D, Coro G, Lelii L, Mangiacrapa F, Marioli V, Pagano P (2014) An infrastructure-oriented approach for supporting biodiversity research. Ecol Inform 26:162–172. doi:10.1016/j.ecoinf.2014.07.006

Cao D, Song L, Zhang Y, Lv K, Hu Z (2011) Environmental preferences of Alopias superciliosus and Alopias vulpinus in waters near Marshall Islands. N Z J Mar Freshw Res 45(1):103–119

Carlens H, Lydersen C, Krafft BA, Kovacs KM (2006) Spring haul-out behavior of ringed seals (Pusa hispida) in Kongsfjorden. Svalbard. Mar Mamm Sci 22(2):379–393

Castelli D, Michel J (2011) D4SCIENCE-II—data infrastructures ecosystem for science. Project final report. Data Infrastructures Ecosystem for Science. Deliverable DNA1.7

Castelli D, Taconet M, Garavelli S, Parker S (2013) iMarine infrastructure for data driven decision making and research: position paper. Presentation at the iMarine e-infrastructure Workshop for data-driven decision making and research, 14–15 May, Brussels, Belgium

Cheung WW, Lam VW, Sarmiento JL, Kearney K, Watson R, Pauly D (2009) Projecting global marine biodiversity impacts under climate change scenarios. Fish Fish 10(3):235–251

Cheung WW, Lam VW, Sarmiento JL, Kearney K, Watson R, Zeller D, Pauly D (2010) Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. Glob Change Biol 16(1):24–35

Cheung WWL, Dunne J, Sarmiento JL, Pauly D (2011) Integrating ecophysiology and plankton dynamics into projected maximum fisheries catch potential under climate change in the Northeast Atlantic. ICES J Mar Sci 68(6):1008–1018

Chin A, Kyne PM, Walker TI, McAULEY R (2010) An integrated risk assessment for climate change: analysing the vulnerability of sharks and rays on Australia's Great Barrier Reef. Glob Change Biol 16(7):1936–1953

Chuine I, Beaubien EG (2001) Phenology is a major determinant of tree species range. Ecol Lett 4(5):500–510

CNR T (2015) The gCube GeoExplorer. https://gcube.wiki.gcube-system.org/gcube/index.php/GeoExplorer

Cohen J et al (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Coro G, Gioia A, Pagano P, Candela L (2013) A service for statistical analysis of marine data in a distributed e-infrastructure. Boll di Geofis Teorica e Appl 54(1):68–70

Coro G, Pagano P, Ellenbroek A (2014) Comparing heterogeneous distribution maps for marine species. GISci Remote Sens 51(5):593–611

Coro G, Candela L, Pagano P, Italiano A, Liccardo L (2015) Parallelizing the execution of native data mining algorithms for computational biology. Concurr Comput Pract Exp 27(17):4630–4644. doi:10.1002/cpe.3435

Corsi F, de Leeuw J, Skidmore A (2000) Modeling species distribution with GIS. Research Techniques in Animal Ecology, Columbia University Press, New York, pp 389–434

Costa GC, Nogueira C, Machado RB, Colli GR (2010) Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. Biodivers Conserv 19(3):883–899

Dawson TP, Jackson ST, House JI, Prentice IC, Mace GM (2011) Beyond predictions: biodiversity conservation in a changing climate. Science 332(6025):53–58

de La Beaujardière J (2004) OGC Web Map Service Interface, version 1.3.0. Open Geospatial Consortium

Dulvy NK, Rogers SI, Jennings S, Stelzenmüller V, Dye SR, Skjoldal HR (2008) Climate change and deepening of the North Sea fish assemblage: a biotic indicator of warming seas. J Appl Ecol 45(4):1029–1039

Erez J, Reynaud S, Silverman J, Schneider K, Allemand D (2011) Coral calcification under ocean acidification and global change. In: Dubinsky Z, Stambler N (eds) Coral reefs: an ecosystem in transition. Springer, Netherlands, pp 151–176

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, pp 226–231

FAO (2015) Fact sheets. http://www.fao.org/newsroom/en/facts/index.html

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

GCube (2015) The GCube Featherweight Stack. http://gcube.wiki.gcube-system.org/gcube/index.php/Featherweight_Stack

Genner MJ, Sims DW, Wearmouth VJ, Southall EJ, Southward AJ, Henderson PA, Hawkins SJ (2004) Regional climatic warming drives long-term community changes of British marine fish. Proc R Soc Lond Ser B Biol Sci 271(1539):655–661

Hare JA, Alexander MA, Fogarty MJ, Williams EH, Scott JD (2010) Forecasting the dynamics of a coastal fishery species using a coupled climate-population model. Ecol Appl 20(2):452–464

Heino M, Engelhard GH, Godo OR (2008) Migrations and hydrography determine the abundance fluctuations of blue whiting (Micromesistius poutassou) in the Barents Sea. Fish Oceanogr 17(2):153–163

