# Improving data quality to build a robust distribution model for *Architeuthis dux*

Gianpaolo Coro[a,1,2,*], Chiara Magliozzi[a], Anton Ellenbroek[b], Pasquale Pagano[a]

[a] *Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa, Italy*

[b] *Food and Agriculture Organization of the United Nations (FAO)*

## Abstract

The giant squid (*Architeuthis*) has been reported since even before the 16th century, and has recently been observed live in its habitat for the first time. Among the species belonging to this genus, *Architeuthis dux* has received special attention from biologists. The distribution of this species is poorly understood, as most of our information stems from stranded animals or stomach remains. Predicting the habitat and distribution of this species, and more in general of difficult to observe species, is important from a biological conservation perspective. In this paper, we present an approach to estimate the potential distribution of *A. dux* at global scale, with relative high resolution (1-degree). Our approach relies on a complex preparation phase, which improves the reliability of presence, absence and environmental data correlated to the species habitat. We compare our distribution with those produced by state-of-the-art approaches (MaxEnt and AquaMaps), and use an expert-drawn map as reference. We demonstrate that our model projec-

[*] Corresponding author

*Email addresses:* coro@isti.cnr.it (Gianpaolo Coro), chiara.magliozzi@isti.cnr.it (Chiara Magliozzi), Anton.Ellenbroek@fao.org (Anton Ellenbroek), pagano@isti.cnr.it (Pasquale Pagano)

[1] Telephone Number: +39 050 315 2978

[2] Fax Number: +39 050 621 3464

tion is in agreement with the expert's map and is also compliant with several biological assessments of the species habitat and with recent observations. Furthermore, we show that our approach can be generalized as a paradigm that is applicable to other rare species.

*Keywords:* Ecological Niche Modelling, AquaMaps, Neural Networks, rare species, Maximum Entropy

## 1. Introduction

In recent years, niche models that estimate species distribution have become widely used in conservation biology (Guisan and Zimmermann, 2000). Rare species are examples where the prediction of suitable habitats is paramount to support fisheries management policies and conservation strategies (Pearce and Boyce, 2006; Márcia Barbosa et al., 2003). Defined by Cao et al. (Cao et al., 1998) as species that occur at lower frequency or in low number in a sample of certain size, rare species have a key role in affecting biodiversity richness and by consequence they are indicators of degradation for aquatic ecosystems (Lyons et al., 1995; Cao et al., 1998). In this context, predictive models can considerably support the qualitative and quantitative criteria used to assign a "status" to a species (IUCN Species Survival Commission and Natural Resources. Species Survival, 2001), by providing accurate, applicable and reliable spatial predictions to species population monitoring and sampling (Guisan et al., 2006). As discussed in many studies, the methodological progresses of Species Distribution Models (SDMs) allow nowadays to apply robust techniques to rare and endangered species (Guisan and Thuiller, 2005; Ferrier, 2002; Gibson et al., 2007; Razgour et al., 2011; Ovaskainen and Soininen, 2011; Rebelo and Jones, 2010; Wisz et al., 2008; Lomba et al., 2010).

Here, we propose a procedure to generate a niche model for a species of the giant squid family (*Architeuthis dux*), based on both presence and estimated absence locations. Our aim is to produce a map that is more accurate with respect to the ones that can be produced by commonly used models. Although giant squids have recently received special attention, little has been published regarding the population demographics and the ecology of these rare species. Most of the records refer to dead stranded animals, individuals captured alive by nets or from the remains found in the stomach of marine mammals (Clarke, 2006). When modelling the distribution of these species, high quality data are crucial but very scarce. This problem is especially important for rare species prediction, where models training is highly dependent on data quality.

Given this context, our study investigates a combination of presence only and presence/absences techniques to identify potentially suitable areas for *A. dux* subsistence. We also expect the results to help defining guidelines for use of SDMs for rare species.

We illustrate our approach using data from authoritative sources of observation records. Furthermore, we use an expert system to produce absence locations. In order to ensure high quality for the environmental variables associated to presence information, we use the Maximum Entropy (MaxEnt) model (Phillips et al., 2006; Berger, 1996) as a filter to select the variables that are important to define the potential habitat of the species. These are the variables that are mostly correlated to the species observations, among those we selected from reference studies. When possible, we make environmental variables values range from 450 to 1000 m, encompassing the deep ocean waters usually inhabited by A.dux (Guerra et al., 2010). Finally, we train an Artificial Neural Network on these datasets and compare the results with (i) a presence-only method, (ii) an expert system and (iii) an expert drawn map.

The paper is organized as follows: Section 2 reports the effort made to model or understand the potential habitat of rare species, and in particular of *A. dux.* Section 3 reports the details of our method and its expandability as a general approach to rare species modelling. Section 4 reports the results of both a qualitative and a quantitative comparison with other distribution maps for *A. dux.* Section 5 discusses the results and Section 6 draws the conclusions.

## 2. Overview

This Section is divided into two subsections. The first reports the current understanding of the distribution of *Architeuthis dux.* The second describes the niche modelling approaches that have been applied or that can be applied to rare species.

### *2.1. Species overview*

The *Architeuthis* genus has been recorded since before the 16th century (Guerra et al., 2011), and has recently been observed live in its natural habitat for the first time (Kubodera and Mori, 2005). Literature studies have recognized up to five species of this genus (Robson, 1933), although Nesis (Nesis, 1987) and Aldrich (Aldrich, 1991) suggested them to be identified as *Architeuthis dux*. Most of the records refer to stranded animals or stomach remains, and are located in the North Atlantic (e.g. Norway), in the North-East Atlantic (off northern Spain), in the South Atlantic (e.g. Namibia and South Africa) and in the South-West Pacific, around New Zealand and Tasmania (Gonzalez et al., 2000; Clarke, 2006; Förch, 1998; Guerra et al., 2004; Bolstad and O'Shea, 2004; Guerra et al., 2004). Most of these animals have been classified as *A. dux* (Cherel, 2003; Clarke, 2006; Bolstad and O'Shea, 2004; Guerra et al., 2010; Clarke, 2006; Nesis, 2003; Aldrich, 1991), but many more refer to the genus level (*Architeuthis* spp.) without further specification (Lordan et al., 1998; Gonzalez et al., 2000; Ré et al., 1998; Arfelli et al., 1991; Kubodera and Mori, 2005; Roeleveld and Lipinski, 1991). In 2003, Nesis (Nesis, 2003) published the distribution of *Architeuthis dux* by correlating latitudinal zones and zoogeographic provinces in the pelagic realm. The identified zonality mainly reflects the general oceanic circulation, and no temperature data was used for the selection of the latitudinal zones. The author identified rate of speciation among the Cephalopoda taxon caused by climatic and orogenic isolation and bi-subtropical species of *Architeuthis dux* in the North Atlantic, the South Pacific and the Southern Ocean. In this paper, we take the map of Nesis as a reference to assess the performance of our models.

Several authors have suggested that *Architeuthis* is an epipelagic/mesopelagic species, living in correspondence with continental slopes, submarine channels or canyons (Roeleveld and Lipinski, 1991; Kubodera and Mori, 2005). Guerra et al. (Guerra et al., 2011) examined the relationship between the number of recorded specimens and some of the main characteristics of the observation areas. The authors report the close association of giant squids with sperm whales sights (Clarke and Pascoe, 1997). They indicate correlation of *Architeuthis* spp. sighting with places presenting high primary production and close to shallow fishing grounds. They also report low incidence of genus sighting, in locations where deep channels or canyons are not present (Guerra et al., 2004). On the basis of the distribution of the strandings, Robson (Robson, 1933) noticed that *Architeuthis* is adapted to temperate waters of about 10 °C. This biological information is in agreement with later studies, that correlate the giant squid presence with the increase of the temperature in some locations (Brix, 1983; Guerra et al., 2004).

In this paper, we demonstrate that our results are in agreement with most of these considerations.

### *2.2. Modelling approaches*

SDMs produce species distributions at global or local scale, by relating species occurrence records with a set of environmental parameters. Many methods are available (Pearson, 2012), some using only presence records and others using both presence and absence records (Ready et al., 2010; Coro et al., 2013b; Guisan and Zimmermann, 2000; Hirzel and Le Lay, 2008). Niche models usually report either the potential or the actual distribution of a species (Elith and Leathwick, 2009; Pearson, 2012). In the case of the potential distribution, the model searches for locations with a suitable habitat, rather than detecting locations where the species is really present (actual distribution).

