

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

3 Gianpaolo Coro^{a,1,2,*}, Chiara Magliozzi^a, Anton Ellenbroek^b, Pasquale
4 Pagano^a

5 ^a*Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa,*
6 *Italy*

7 ^b*Food and Agriculture Organization of the United Nations (FAO)*

8 **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

Preprint submitted to Ecological Modelling
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) March 17, 2015

¹Telephone Number: +39 050 315 2978

²Fax 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.

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

11 **1. Introduction**

12 In recent years, niche models that estimate species distribution have be-
13 come widely used in conservation biology (Guisan and Zimmermann, 2000).
14 Rare species are examples where the prediction of suitable habitats is paramount
15 to support fisheries management policies and conservation strategies (Pearce
16 and Boyce, 2006; Márcia Barbosa et al., 2003). Defined by Cao et al. (Cao
17 et al., 1998) as species that occur at lower frequency or in low number in a
18 sample of certain size, rare species have a key role in affecting biodiversity
19 richness and by consequence they are indicators of degradation for aquatic
20 ecosystems (Lyons et al., 1995; Cao et al., 1998). In this context, predictive
21 models can considerably support the qualitative and quantitative criteria
22 used to assign a “status” to a species (IUCN Species Survival Commission
23 and Natural Resources. Species Survival, 2001), by providing accurate, ap-
24 plicable and reliable spatial predictions to species population monitoring and
25 sampling (Guisan et al., 2006). As discussed in many studies, the method-
26 ological progresses of Species Distribution Models (SDMs) allow nowadays to

27 apply robust techniques to rare and endangered species (Guisan and Thuiller,
28 2005; Ferrier, 2002; Gibson et al., 2007; Razgour et al., 2011; Ovaskainen and
29 Soininen, 2011; Rebelo and Jones, 2010; Wisz et al., 2008; Lomba et al., 2010).

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

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

46 We illustrate our approach using data from authoritative sources of ob-
47 servation records. Furthermore, we use an expert system to produce absence
48 locations. In order to ensure high quality for the environmental variables
49 associated to presence information, we use the Maximum Entropy (MaxEnt)

50 model (Phillips et al., 2006; Berger, 1996) as a filter to select the variables
51 that are important to define the potential habitat of the species. These are
52 the variables that are mostly correlated to the species observations, among
53 those we selected from reference studies. When possible, we make environ-
54 mental variables values range from 450 to 1000 m, encompassing the deep
55 ocean waters usually inhabited by *A.dux* (Guerra et al., 2010). Finally, we
56 train an Artificial Neural Network on these datasets and compare the results
57 with (i) a presence-only method, (ii) an expert system and (iii) an expert
58 drawn map.

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

66 2. Overview

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

71 2.1. Species overview

72 The *Architeuthis* genus has been recorded since before the 16th century
73 (Guerra et al., 2011), and has recently been observed live in its natural habi-
74 tat for the first time (Kubodera and Mori, 2005). Literature studies have
75 recognized up to five species of this genus (Robson, 1933), although Nesis
76 (Nesis, 1987) and Aldrich (Aldrich, 1991) suggested them to be identified as
77 *Architeuthis dux*. Most of the records refer to stranded animals or stomach
78 remains, and are located in the North Atlantic (e.g. Norway), in the North-
79 East Atlantic (off northern Spain), in the South Atlantic (e.g. Namibia and
80 South Africa) and in the South-West Pacific, around New Zealand and Tas-
81 mania (Gonzalez et al., 2000; Clarke, 2006; Förch, 1998; Guerra et al., 2004;
82 Bolstad and O’Shea, 2004; Guerra et al., 2004). Most of these animals have
83 been classified as *A. dux* (Cherel, 2003; Clarke, 2006; Bolstad and O’Shea,
84 2004; Guerra et al., 2010; Clarke, 2006; Nesis, 2003; Aldrich, 1991), but many
85 more refer to the genus level (*Architeuthis* spp.) without further specifica-
86 tion (Lordan et al., 1998; Gonzalez et al., 2000; Ré et al., 1998; Arfelli et al.,
87 1991; Kubodera and Mori, 2005; Roeleveld and Lipinski, 1991). In 2003,
88 Nesis (Nesis, 2003) published the distribution of *Architeuthis dux* by corre-
89 lating latitudinal zones and zoogeographic provinces in the pelagic realm.
90 The identified zonality mainly reflects the general oceanic circulation, and
91 no temperature data was used for the selection of the latitudinal zones. The
92 author identified rate of speciation among the Cephalopoda taxon caused by
93 climatic and orogenic isolation and bi-subtropical species of *Architeuthis dux*

94 in the North Atlantic, the South Pacific and the Southern Ocean. In this
95 paper, we take the map of Nesis as a reference to assess the performance of
96 our models.