Hey AJ, Tansley S, Tolle KM et al (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft Research Redmond, WA

Hiddink J, Ter Hofstede R (2008) Climate induced increases in species richness of marine fishes. Glob Change Biol 14(3):453–460

Hsieh CH, Kim HJ, Watson W, Di Lorenzo E, Sugihara G (2009) Climate-driven changes in abundance and distribution of larvae of oceanic fishes in the southern California region. Glob Change Biol 15(9):2137–2152

Hughes TP, Bellwood DR, Folke C, Steneck RS, Wilson J (2005) New paradigms for supporting the resilience of marine ecosystems. Trends Ecol Evolut 20(7):380–386

Hyrenbach KD, Veit RR (2003) Ocean warming and seabird communities of the southern California Current System (1987–98): response at multiple temporal scales. Deep Sea Res Part II Top Stud Oceanogr 50(14):2537–2565

IUCN (2015) The IUCN Red List of species. www.iucnredlist.org

Jordà G, Marbà N, Duarte CM (2012) Mediterranean seagrass vulnerable to regional climate warming. Nat Clim Change 2(11):821–824

Kaschner K, Watson R, Trites A, Pauly D (2006) Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. Mar Ecol Prog Ser 316:285–310

Knights B (2003) A review of the possible impacts of long-term oceanic and climate changes and fishing mortality on recruitment of anguillid eels of the Northern Hemisphere. Sci Total Environ 310(1):237–244

Koch M, Bowes G, Ross C, Zhang XH (2013) Climate change and ocean acidification effects on seagrasses and marine macroalgae. Glob Change Biol 19(1):103–132

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 1:159–174

Lassalle G, Rochard E (2009) Impact of twenty-first century climate change on diadromous fish spread over europe, north africa and the middle east. Glob Change Biol 15(5):1072–1089

Lawler IR, Parra G, Noad M (2007) Vulnerability of marine mammals in the Great Barrier Reef to climate change. http://hdl.handle.net/11017/548

Levin SA, Lubchenco J (2008) Resilience, robustness, and marine ecosystem-based management. Bioscience 58(1):27–32

MacLeod CD (2009) Global climate change, range changes and potential implications for the conservation of marine cetaceans: a review and synthesis. Endanger Species Res 7(2):125–136

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 14. California, USA, pp 281–297

Marzec RJ, Kim Y, Powell EN (2010) Geographical trends in weight and condition index of surfclams (Spisula solidissima) in the Mid-Atlantic Bight. J Shellfish Res 29(1):117–128

Mueter FJ, Litzow MA (2008) Sea ice retreat alters the biogeography of the Bering Sea continental shelf. Ecol Appl 18(2):309–320

Nakicenovic N, Swart R (2000) Special report on emissions scenarios. Special report on emissions scenarios, In: Nakicenovic N, Swart R (ed) pp 612 ISBN 0521804930, Cambridge University Press, Cambridge, UK, 1 July 2000

Nye JA, Link JS, Hare JA, Overholtz WJ (2009) Changing spatial distribution of fish stocks in relation to climate and population size on the Northeast United States continental shelf. Mar Ecol Prog Ser 393:111–129

Parmesan C (2006) Ecological and evolutionary responses to recent climate change. Annu Rev Ecol Evolut Syst 1:637–669

Pearson RG (2012) Species distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. http://ncep.amnh.org

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob Ecol Biogeogr 12(5):361–371

Pelleg D, Moore AW (2000) X-means: extending K-means with efficient estimation of the number of clusters. In: ICML, pp 727–734

Pellissier L, Bråthen KA, Vittoz P, Yoccoz NG, Dubuis A, Meier ES, Zimmermann NE, Randin CF, Thuiller W, Garraud L, Van Es J, Guisan A (2013) Thermal niches are more conserved at cold than warm limits in arctic-alpine plant species. Glob Ecol Biogeogr 22(8):933–941

Perry AL, Low PJ, Ellis JR, Reynolds JD (2005) Climate change and distribution shifts in marine fishes. Science 308(5730):1912–1915

Ready J, Kaschner K, South AB, Eastwood PD, Rees T, Rius J, Agbayani E, Kullander S, Froese R (2010) Predicting the distributions of marine organisms at the global scale. Ecol Model 221(3):467–478. doi:10.1016/j.ecolmodel.2009.10.025

Reyes K (2015) AquaMaps: algorithm and data sources for aquatic organisms. http://www.aquamaps.org/main/FB_Book_KReyes_AquaMaps_JG.pdf

Roeckner E, Arpe K, Bengtsson L, Brinkop S, Dümenil L, Esch M, Kirk E, Lunkeit F, Ponater M, Rockel B et al (1992) Simulation of the present-day climate with the ECHAM model: impact of model physics and resolution. Max-Planck-Institut für Meteorologie, Hamburg