Presence-absence methods have been recognized to be the best in producing the potential niche of a species, especially for wide-ranging and tolerant species when the quality of the data is high (Elith and Leathwick, 2009; Brotons et al., 2004). Nevertheless, scarcity of data is a common issue when modelling rare species: few records are present in biodiversity databases, and often scarce in both quality and geospatial reliability (Engler et al., 2004). Providing reliable presence and absence data, enhances the performance of niche models (Guisan and Zimmermann, 2000). However, the identification of absences should be carefully addressed, since they bear strong imprints of biotic interactions, dispersal constraints and disturbances (Pulliam, 2000; Gibson et al., 2007; Hirzel and Le Lay, 2008; Cianfrani et al., 2010).

In this paper, we use different approaches to model the potential distribution of *A. dux*. We take the AquaMaps expert system as reference for the comparison. The AquaMaps algorithms (Kaschner et al., 2006, 2008) are presence-only models that include scientific expert knowledge into species habitats modelling (Ready et al., 2010). The AquaMaps algorithms include two models: AquaMaps Suitable and AquaMaps Native, addressing the potential and the actual distribution of a species respectively. 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). AquaMaps combines mechanistic assumptions and automatic procedures for habitat parameters and species values estimations, making the modelling approach usually reliable, but less accurate when expert knowledge at global scale is missing. In the experiment for this paper, we used AquaMaps Native to produce absence locations and AquaMaps Suitable as reference to assess the performance of the other models.

One largely used presence-only technique is Maximum Entropy (MaxEnt) (Phillips et al., 2006; Phillips and Dudik, 2008). The general idea of MaxEnt is to approximate a probability density function, defined on an environmental features vectorial space, ensuring that this function is compliant with the mean values at the presence locations, and that the entropy of the probability distribution is maximum (Elith et al., 2011). The algorithm relies on unbiased samples, so effort in collecting a set of high quality presence records is critical to avoid estimation errors (Elith and Leathwick, 2009). We used MaxEnt as a reference model to assess the performance of our approach. On the other hand, MaxEnt is a fundamental part of our approach, because we used it to help a presence-absence model by providing features that are important to assess habitat suitability. We give more details about our MaxEnt usage in Section 3.3.

Among the many presence/absence models, Artificial Neural Networks (ANNs) have demonstrated to gain good performance with respect to other approaches, especially for rare species (Pearson et al., 2002; Coro et al., 2013b). ANNs try to automatically simulate the probability of occurrence of a species, given certain environmental conditions. They learn on the basis of the environmental characteristics of positive and negative examples. We used ANNs to combine the outputs of our presence/absence data production and of the environmental features filtering phase. In Section 3.5 we give details about our usage of ANNs.

## 3. Method

In this Section we describe the technology which supported the experiments, and we also report our procedures for data preparation and environmental features selection. Furthermore, we explain our presence/absence approach to model the distribution of *A.dux* and its relevance for other rare species.

### 3.1. Technology and tools

Preparing an experimental setup to model the distribution of a rare species requires expertise in several disciplines. The model requires highly reliable presence records. The environmental features describing the ecological niche of the species should be of high quality and with the appropriate spatial resolution (Kamino et al., 2012; Elith and Leathwick, 2009). Since environmental features are distributed as geospatial datasets, their projections should be perfectly aligned in order to correctly retrieve correspondent values. During the training phase, different models need to be tested and reapplied to avoid problems of local minimum of the fitting curve (Bishop, 1995) and if several models are combined, the output of a model must agree with the input of the next.

We overcame these issues of high quality environmental features sets and their alignment by using an e-Infrastructure for biodiversity conservation (D4science) (Candela et al., 2009). D4Science supplies several models as-a-service. The model compatibility is guaranteed by specialized e-Infrastructure services. Furthermore, D4Science uses Cloud computing to speed processing up (Coro et al., 2013b; Candela et al., 2013). D4Science provides automatic alignment and comparison of geospatial datasets (Coro, 2014), by re-projecting environmental features into a common coordinates system.

D4Science hosts a large variety of environmental features at global scale, with resolution varying from 0.01 degrees to 1 degree (Castelli et al., 2013).

D4Science also allows retrieving species presence information from heterogeneous biodiversity data collections (e.g. OBIS (Berghe et al., 2010), GBIF (Edwards et al., 2000) and the Catalog of Life (Wilson, 2003)), under the same format (Candela et al., 2014). Information is attached to each presence record, to indicate the ownership of the observation, its source (e.g. human observation, specimen etc.) and possibly if the record underwent expert review.

### *3.2. Occurrence data preparation*

We used a presence-absence modelling approach, to find correlation between the presence records of *Architeuthis dux* and a multidimensional space made up of environmental features. We decided to use high quality presence points and reliable absence locations as input to our models, according to the indications reported in Section 2.2. Using the D4Science web services (Candela et al., 2014), we retrieved human observations for *A. dux* from authoritative sources. We came up with 11 records from OBIS and 1 from GBIF. The records are reported in Table 1, along with the name of the subcollection hosting each record. The records had indication about the experts that identified the species. Most points belong to the area around the Gulf of Mexico and one is in North-West Atlantic. The point from GBIF is in agreement with the records from OBIS, thus we decided to use it. We limited the records to the ones for *A. dux* only. In the context of improving data quality, we did not include the other *Architeuthis* species.

It is notable that both OBIS and GBIF contain few of the recent live observations of *Architeuthis dux*. In particular, the observations from CephBase in Table 1 are the only direct observations, whereas the records from the Smithsonian Institute and the Florida Museum of Natural History come from specimens that have been found in the stomach of sperm whales or floating on the sea surface. The other observation records are reliable estimates from the Biodiversity of the Gulf of Mexico Database, derived from literary studies or unregistered observations that have been later validated by experts. The points in Table 1 are associated to the species presence in a depth range between 700 and 475 meters. In our SDM, we used a large resolution of 1° and this softens errors due to the usage of non-exact presence locations. Thus we decided to employ all the points in Table 1 in our model. On the other hand, we used recent live observations of *A. dux*, not included in OBIS and GBIF, to validate our model (see Section 4.1).

Data retrieved using D4Science follow the Darwin Core format (Wieczorek et al., 2012) and can be provided as input to the D4Science models directly. All models accept the same format of input data of presence records, which makes the data preparation phase faster.

### *3.3. Environmental data selection*

The environmental characteristics in our model refer to geospatially explicit chemical and physical measurements. During its training session, our model learns from positive and negative examples that are based only on environmental features. In the subsequent projection session, a real value from 0 to 1 is associated to several locations to assess their habitat suitability. A well performing model is one having good projection on the locations of the training set and, at the same time, not suffering of overfitting issues on the training values (Bishop, 1995).

Environmental features selection requires attention (see Section 2.2) to ensure they are not highly correlated: adding a feature that is dependent on previous ones would not bring more information to the model, but it could add noise during the training session. Furthermore, the spatial resolution should fit the precision of the projection: a model that has to produce a map with resolution 0.5 degrees, should rely on environmental information with the same resolution. This allows not using values coming from rescaling processes or kriging that would add uncertainty to the measurements. Global scale maps also contain estimated values, but these have been produced by experts. Thus, we recommend using the native resolution of the environmental datasets in global scale modelling. Furthermore, the reliability of the data is crucial. This depends on the data provider, as some providers require the dataset to pass a data quality process in order to be published (e.g. MyOcean (Bahurel et al., 2010) and the World Ocean Atlas (Locarnini et al., 2006)).

Features selection methods analyse the features space. Several approaches try to reduce the dimensions of this space, for example by recovering the most independent features or combining them into new features (Jolliffe, 2005; MacLeod, 2010). In our approach, instead, we wanted to reduce the dimension of the number of features to use, but at the same time we wanted to take the correlation between presence points and random points (background points) into account. To such aim, we used the MaxEnt model as a features filter.

We collected environmental features that could *a priori* influence the habitat suitability for *A.dux*, according to the studies we have reported in Section 2.1. We chose the parameters reported in Table 2, averaged on annual values. Based on the depth range of our presence points and on indications from literature (Guerra et al., 2010), we took parameters values in the following ranges: (i) in the entire water column, (ii) averaged between 450 and 1000 meters, (iii) at surface level. In particular, we used the 450-1000 m range when the data provider reported information at several depth ranges. Table 2 indicates the ranges we used for each parameter. The parameters layers come with different projections and reference systems, but the MaxEnt implementation on D4Science automatically accounts for making the layers projections and reference systems uniform, before training the models. In our experiments, the layers from MyOcean and the World Ocean Atlas were available in the e-Infrastructure as GIS layers, while we provided the others as external datasets, in one of the accepted D4Science input formats (Coro, 2014).