97 Several authors have suggested that *Architeuthis* is an epipelagic/mesopelagic
98 species, living in correspondence with continental slopes, submarine channels
99 or canyons (Roeleveld and Lipinski, 1991; Kubodera and Mori, 2005). Guerra
100 et al. (Guerra et al., 2011) examined the relationship between the number
101 of recorded specimens and some of the main characteristics of the observa-
102 tion areas. The authors report the close association of giant squids with
103 sperm whales sights (Clarke and Pascoe, 1997). They indicate correlation of
104 *Architeuthis* spp. sighting with places presenting high primary production
105 and close to shallow fishing grounds. They also report low incidence of genus
106 sighting, in locations where deep channels or canyons are not present (Guerra
107 et al., 2004). On the basis of the distribution of the strandings, Robson (Rob-
108 son, 1933) noticed that *Architeuthis* is adapted to temperate waters of about
109 10 °C. This biological information is in agreement with later studies, that
110 correlate the giant squid presence with the increase of the temperature in
111 some locations (Brix, 1983; Guerra et al., 2004).

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

114 2.2. *Modelling approaches*

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

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

136 In this paper, we use different approaches to model the potential distri-

137 bution of *A. dux*. We take the AquaMaps expert system as reference for the
138 comparison. The AquaMaps algorithms (Kaschner et al., 2006, 2008) are
139 presence-only models that include scientific expert knowledge into species
140 habitats modelling (Ready et al., 2010). The AquaMaps algorithms include
141 two models: AquaMaps Suitable and AquaMaps Native, addressing the po-
142 tential and the actual distribution of a species respectively. Expert knowl-
143 edge is used in modelling species-habitat relations at global scale with 0.5°
144 resolution, relying on the following environmental variables: depth, salinity,
145 temperature, primary production, distance from land and sea ice concentra-
146 tion (Corsi et al., 2000). AquaMaps combines mechanistic assumptions and
147 automatic procedures for habitat parameters and species values estimations,
148 making the modelling approach usually reliable, but less accurate when ex-
149 pert knowledge at global scale is missing. In the experiment for this paper, we
150 used AquaMaps Native to produce absence locations and AquaMaps Suitable
151 as reference to assess the performance of the other models.

152 One largely used presence-only technique is Maximum Entropy (MaxEnt)
153 (Phillips et al., 2006; Phillips and Dudik, 2008). The general idea of MaxEnt
154 is to approximate a probability density function, defined on an environmental
155 features vectorial space, ensuring that this function is compliant with the
156 mean values at the presence locations, and that the entropy of the probability
157 distribution is maximum (Elith et al., 2011). The algorithm relies on unbiased
158 samples, so effort in collecting a set of high quality presence records is critical
159 to avoid estimation errors (Elith and Leathwick, 2009). We used MaxEnt as

160 a reference model to assess the performance of our approach. On the other
161 hand, MaxEnt is a fundamental part of our approach, because we used it to
162 help a presence-absence model by providing features that are important to
163 assess habitat suitability. We give more details about our MaxEnt usage in
164 Section 3.3.

165 Among the many presence/absence models, Artificial Neural Networks
166 (ANNs) have demonstrated to gain good performance with respect to other
167 approaches, especially for rare species (Pearson et al., 2002; Coro et al.,
168 2013b). ANNs try to automatically simulate the probability of occurrence of
169 a species, given certain environmental conditions. They learn on the basis of
170 the environmental characteristics of positive and negative examples. We used
171 ANNs to combine the outputs of our presence/absence data production and
172 of the environmental features filtering phase. In Section 3.5 we give details
173 about our usage of ANNs.