Sardella BA, Sanmarti E, Kültz D (2008) The acute temperature tolerance of green sturgeon (Acipenser medirostris) and the effect of environmental salinity. J Exp Zool Part A Ecol Genet Physiol 309(8):477–483

Schwartz MW (2012) Using niche models with climate projections to inform conservation management decisions. Biol Conserv 155:149–156

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Somero GN (2012) The physiology of global change: linking patterns to mechanisms. Ann Rev Mar Sci 4:39–61

Sorvari S, Brus M (2012) ENVRI—overcoming the environmental challenges with common solutions. EPOS newsletter, February

Sunday JM, Bates AE, Dulvy NK (2012) Thermal tolerance and the global redistribution of animals. Nat Clim Change 2(9):686–690

Sydeman W, García-Reyes M, Schoeman D, Rykaczewski R, Thompson S, Black B, Bograd S (2014) Climate change and wind intensification in coastal upwelling ecosystems. Science 345(6192):77–80

Thompson KF, Millar CD, Baker CS, Dalebout M, Steel D, van Helden AL, Constantine R (2013) A novel conservation approach provides insights into the management of rare cetaceans. Biol Conserv 157:331–340

Thuiller W (2004) Patterns and uncertainties of species' range shifts under climate change. Glob Change Biol 10(12):2020–2027

Thuiller W, Lavorel S, Araújo MB, Sykes MT, Prentice IC (2005) Climate change threats to plant diversity in Europe. Proc Natl Acad Sci USA 102(23):8245–8250

Tzeng WN, Tseng YH, Han YS, Hsu CC, Chang CW, Di Lorenzo E, Hsieh Ch (2012) Evaluation of multi-scale climate effects on annual recruitment levels of the Japanese eel, Anguilla japonica, to Taiwan. PLoS One 7(2):e30,805

Waldrop MM (2008) Science 2.0. Sci Am 298(5):68–73

Wassmann P, Duarte CM, Agusti S, Sejr MK (2011) Footprints of climate change in the Arctic marine ecosystem. Glob Change Biol 17(2):1235–1249

Wootton JT, Pfister CA, Forester JD (2008) Dynamic patterns and ecological impacts of declining ocean pH in a high-resolution multi-year dataset. Proc Natl Acad Sci 105(48):18,848–18,853

Gianpaolo Coro is a Physicist with a Ph.D. in Computer Science. His research focuses on Artificial Intelligence and Data Mining. He has been working on Machine Learning and Signal Processing with applications to Computational Biology, Brain Computer Interfaces, Language Technologies and Cognitive Sciences. The aim of his research is the study and experimentation of models and methodologies to process biological data and to apply the results to fields in Ecological Modelling, Vessel Monitoring Systems and Ecological Niche Modelling. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Grid and Cloud based technologies.

Chiara Magliozzi is a Environmental Scientist and she is interested in the ability of natural systems to deliver ecosystem services. She is Ph.D. student at Cranfield University in collaboration with the River Restoration Centre. Her research concerns river restoration measures and their impact on the hyporheic zone. In particular, her work focuses on assessing the effectiveness of physical restoration under varying hyporheic conditions for ecological and habitat improvement.

Anton Ellenbroek joined FAO in June 2000, and the Fisheries department 28 in 2010. He is based at the FAO HQ in Rome, Italy. FAO leads iMarine Governance and Policy Development, and Anton coordinates the activities of the project related to the Community behind Ecosystem Approach to Fisheries and the Conservation of Marine Living Resources. In FAO, he has held data management positions for a large number of projects to deliver solutions for data collection, analysis, and dissemination. Before joining FAO in 2000, he was an Oracle information analyst and database developer based in Amsterdam. From 1992 to 1995 he worked with IUCN-NL on a variety of data projects, such as the Netherlands and the World Ecology, and the Ecological Footprint of the Netherlands.

Kristin Kaschner is a marine scientist affiliated to the Department of Biometry and Environmental Systems Analysis of the Albert-Ludwigs University in Freiburg, Germany. Her main research interests are global conservation, biogeography and macroecology of marine organisms, anthropogenic impacts on marine mammals, spatial planning, marine protected areas, species distribution modeling and meta-analyses. She is the main developer of the AquaMaps species distribution model.

Pasquale Pagano Senior Researcher at the Networked Multimedia Information Systems Laboratory of the "Istituto di Scienza e Tecnologie della Informazione A. Faedo" (ISTI) of the Italian National Research Council (CNR). The aim of his research is the study and experimentation of models, methodologies and techniques for the design and development of distributed virtual research environments (VREs), which require the handling of heterogeneous resources—data, computing and storage—accessed via Grid and Cloud based technologies. He has a background in distributed computing and data integration, harmonization, and access.