During the training phase, MaxEnt minimizes the relative entropy of the features at the presence locations, with respect to the features of random points (Phillips et al., 2006). Presence points are taken as constraints during this minimization. The model uses a linear combination of the features, where the coefficients of the combination are adapted to reflect the "importance" of each variable in predicting the distribution of the species. After the training phase, MaxEnt also reports these coefficients. We relied on these to select the features that provided the most information about the species' habitat preferences, from the point of view of a machine learning model. In other words, we used MaxEnt to filter out the features that could bring noise or that did not bring more information to a model for *A. dux*. We set a non-strict cut-off threshold, taking all the features that had coefficients values higher than the 5% of the maximum coefficient value. In the end, MaxEnt produced the following list of features from the ones in Table 2, ranked according to a decreasing importance: (i) mole concentration of Silicate, (ii) depth, (iii) maximum temperature in the water column, (iv) ph, (v) mole concentration of Nitrate, (vi) range of temperature in the water column, (vii) distance from land, (viii) mass concentration of Chlorophyll.

### *3.4. Absence points*

In order to improve data quality, we searched for a method to produce robust absence locations. Several methods exist to estimate absence locations (Pearson, 2012), but we avoided introducing biases by using other machine learning models. One approach that proved to be effective, is to use an expert system to generate absence locations (Coro et al., 2013b,a). Expert systems combine automatic processing with expert indications and can be used to simulate expert opinion. Thus, we used AquaMaps Native (see Section 2.2) to retrieve absence areas by looking at locations having probability lower than 0.2 but higher than 0. Setting the threshold over zero, selects areas having low values for several environmental envelopes. This approach simulates locations where an expert asserts that the habitat is particularly unsuited for the species. Furthermore, these locations are reported at a relatively high resolution of 0.5 degrees at global scale. From the AquaMaps Native distribution, we extracted absence scattered locations, because this allows having a wider range of environmental characteristics for low probability locations. We took only absences that were two degrees distant at least. In another work (Coro et al., 2013a), we demonstrated that this method results in better performance than using concentrated absence records.

In order to balance the number of presence and absence records, we limited the absence locations to 25 points, slightly more than two times the presence points. These points gave us a wide range of absence environmental features and, at the same time, limited possible over-prediction tendency by niche models. Figure 1 reports the AquaMaps Native distribution for *Architeuthis dux*, and the presence/absence dataset resulting from our selection.

### *3.5. Modelling*

In order to produce distribution maps for *Architeuthis dux*, we used both MaxEnt and Artificial Neural Networks. As input data, we used the presence dataset described in Section 3.2, the pseudo-absences extracted from AquaMaps (see Section 3.4) and the filtered environmental features described in Section 3.3. We assumed that this input was of sufficient quality to ensure the reliability of the models.

We used the MaxEnt model as benchmark to evaluate the performance of an Artificial Neural Network. Our aim was to compare a state-of-the-art model (MaxEnt) that has been yet used to model rare species (Wisz et al., 2008; Elith et al., 2011; Phillips and Dudik, 2008), with a new approach using MaxEnt only to filter out noisy environmental features. In our experiment, we used the MaxEnt implementation of D4Science (Coro, 2014), which is based on the one by the Phillips et al. (Phillips et al., 2006). We trained the model at global scale, with 1-degree of resolution, since this was the highest degree available for our layers and we wanted to avoid resampling. Consequently, also the projection of the model had a 1-degree resolution. We assumed a 0.5 value for the default species prevalence parameter and executed 1000 learning iterations. We performed several training sessions to ensure that the model consistently converged to the same parameters estimation.

In order to evaluate the performance of MaxEnt in distinguishing between absences and presences in the training dataset, we referred to the AUC curve of the model. This indicates the probability threshold to assert a location is suitable to a species. We found that this probability threshold was 0.03 for our model. Thus, we assumed that all probabilities above this threshold identified a location viable for *A. dux* to a certain degree. The resulting distribution map is displayed in Figure 2.

Artificial Neural Networks, in particular Feed Forward Neural Networks (FFNNs) (Bebis and Georgiopoulos, 1994), have proven good performance in niche modelling and have been applied to model the distribution of rare species (Pearson, 2012; Coro et al., 2013b). Furthermore, with respect to alternative models, they have proven to perform better when the quality of the data is high (Coro et al., 2013b). The aim of an FFNN is to build a hierarchical multi-layered network, made up of interconnected nodes, which simulates a complex function. The complexity of the function depends on the number of layers and neurons in the network. During a training session, the weights of the network connections are adapted to produce expected values on the training dataset. In our case, the training set consisted of the environmental features at presence and absence locations, where features were extracted at 1-degree resolution. The FFNN performance depends only on the values assumed by the features on the training set, differently from MaxEnt. For presences, the expected value was set to 1 and for absences it was set to 0. In order to define the optimal number of layers and neurons per layer to use in the network, we adopted a *growing* strategy (Bishop, 1995). We added neurons and layers as far as the error with respect to the training set decreased after a training session (up to a certain threshold). The threshold was empirically set to 0.01 in order to avoid overfitting. We executed the Network training session 10 times for each topology and eventually took the one with the best learning result, i.e. with the lowest mean error with respect to the training points. This process ended in two Networks achieving good learning capacity: one having two layers, with 10 neurons in the first layer and 2 in the second, the other having two layers too, with 100 neurons in the first layer and 2 in the second. We will refer to the first as the "simple topology FFNN" and to the second as the "complex topology FFNN". One characteristic of the second FFNN is that the learning process is more stable, i.e. it usually ends in the same distance from the training set. On the other hand, using simpler topologies is better especially to avoid overfitting issues. Indeed, in Section 4 we demonstrate that the simpler topology gains overall better performance. In the same way we did for MaxEnt, we calculated that for the FFNNs the best threshold to filter out too low habitat suitability was 0.1. Figure 3 reports the maps associated to the two FFNN topologies when we projected the models at global scale, with 1-degree resolution.

### *3.6. Applicability to other species*

Our approach can be generalized and applied to rare species and to data-limited scenarios that satisfy certain conditions. The main steps and the conditions of this generalized process are the following:

1. Retrieve high quality presence locations by relying on the metadata of the records,
2. Select a number of environmental characteristics correlated to the species presence,

3. Use MaxEnt to filter the environmental characteristics that are really important with respect to the presence points,
4. Use expert knowledge or an expert system to detect absence locations. Select absence locations as widespread as possible,
5. Train a Feed Forward Neural Network on presence and absence locations and select the best learning topology,
6. Project the FFNN at global scale, using the a resolution equal to the maximum in the environmental features,
7. Train a MaxEnt model as comparison system.

## 4. Results

In this Section we describe the qualitative and quantitative approaches we used to compare the trained models with existing literature data. First, we report a "qualitative" comparison on coarse presence locations reported in literature for *Architeuthis dux* and *Architeuthis* spp. In order to investigate the differences between the models in detail, we also report the results of a quantitative comparison, with respect to a map drawn by an expert (Nesis, 2003).

### *4.1. Qualitative evaluation*

We used *Architeuthis dux* and *Architeuthis* spp. records reported by different authors (Kjennerud, 1958; Aldrich, 1991; Arfelli et al., 1991; Roeleveld and Lipinski, 1991; Lordan et al., 1998; Ré et al., 1998; Gonzalez et al., 2000;

Cherel, 2003; Kubodera and Mori, 2005; Clarke, 2006; Guerra et al., 2010) in a qualitative analysis of the models performance. The list of reference areas resulting from this analysis is reported in Table 3. *Architeuthis dux* was identified in six areas, while the other eight locations refer to the generic *Architeuthis* spp. We compared our models on these areas, reporting 1 when there was at least one location having non-zero probability and 0 otherwise. Since our models produce potential niche estimations, we also added the AquaMaps Suitable model to the comparison, which is depicted in Figure 4. In this scenario, the performance of the FFNNs is the same, because they predict habitat suitability in almost all the areas where *A. dux* was recorded, and in six of the eight areas where only the genus was reported. Differences between the behaviours of the two FFNNs are in Kerguelen Islands and off the bay of Biscay. It seems that MaxEnt performs slightly better than the FFNNs and AquaMaps, because it matches several areas for both *A. dux* and *A.* spp. On the other hand, in many locations the probabilities indicated by the model are low.

When we set a probability threshold to filter out values lower than 0.8, the maps highlight only the places with high habitat suitability. In this case, the results of the assessments by the models are reported in Table 4. We notice that the FFNN with the simple topology and AquaMaps Suitable still present high performance. In particular, the FFNN predicts species presence in Newfoundland, Norway Sea, South America, South-Eastern Africa and in the Mediterranean Sea. Conversely, the AquaMaps Suitable model covers the Eastern-North Atlantic, the Kerguelen Islands, the New Zealand coasts and the Tasman Sea. Using this probability threshold, the complex topology FFNN and the MaxEnt model predict very few suitable areas, especially for *Architeuthis* spp. This means that, overall, the FFNN with the simple topology is more stable and reliable. One evident difference between the FFNNs and the AquaMaps model is that, according to AquaMaps, the species is not present in open ocean but only prefers coastal areas. In order to explore more such difference, we used a quantitative discrepancy analysis.