174 **3. Method**

175 In this Section we describe the technology which supported the exper-
176 iments, and we also report our procedures for data preparation and envi-
177 ronmental features selection. Furthermore, we explain our presence/absence
178 approach to model the distribution of *A. dux* and its relevance for other rare
179 species.

180 3.1. *Technology and tools*

181 Preparing an experimental setup to model the distribution of a rare
182 species requires expertise in several disciplines. The model requires highly
183 reliable presence records. The environmental features describing the ecolog-
184 ical niche of the species should be of high quality and with the appropriate
185 spatial resolution (Kamino et al., 2012; Elith and Leathwick, 2009). Since
186 environmental features are distributed as geospatial datasets, their projec-
187 tions should be perfectly aligned in order to correctly retrieve correspondent
188 values. During the training phase, different models need to be tested and
189 reapplied to avoid problems of local minimum of the fitting curve (Bishop,
190 1995) and if several models are combined, the output of a model must agree
191 with the input of the next.

192 We overcame these issues of high quality environmental features sets
193 and their alignment by using an e-Infrastructure for biodiversity conserva-
194 tion (D4science) (Candela et al., 2009). D4Science supplies several mod-
195 els as-a-service. The model compatibility is guaranteed by specialized e-
196 Infrastructure services. Furthermore, D4Science uses Cloud computing to
197 speed processing up (Coro et al., 2013b; Candela et al., 2013). D4Science
198 provides automatic alignment and comparison of geospatial datasets (Coro,
199 2014), by re-projecting environmental features into a common coordinates
200 system.

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

203 D4Science also allows retrieving species presence information from heteroge-
204 neous biodiversity data collections (e.g. OBIS (Berghe et al., 2010), GBIF
205 (Edwards et al., 2000) and the Catalog of Life (Wilson, 2003)), under the
206 same format (Candela et al., 2014). Information is attached to each presence
207 record, to indicate the ownership of the observation, its source (e.g. hu-
208 man observation, specimen etc.) and possibly if the record underwent expert
209 review.

210 3.2. Occurrence data preparation

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

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

239 Data retrieved using D4Science follow the Darwin Core format (Wiec-
 240 zorek et al., 2012) and can be provided as input to the D4Science models di-
 241 rectly. All models accept the same format of input data of presence records,
 242 which makes the data preparation phase faster.

243 3.3. Environmental data selection

244 The environmental characteristics in our model refer to geospatially ex-
 245 plicit chemical and physical measurements. During its training session, our
 246 model learns from positive and negative examples that are based only on en-

247 vironmental features. In the subsequent projection session, a real value from
248 0 to 1 is associated to several locations to assess their habitat suitability. A
249 well performing model is one having good projection on the locations of the
250 training set and, at the same time, not suffering of overfitting issues on the
251 training values (Bishop, 1995).

252 Environmental features selection requires attention (see Section 2.2) to
253 ensure they are not highly correlated: adding a feature that is dependent on
254 previous ones would not bring more information to the model, but it could
255 add noise during the training session. Furthermore, the spatial resolution
256 should fit the precision of the projection: a model that has to produce a map
257 with resolution 0.5 degrees, should rely on environmental information with
258 the same resolution. This allows not using values coming from rescaling pro-
259 cesses or kriging that would add uncertainty to the measurements. Global
260 scale maps also contain estimated values, but these have been produced by
261 experts. Thus, we recommend using the native resolution of the environ-
262 mental datasets in global scale modelling. Furthermore, the reliability of the
263 data is crucial. This depends on the data provider, as some providers require
264 the dataset to pass a data quality process in order to be published (e.g. My-
265 Ocean (Bahurel et al., 2010) and the World Ocean Atlas (Locarnini et al.,
266 2006)).

267 Features selection methods analyse the features space. Several approaches
268 try to reduce the dimensions of this space, for example by recovering the
269 most independent features or combining them into new features (Jolliffe,

270 2005; MacLeod, 2010). In our approach, instead, we wanted to reduce the
271 dimension of the number of features to use, but at the same time we wanted to
272 take the correlation between presence points and random points (background
273 points) into account. To such aim, we used the MaxEnt model as a features
274 filter.