### *4.2. Quantitative evaluation*

In order to quantitatively compare the similarity between the maps, we used also a distribution map drawn by an expert, which is depicted in Figure 5. Nesis (Nesis, 2003) mapped the distribution of *Architeuthis dux* relying on his knowledge about the species: he identified three main areas corresponding to the species presence, i.e. North Atlantic Ocean, North Pacific Ocean and Southern Ocean. In order to make a numeric comparison, we georeferenced this map using QGIS (Quantum GIS, 2011) and obtained a polygonal representation of the distribution. We assigned probability 1 to the regions indicated in the map and forced a 0 value to absence areas that did not contain locations reported in the qualitative analysis, i.e. the Arabian Sea, the Indian Ocean and the South Atlantic Ocean. The map by Nesis does not have high precision, thus we did not expect a full agreement by the models, but it gives a common field for an overall comparison of the maps.

We assumed that the map closest to this was the most reliable.

In order to quantitatively measure the distance between the maps, we used the maps comparison process described in (Coro et al., 2014). This process performs a point-to-point comparison between two maps at a given resolution and calculates indicators of their similarity. Among the measurements produced by this process, we concentrated on "accuracy", i.e. the ratio of locations where the probabilities by two models give the same value, according to a certain tolerance threshold. We used several tolerance thresholds to vary the strictness with respect to presence and absence locations. A threshold of 0.3, means that two probability values for a certain location are considered as having the same value if they differ less than 0.3. We performed this point-to-point comparison at 1-degree resolution.

Table 5 reports the performance using several thresholds: 0.8, 0.5 and 0.3. Furthermore, we made three comparisons with the map of Nesis using presence-only, absence-only and presence-absence polygons separately. In this way we observed that, even if one model can be in good agreement with either presences or absences, it can be in lower agreement with respect to both. The FFNN with the simple topology has lower agreement with absence locations, but overall is the closest to the expert drawn map, according to all the probability thresholds.

## 5. Discussion

The results demonstrate that, according to a qualitative analysis, the simple topology FFNN gives the most promising results. In this scenario, the AquaMaps Suitable model is indeed the most stable. On the other hand, if we move to a quantitative evaluation with respect to an expert-drawn map, we better understand the differences between AquaMaps and the FFNN. AquaMaps presents few points in open ocean, because the model assigns more weight to the proximity of land, while the expert's map indicates many of these points as suitable locations. This discrepancy is reflected in the overall better similarity between the expert's map and the FFNN map. MaxEnt gains good performance too, but it overestimates absence locations, thus the overall accuracy is lower than the FFNN one.

FFNN identifies suitable habitat for *Architeuthis dux* in the Northern and Eastern Atlantic Ocean (i.e around Newfoundland and in the Norway Sea). This agrees with literature studies that indicate Newfoundland as the original centre of dispersal for the European population of *A. dux* (Robson, 1933). Our model also agrees with other studies (Roeleveld and Lipinski, 1991; Kubodera and Mori, 2005) reporting records in the North Atlantic Ocean (Sweeney and Roper, 2001) and predicts habitat suitability in correspondence of continental slopes, canyons and abyssal plains.

The FFNN is the model that better resembles the expert's map, but more information is needed to ensure its reliability: there are some discrepancy locations, like the South Africa coasts, that need further investigation. The highest discrepancy with respect to the expert's map is in the South-West coast of South Africa, in the Indian Ocean and in North Australia. This discrepancy could be explained by the fact that the FFNN predicts potential habitat, while the expert indicates the known (actual) habitat. On the other hand, there are studies supporting the indications by the FFNN map: *Architeuthis* specimens were captured in South-West Pacific Ocean, and around Australian coasts, especially off the West coasts (Jackson, 1991; Sweeney and Roper, 2001). As for the Indian Ocean, several studies report the presence of *Architeuthis* near the Reunion Island, the Mauritius Islands and generally in the South-Western Indian Ocean (Sweeney and Roper, 2001; Guerra et al., 2011; Cherel, 2003; Mikhalev et al., 1981). In some Indian survey works, it is reported that *Architeuthis* species are present off the west coasts of India (Silas, 1968, 1985).

Some scientists stress out that different species of *Architeuthis* cannot have overlapping populations (Roeleveld and Lipinski, 1991). Although it has been suggested that the West coast of South Africa is a "natural" habitat for *Architeuthis*, no certified record of *A. dux* has been reported yet.

In summary, even if we cannot demonstrate the effectiveness of the FFNN model in this case, we can state that there are good hints about its better reliability with respect to AquaMaps and MaxEnt. This effect is due to the abstraction power of this presence/absence model (Coro et al., 2013b), and also to the data preparation phase of our approach.

## 6. Conclusions

In this paper, we have described a method to predict the distribution of *Architeuthis dux* at global scale. We have used a presence-only model to identify important environmental features possibly extracted at *Architeuthis* depth ranges indicated by other studies, we have generated absence locations using an expert system and we have retrieved presence records from two authoritative data sources. By means of a presence/absence model based on an Artificial Neural Network, we have produced a potential habitat distribution for *A. dux* having reasonably good reliability. This distribution is the one that is most in agreement with the opinion of an expert. Common traits in the expert's map and in the Neural Network map are visible, e.g. there is a common strip of absences from Brazil to the coasts of Guinea-Sierra Leone. Agreement between the maps in other regions is lower (e.g. in the Indian Ocean), but overall the simple topology FFNN is the best model compared to the maps produced with AquaMaps Suitable and MaxEnt. As discussed in Section 5, the Neural Network map correctly predicts some known species habitat and depicts the potential (not the actual) distribution of the species. It covers locations where the species was observed, but that were not included in the training set, and it neglects other locations where the observations probably did not refer strictly to *A. dux*.

In summary, maximising the reliability of presence, absence and environmental parameters gives good estimate of the distribution of *A. dux*. This maximisation determines reliable patterns of occurrence related to environmental gradients, as also supported by other studies (Segurado and Araujo, 2004; Franklin, 2010). A large scale distribution for *A. dux* can also help understanding the role of this species on a broader geographic perspective (Lordan et al., 2001).

The work reported in this paper builds on our previous experience on modelling the distribution of the Coelacanth (Coro et al., 2013b). In our previous work, we used a model combining a Neural Network with absence information produced from AquaMaps. The model was trained using only observation records near Madagascar and the same environmental parameters used by AquaMaps. The approach was promising, because it predicted habitat suitability in some locations in Indonesia were a variant of the Coelacanth has been really observed. In this paper we have enhanced this model, because we (i) use other environmental parameters, (ii) select the most influential parameters and (iii) suggest a method to compare the results with other maps and understand complementarity. Furthermore, we have explained how our approach can be generalized and extended to other rare species.

Generally speaking, the presented work can be useful in species conservation. In fact, model-based approaches for rare species that count on data quality have proved to be valuable when used in population management and conservation strategies (Austin, 2007). In particular, many conservation projects need a complete description of species' geographical distributions, and modelling techniques (e.g. MaxEnt, Artificial Neural Networks and AquaMaps) have already proved to reliably support this activity (Ficetola et al., 2007; Ward, 2007; Hijmans and Graham, 2006; Fitzpatrick et al., 2008; Thorn et al., 2009; Wollan et al., 2008; Echarri et al., 2009; Cordellier and Pfenninger, 2009). The produced maps can be also used in fisheries, because producing a potential distribution for a rare species like the giant squid can help locating vulnerable marine ecosystems (Auster et al., 2010; Stevens et al., 2000; Tittensor et al., 2009; Stevens et al., 2000).

The D4Science e-Infrastructure enabled the prediction of the distribution of *A. dux* with powerful modelling resources, automated data retrieval and results sharing. Furthermore, the experiment is fully reproducible. This experiment demonstrates how e-Infrastructures can support species distribution modelling of rare species.

## 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 Giovanisì project of the Presidency of the Tuscan Regional Government.

## References

Aldrich, F., 1991. Some aspects of the systematics and biology of squid of the genus architeuthis based on a study of specimens from newfoundland waters. Bulletin of Marine Science 49 (1874), 457–481.

Arfelli, C., De Amorim, A. F., Tomas, A., 1991. First record of a giant squid Architeuthis sp. Steenstrup, 1857 (Cephalopoda, Architeuthidae) in Brazilian waters. Bol. Inst. Pesca Sao Paulo 18, 83–88.

Auster, P. J., Gjerde, K., Heupel, E., Watling, L., Grehan, A., Rogers, A. D., 2010. Definition and detection of vulnerable marine ecosystems on the high seas: problems with the "move-on" rule. ICES Journal of Marine Science: Journal du Conseil, fsq074.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological modelling 200 (1), 1–19.

Bahurel, P., Adragna, F., Bell, M. J., Jacq, F., Johannessen, J. A., Le Traon, P.-Y., Pinardi, N., She, J., 2010. Ocean monitoring and forecasting core services: The european myocean example. Proceedings of OceanObis 9, 02.

Bebis, G., Georgiopoulos, M., 1994. Feed-forward neural networks. Potentials, IEEE 13 (4), 27–31.

Berger, A., 1996. A brief maxent tutorial. http://www-2.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html 25.

Berghe, E. V., Stocks, K. I., Grassle, J. F., 2010. Data integration: The ocean biogeographic information system. Life in the World's Oceans: Diversity, Distribution, and Abundance, 333.

Bishop, C. M., 1995. Neural networks for pattern recognition. Clarendon press Oxford.

Bolstad, K. S., O'Shea, S., 2004. Gut contents of a giant squid Architeuthis dux (Cephalopoda: Oegopsida) from New Zealand waters. New Zealand Journal of Zoology 31 (1), 15–21.

Brix, O., 1983. Giant squids may die when exposed to warm water currents. Nature 303, 422–423.

Brotons, L., Thuiller, W., Araujo, M. B., Hirzel, A. H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448.

Candela, L., Castelli, D., Coro, G., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., 2014. An infrastructure-oriented approach for supporting biodiversity research. Ecological Informatics n/a (0), n/a.

Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F., 2013. Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience, n/a.
URL `http://dx.doi.org/10.1002/cpe.3030`

Candela, L., Castelli, D., Pagano, P., 2009. D4science: an e-infrastructure for supporting virtual research environments. In: IRCDL. pp. 166–169.

Cao, Y., Williams, D. D., Williams, N. E., 1998. How important are rare species in aquatic community ecology and bioassessment? Limnology and Oceanography 43 (7), 1403–1409.

Castelli, D., Pagano, P., Candela, L., Coro, G., 2013. The imarine data bonanza: Improving data discovery and management through an hybrid data infrastructure. Bollettino di Geofisica Teorica ed Applicata 54, 105–107.

Cherel, Y., 2003. New records of the giant squid Architeuthis dux in the southern Indian Ocean. Journal of the Marine Biological Association of the UK 83 (6), 1295–1296.

Cianfrani, C., Le Lay, G., Hirzel, A. H., Loy, A., 2010. Do habitat suitability models reliably predict the recovery areas of threatened species? Journal of Applied Ecology 47 (2), 421–430.

Clarke, M., 2006. Oceanic cephalopod distribution and species diversity in the eastern north atlantic malcolm r. clarke. Life and marine Sciences 23A, 27–46.

Clarke, M., Pascoe, P., 1997. Cephalopod Species in the Diet of a Sperm Whale (Physeter Catodon) Stranded at Penzance, Cornwall. Journal of the Marine Biological Association of the United Kingdom 77 (04), 1255.

Cordellier, M., Pfenninger, M., 2009. Inferring the past to predict the future: climate modelling predictions and phylogeography for the freshwater gastropod radix balthica (pulmonata, basommatophora). Molecular Ecology 18 (3), 534–544.

Coro, G., 2014. The D4Science MaxEnt implementation. http://wiki.i-marine.eu/index.php/MaxEnt.

Coro, G., Pagano, P., Ellenbroek, A., 2013a. Automatic procedures to assist in manual review of marine species distribution maps. In: Adaptive and Natural Computing Algorithms. Springer, pp. 346–355.

Coro, G., Pagano, P., Ellenbroek, A., 2013b. Combining simulated expert knowledge with Neural Networks to produce Ecological Niche Models for *Latimeria chalumnae*. Ecological Modelling 268, 55–63.

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

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

Echarri, F., Tambussi, C., Hospitaleche, C. A., 2009. Predicting the distribution of the crested tinamous, eudromia spp.(aves, tinamiformes). Journal of Ornithology 150 (1), 75–84.

Edwards, J. L., Lane, M. A., Nielsen, E. S., 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289 (5488), 2312–2314.

Elith, J., Leathwick, J. R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40 (1), 677–697.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., Yates, C. J., Jan. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17 (1), 43–57.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41, 263–274.

Ferrier, S., 2002. Mapping Spatial Pattern in Biodiversity for Regional Conservation Planning: Where to from Here? Syst. Biol 51 (2), 331–363.

Ficetola, G. F., Thuiller, W., Miaud, C., 2007. Prediction and validation of the potential global distribution of a problematic alien invasive species—the american bullfrog. Diversity and Distributions 13 (4), 476–485.

Fitzpatrick, M. C., Gove, A. D., Sanders, N. J., Dunn, R. R., 2008. Climate change, plant migration, and range collapse in a global biodiversity hotspot: the banksia (proteaceae) of western australia. Global Change Biology 14 (6), 1337–1352.

Förch, E., 1998. The marine fauna of New Zealand: Cephalopoda: Oegopsida: Architeuthidae (Giant Squid). NIWA Biodiversity Memoir 110, 1–113.

Franklin, J., 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press.

Gibson, L., Barrett, B., Burbidge, A., 2007. Dealing with uncertain absences in habitatmodelling: a case study of a rare ground-dwelling parrot. Diversity and Distributions 13, 704–713.

Gonzalez, M., Fernandez-Casado, M., Rodriguez, P., Segura, A., Martin, J. J., 2000. First record of the giant squid Architeuthis sp . (Architeuthidae) in the Mediterranean Sea ¨. J.Mar.Biol.Ass.U.K. 80, 745–746.

Guerra, A., Gonzalez, A. F., Dawe, E. G., Rocha, F., 2004. Records of giant squid in the north-eastern Atlantic, and two records of male Architeuthis sp. of the Iberian Peninsula. J.Mar.Biol.Ass.U.K. 84, 426–431.

Guerra, A., González, A. F., Pascual, S., Dawe, E. G., 2011. The giant squid *Architeuthis*: An emblematic invertebrate that can represent concern for the conservation of marine biodiversity. Biological Conservation 144 (7), 1989–1997.

Guerra, Á., Rodríguez-Navarro, A. B., González, Á. F., Romanek, C. S., Álvarez-Lloret, P., Pierce, G. J., 2010. Life-history traits of the giant squid architeuthis dux revealed from stable isotope signatures recorded in beaks. ICES Journal of Marine Science: Journal du Conseil, fsq091.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N. G., Lehmann, A., Zimmermann, N. E., 2006. Using Niche-Based Models to Improve the Sampling of Rare Species. Conservation Biology 20 (2), 501–511.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8 (9), 993–1009.

Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Hijmans, R. J., Graham, C. H., 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Global change biology 12 (12), 2272–2281.

Hirzel, A. H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. Journal of Applied Ecology 45 (5), 1372–1381.

IEDA, 2014. The Marine Geoscience website. http://www.marine-geo.org/tools/maps_grids.php.

IUCN Species Survival Commission and Natural Resources. Species Survival, 2001. IUCN Red List Categories and Criteria. Osprey Publishing.

Jackson, G. D., 1991. Age, growth and population dynamics of tropical squid and sepioid populations in waters off townsville, north queensland, australia. Ph.D. thesis, James Cook University.

Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.

Kamino, L., Stehmann, J., Amaral, S., De Marco, P., Rangel, T., de Siqueira, M., De Giovanni, R., Hortal, J., 2012. Challenges and perspectives for species distribution modelling in the neotropics. Biology letters 8 (3), 324–326.

Kaschner, K., Ready, J., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood, P., South, A., Kullander, S., Rees, T., Close, C., et al., 2008. Aquamaps: Predicted range maps for aquatic species. World wide web electronic publication, www. aquamaps. org, Version 8, 2010.

Kaschner, K., Watson, R., Trites, A., Pauly, D., 2006. Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. Marine Ecology Progress Series 316, 285–310.

Kjennerud, J., 1958. Description of a giant squid, Architeuthis, stranded on the west coast of Norway. Grieg.

Kubodera, T., Mori, K., 2005. First-ever observations of a live giant squid in the wild. Proceedings of the Royal Society B: Biological Sciences 272 (1581), 2583–2586.

Locarnini, R. A., Mishonov, A., Antonov, J., Boyer, T., Garcia, H., Levitus, S., et al., 2006. World ocean atlas 2005 volume 1: Temperature [+ dvd]. Noaa atlas nesdis 61 (1).

Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J., Guisan, A., 2010. Overcoming the rare species modelling paradox: a novel hierarchical framework applied to an iberian endemic plant. Biological Conservation 143 (11), 2647–2657.

Lordan, C., Collins, M., Perales-Raya, C., 1998. Observations on Morphology, Age and Diet of Three Architeuthis Caught Off the West Coast of Ireland in 1995. Journal of the Marine Biological Association of the United Kingdom 78, 903–917.

Lordan, C., Warnes, S., Cross, T. F., Burnell, G. M., 2001. The distribution and abundance of cephalopod species caught during demersal trawl surveys west of ireland and in the celtic sea. Marine Institute Open Access Repository.

Lyons, J., Navarro-Perz, S., Cochran, P. A., Santana-C., E., Guzman-Arroyo, M., 1995. Index of biotic integrity based on fish assemblages for the conservation of streams and rivers in west-central Mexico. Conserv. Biol. 9, 569–584.

MacLeod, C. D., Jun. 2010. Habitat representativeness score (HRS): a novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. Journal of the Marine Biological Association of the United Kingdom 90 (07), 1269–1277.

Márcia Barbosa, Real, R., Olivero, J., Mario Vargas, J., Dec. 2003. Otter (Lutra lutra) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. Biological Conservation 114 (3), 377–387.

Mikhalev, J., Savusin, V., Kishiyan, N., Ivashin, M., 1981. To the problem of the feeding of sperm whales from the southern hemisphere. Reports of the International Whaling Commission 31, 737–745.

Nesis, K. N., 1987. Cephalopods of the World. TFH Publications, Inc, Neptune City, New Jersey Nesis.

Nesis, K. N., 2003. Distribution of recent cephalopoda and implications for plio-pleistocene events. Coleoid cephalopods through time 3 (4), 199–224.

Ovaskainen, O., Soininen, J., 2011. Making more out of sparse data: hierarchical modeling of species communities. Ecology 92 (2), 289–295.

Pearce, J. L., Boyce, M. S., Jun. 2006. Modelling distribution and abundance with presence-only data. Journal of Applied Ecology 43 (3), 405–412.

Pearson, R. G., 2012. Species distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.

Pearson, R. G., Dawson, T. P., Berry, P. M., Harrison, P., 2002. SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species. Ecological Modelling 154 (3), 289–300.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4), 231–259.

Phillips, S. J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Pulliam, H., 2000. On the relationship between niche and distribution. Ecology Letters 3, 349–361.

Quantum GIS, 2011. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. URL:[http://qgis. osgeo. org].

Razgour, O., Hanmer, J., Jones, G., 2011. Using multi-scale modelling to predict habitat suitability for species of conservation concern: the grey long-eared bat as a case study. Biological Conservation 144 (12), 2922–2930.

Ré, M., Baron, P., Beron, J., Gosztonyi, A., Kuba, L., Monsalve, M., Sardella, N., 1998. A giant squid Architeuthis sp. (Mol-lusca, Cephalopoda) stranded on the Patagonian shore of Argentina. Cephalopod Biodiversity, Ecology and Evolution. S. Afr. J. Mar. Sci. 20, 109–122.

Ready, J., Kaschner, K., South, A. B., Eastwood, P. D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R., 2010. Predicting the distributions of marine organisms at the global scale. Ecological Modelling 221 (3), 467 – 478.

Rebelo, H., Jones, G., 2010. Ground validation of presence-only modelling with rare species: a case study on barbastelles barbastella barbastellus (chiroptera: Vespertilionidae). Journal of Applied Ecology 47 (2), 410–420.

Robson, G., 1933. On architeuthis clarkei, a new species of giant squid, with observations on the genus. In: Proceedings of the Zoological Society of London. Vol. 103. Wiley Online Library, pp. 681–697.

Roeleveld, M., Lipinski, M., 1991. The giant squid Architeuthis in southern African waters. J.Zool.Lond. 224, 431–477.

Segurado, P., Araujo, M. B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31 (10), 1555–1568.

Silas, E., 1968. Cephalopoda of the west coast of india collected during the cruises of the research vessel varuna, with a catalogue of the species known from the indian ocean. In: Proceedings of the Symposium on Mollusca. Vol. 1. pp. 277–359.

Silas, E., 1985. Cephalopod fisheries of india—an introduction to the subject with methodologies adopted for this study. CMFRI Bulletin 37, 1–4.

Stevens, J., Bonfil, R., Dulvy, N., Walker, P., 2000. The effects of fishing on sharks, rays, and chimaeras (chondrichthyans), and the implications for marine ecosystems. ICES Journal of Marine Science: Journal du Conseil 57 (3), 476–494.

Sweeney, M., Roper, C., 2001. Records of architeuthis specimens from published reports. Avaliable online at: www.mnh.si.edu/cephs/archirec.pdf.

The AquaMaps Consortium, 2014. The AquaMaps website. Www.aquamaps.org.

Thorn, J. S., Nijman, V., Smith, D., Nekaris, K., 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for asian slow lorises (primates: Nycticebus). Diversity and Distributions 15 (2), 289–298.

Tittensor, D. P., Baco, A. R., Brewin, P. E., Clark, M. R., Consalvey, M., Hall-Spencer, J., Rowden, A. A., Schlacher, T., Stocks, K. I., Rogers, A. D., 2009. Predicting global habitat suitability for stony corals on seamounts. Journal of Biogeography 36 (6), 1111–1128.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., De Clerck, O., 2012. Bio-oracle: a global environmental dataset for marine species distribution modelling. Global Ecology and Biogeography 21 (2), 272–281.

Ward, D. F., 2007. Modelling the potential geographic distribution of invasive ant species in new zealand. Biological Invasions 9 (6), 723–735.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin core: An evolving community-developed biodiversity data standard. PLoS One 7 (1), e29715.

Wilson, E. O., 2003. The encyclopedia of life. Trends in Ecology & Evolution 18 (2), 77–80.

Wisz, M. S., Hijmans, R., Li, J., Peterson, A. T., Graham, C., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14 (5), 763–773.

Wollan, A. K., Bakkestuen, V., Kauserud, H., Gulden, G., Halvorsen, R., 2008. Modelling and predicting fungal distribution patterns using herbarium data. Journal of Biogeography 35 (12), 2298–2310.

| Data collection | Collection code | Last update | Locality | Lat. | Long. |
|---|---|---|---|---|---|
| OBIS | USNM | 11/05/2010 | Gulf of Mexico | 26.98 | -90.37 |
| OBIS | HRI | 11/12/2009 | WSW Gulf of Mexico | 22.45 | -97.31 |
| OBIS | HRI | 11/12/2009 | ESE Gulf of Mexico | 23.04 | -82.93 |
| OBIS | HRI | 11/12/2009 | NNW Gulf of Mexico | 27.69 | -91.75 |
| OBIS | HRI | 11/12/2009 | NNE Gulf of Mexico | 29.47 | -87.17 |
| OBIS | HRI | 11/12/2009 | SSE Gulf of Mexico | 23.64 | -89.18 |
| OBIS | HRI | 11/12/2009 | ENE Gulf of Mexico | 26.91 | -84.71 |
| OBIS | HRI | 11/12/2009 | WNW Gulf of Mexico | 26.96 | -96.08 |
| OBIS | HRI | 11/12/2009 | SSW Gulf of Mexico | 19.24 | -93.51 |
| OBIS | 343 | n/a | South Carolina coast | 31 | -76 |
| OBIS | 343 | n/a | Newfoundland | 48.16 | -49.33 |
| GBIF | FLMNH | n/a | Florida coast | 27.26 | -80.01 |

Table 1: Occurrence records from the OBIS and GBIF data collections. The collection codes refer to the OBIS and GBIF codes for the following sub-collections: Biodiversity of the Gulf of Mexico Database (HRI), Invertebrate Zooology Collections (Smithsonian Institute, USNM), CephBase (343), Florida Museum of Natural History (FLMNH).

| Parameter | Spatial Resolution | Unit of Measure | Provider |
|---|---|---|---|
| Minimum temperature (in the water column) | 1° | K | World Ocean Atlas |
| Maximum temperature (in the water column) | 1° | K | World Ocean Atlas |
| Range of temperature (in the water column) | 1° | K | World Ocean Atlas |
| Salinity (avg 450-1000 m) | 1° | - | World Ocean Atlas |
| Ph (avg in the water column) | 0.083° | - | Bio-Oracle |
| Mass concentration of Chlorophyll (avg 450-1000 m) | 0.5° | m g/$m^3$ | MyOcean |
| Mole concentration of Nitrate (avg 450-1000 m) | 0.5° | m mol/$m^3$ | MyOcean |
| Dissolved Oxygen (avg 450-1000 m) | 1° | m g/l | World Ocean Atlas |
| Mole concentration of Phosphate (avg 450-1000 m) | 1° | $\mu$ mol/l | World Ocean Atlas |
| Mole concentration of Silicate (avg 450-1000 m) | 1° | $\mu$ mol/l | World Ocean Atlas |
| Wind stress (surface level) | 0.25° | Pa | MyOcean |
| Depth (max in a 0.14° sqr. cell) | 0.14° | $m$ | Marine Geoscience |
| Distance from land (centre of a 0.5° sqr. cell) | 0.5° | $m$ | AquaMaps |

Table 2: Complete list of environmental characteristics related to the *Architeuthis dux* distribution we used in our features selection phase. The datasets come from several and heterogeneous sources: MyOcean (Bahurel et al., 2010), World Ocean Atlas (Locarnini et al., 2006), Bio-Oracle (Tyberghein et al., 2012), Marine Geoscience website (IEDA, 2014) and the AquaMaps website (The AquaMaps Consortium, 2014).

| **Areas** | **Species or Genus** | **FFNN (100-2)** | **FFNN (10-2)** | **MaxEnt** | **AquaMaps Suitable** |
|---|---|---|---|---|---|
| KERGUELEN ISLANDS | *A.dux* | 1 | 0 | 0 | 1 |
| NEW ZEALAND-TASMAN SEA | *A.dux* | 1 | 1 | 1 | 1 |
| BAY OF BISCAY | *A.dux* | 0 | 1 | 1 | 1 |
| NORTH-EAST ATLANTIC | *A.dux* | 0 | 0 | 1 | 1 |
| NEWFOUNDLAND | *A.dux* | 1 | 1 | 1 | 0 |
| NORWEGIAN SEA | *A.dux* | 1 | 1 | 1 | 1 |
| IRELAND COASTS | *Architeuthis spp.* | 1 | 1 | 1 | 1 |
| PATAGONIA | *Architeuthis spp.* | 1 | 1 | 1 | 0 |
| BRAZIL | *Architeuthis spp.* | 1 | 1 | 1 | 1 |
| JAPAN | *Architeuthis spp.* | 1 | 1 | 1 | 1 |
| SOUTH AFRICA-ORANGE RIVER | *Architeuthis spp.* | 0 | 0 | 0 | 0 |
| SOUTH AFRICA-TABLE BAY | *Architeuthis spp.* | 0 | 0 | 0 | 0 |
| SOUTH AFRICA-DURBAN | *Architeuthis spp.* | 1 | 1 | 1 | 1 |
| FUENGIROLA BEACH-MEDITERRANEAN SEA | *Architeuthis spp.* | 1 | 1 | 1 | 1 |
| ACCURACY | | 71.4% | 71.4% | 78.6% | 71.4% |

Table 3: Comparison between the predictions of *Architeuthis dux* presence on indicative presence areas. The second column indicates if the species was reported at genus or species level. FFNN (x-y) indicates a Feed-Forward Artificial Neural Network having 2 layers, with x neurons in the first layer and y neurons in the second. Values equal to 1 indicate that the models report sensibly non-zero values in that area.

| **Areas** | **Species or Genus** | **FFNN (100-2)** | **FFNN (10-2)** | **MaxEnt** | **AquaMaps Suitable** |
|---|---|---|---|---|---|
| KERGUELEN ISLANDS | *A.dux* | 0 | 0 | 0 | 1 |
| NEW ZEALAND-TASMAN SEA | *A.dux* | 0 | 0 | 0 | 1 |
| BAY OF BISCAY | *A.dux* | 0 | 0 | 0 | 1 |
| NORTH-EAST ATLANTIC | *A.dux* | 0 | 0 | 0 | 1 |
| NEWFOUNDLAND | *A.dux* | 0 | 1 | 0 | 0 |
| NORWEGIAN SEA | *A.dux* | 0 | 1 | 1 | 1 |
| IRELAND COASTS | *Architeuthis spp.* | 0 | 0 | 0 | 1 |
| PATAGONIA | *Architeuthis spp.* | 0 | 1 | 0 | 0 |
| BRAZIL | *Architeuthis spp.* | 0 | 1 | 0 | 1 |
| JAPAN | *Architeuthis spp.* | 1 | 1 | 0 | 1 |
| SOUTH AFRICA-ORANGE RIVER | *Architeuthis spp.* | 0 | 0 | 0 | 0 |
| SOUTH AFRICA-TABLE BAY | *Architeuthis spp.* | 0 | 0 | 0 | 0 |
| SOUTH AFRICA-DURBAN | *Architeuthis spp.* | 0 | 1 | 0 | 1 |
| FUENGIROLA BEACH-MEDITERRANEAN SEA | *Architeuthis spp.* | 0 | 1 | 1 | 1 |
| ACCURACY | | 7.1% | 50% | 14.3% | 71.4% |

Table 4: Comparison between the predictions of *Architeuthis dux* presence on indicative presence areas, when the probability threshold for sensibly non-zero values is set to 0.8. The second column indicates if the species was reported at genus or species level. FFNN (x-y) indicates a Feed-Forward Artificial Neural Network having 2 layers, with x neurons in the first layer and y neurons in the second. Values equal to 1 indicate that the models report sensibly non-zero values in that area.

| **Accuracy with resp. to Nesis (Nesis, 2003).** | | | |
|---|---|---|---|
| | **Comparison thresholds** | | |
| | **0.8** | **0.5** | **0.3** |
| **Presences and Absences** | | | |
| FFNN (10-2) | **42.83%** | **30.56%** | **26.81%** |
| MaxEnt | 21.68% | 18.36% | 17.65% |
| AquaMaps Suitable | 22.01% | 20.19% | 18.83% |
| FFNN (100-2) | 29.85% | 20.56% | 16.3% |
| **Presences-only** | | | |
| FFNN (10-2) | **44.42%** | **31.42%** | **27.81%** |
| MaxEnt | 4.72% | 0.78% | 0.19% |
| AquaMaps Suitable | 5.35% | 3.95% | 2.61% |
| FFNN (100-2) | 17.91% | 9.24% | 6.42% |
| **Absences-only** | | | |
| FFNN (10-2) | 38.27% | 29.53% | 25.09% |
| MaxEnt | **100%** | **100%** | **99.21%** |
| AquaMaps Suitable | 99.46% | 95.78% | 94.35% |
| FFNN (100-2) | 87.77% | 75.55% | 64.5% |

Table 5: Accuracy of a point-to-point maps comparison process at 1-degree resolution (Coro et al., 2014), using presence and absence locations indicated by Nesis (Nesis, 2003). The performance is reported also on presence and absence locations separately.

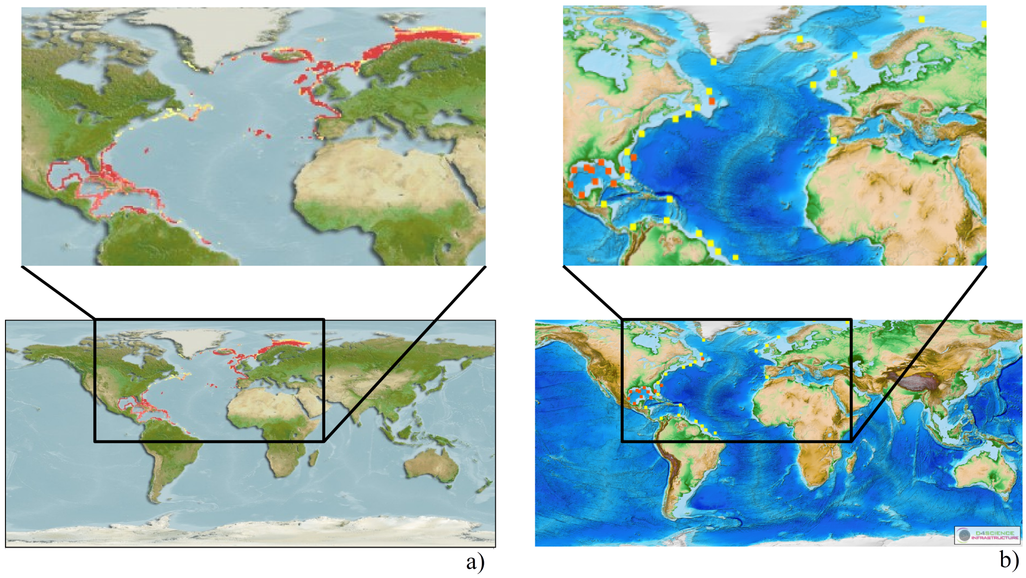



Figure 1: a. The AquaMaps Native distribution for *Architeuthis dux*. Darker colours refer to higher probability locations. b. The presences/absence points resulting from our process. Darker colours refer to presence locations.

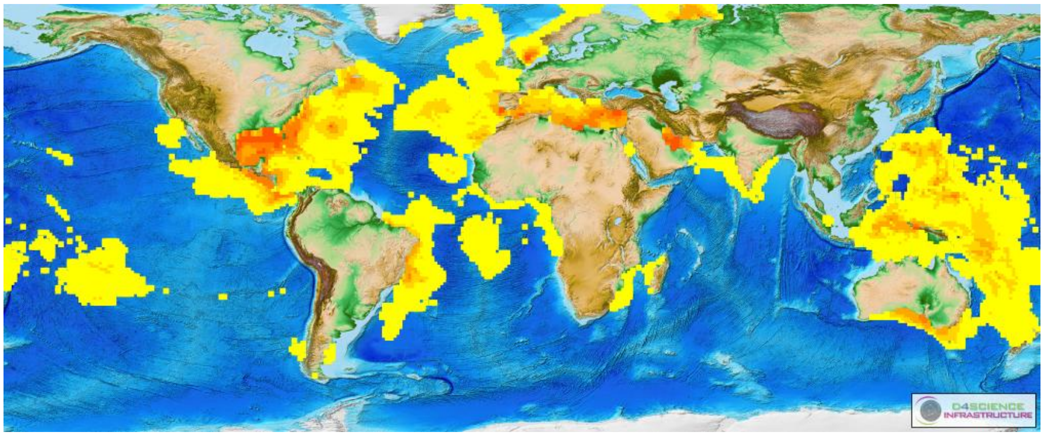
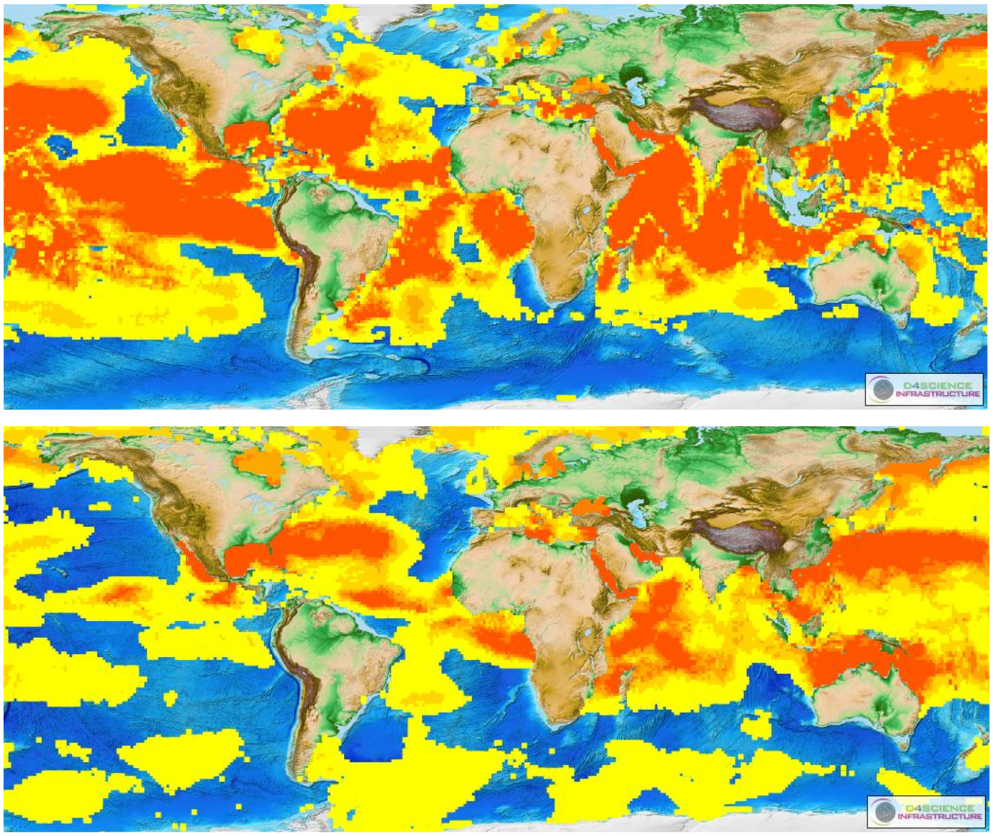



Figure 2: Distribution of *A. dux* produced with the MaxEnt model, trained using our filtered environmental features.




a) b)

Figure 3: Distribution of *A. dux* produced by two Artificial Feed Forward Neural Networks: (a) with 2 layers, containing 10 neurons in the first layer and 2 in the second; (b) with 2 layers, containing 100 neurons in the first layer and 2 in the second.

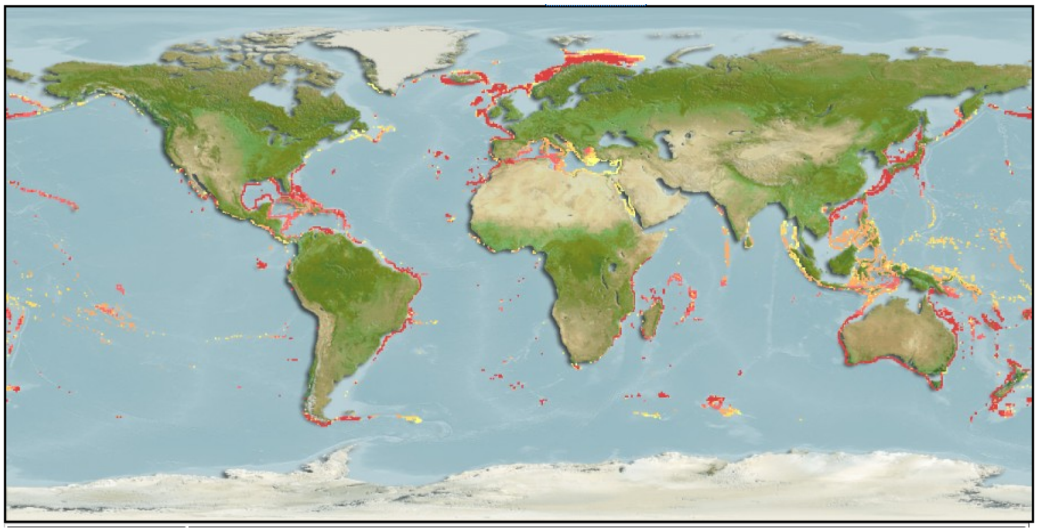


Figure 4: Distribution of *A. dux* produced with the AquaMaps Suitable model (Kaschner et al., 2008).

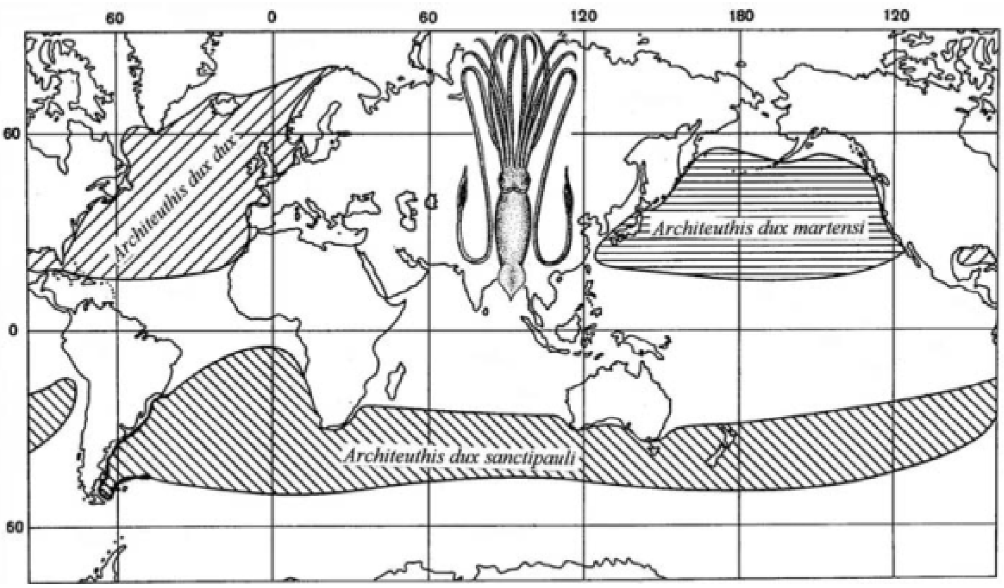



Figure 5: Distribution of *A. dux* reported by Nesis (2003).