275 We collected environmental features that could *a priori* influence the
276 habitat suitability for *A. dux*, according to the studies we have reported in
277 Section 2.1. We chose the parameters reported in Table 2, averaged on
278 annual values. Based on the depth range of our presence points and on
279 indications from literature (Guerra et al., 2010), we took parameters values
280 in the following ranges: (i) in the entire water column, (ii) averaged between
281 450 and 1000 meters, (iii) at surface level. In particular, we used the 450-1000
282 m range when the data provider reported information at several depth ranges.
283 Table 2 indicates the ranges we used for each parameter. The parameters
284 layers come with different projections and reference systems, but the MaxEnt
285 implementation on D4Science automatically accounts for making the layers
286 projections and reference systems uniform, before training the models. In
287 our experiments, the layers from MyOcean and the World Ocean Atlas were
288 available in the e-Infrastructure as GIS layers, while we provided the others
289 as external datasets, in one of the accepted D4Science input formats (Coro,
290 2014).

291 During the training phase, MaxEnt minimizes the relative entropy of the
292 features at the presence locations, with respect to the features of random

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

309 3.4. *Absence points*

310 In order to improve data quality, we searched for a method to produce
 311 robust absence locations. Several methods exist to estimate absence locations
 312 (Pearson, 2012), but we avoided introducing biases by using other machine
 313 learning models. One approach that proved to be effective, is to use an expert
 314 system to generate absence locations (Coro et al., 2013b,a). Expert systems

315 combine automatic processing with expert indications and can be used to
316 simulate expert opinion. Thus, we used AquaMaps Native (see Section 2.2)
317 to retrieve absence areas by looking at locations having probability lower than
318 0.2 but higher than 0. Setting the threshold over zero, selects areas having
319 low values for several environmental envelopes. This approach simulates
320 locations where an expert asserts that the habitat is particularly unsuited
321 for the species. Furthermore, these locations are reported at a relatively
322 high resolution of 0.5 degrees at global scale. From the AquaMaps Native
323 distribution, we extracted absence scattered locations, because this allows
324 having a wider range of environmental characteristics for low probability
325 locations. We took only absences that were two degrees distant at least. In
326 another work (Coro et al., 2013a), we demonstrated that this method results
327 in better performance than using concentrated absence records.

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

334 3.5. Modelling

335 In order to produce distribution maps for *Architeuthis dux*, we used both
336 MaxEnt and Artificial Neural Networks. As input data, we used the pres-

ence 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

360 identified a location viable for *A. dux* to a certain degree. The resulting
361 distribution map is displayed in Figure 2.

362 Artificial Neural Networks, in particular Feed Forward Neural Networks
363 (FFNNs) (Bebis and Georgiopoulos, 1994), have proven good performance
364 in niche modelling and have been applied to model the distribution of rare
365 species (Pearson, 2012; Coro et al., 2013b). Furthermore, with respect to al-
366 ternative models, they have proven to perform better when the quality of the
367 data is high (Coro et al., 2013b). The aim of an FFNN is to build a hierarchi-
368 cal multi-layered network, made up of interconnected nodes, which simulates
369 a complex function. The complexity of the function depends on the number
370 of layers and neurons in the network. During a training session, the weights
371 of the network connections are adapted to produce expected values on the
372 training dataset. In our case, the training set consisted of the environmental
373 features at presence and absence locations, where features were extracted
374 at 1-degree resolution. The FFNN performance depends only on the values
375 assumed by the features on the training set, differently from MaxEnt. For
376 presences, the expected value was set to 1 and for absences it was set to
377 0. In order to define the optimal number of layers and neurons per layer
378 to use in the network, we adopted a *growing* strategy (Bishop, 1995). We
379 added neurons and layers as far as the error with respect to the training set
380 decreased after a training session (up to a certain threshold). The threshold
381 was empirically set to 0.01 in order to avoid overfitting. We executed the
382 Network training session 10 times for each topology and eventually took the

383 one with the best learning result, i.e. with the lowest mean error with respect
 384 to the training points. This process ended in two Networks achieving good
 385 learning capacity: one having two layers, with 10 neurons in the first layer
 386 and 2 in the second, the other having two layers too, with 100 neurons in
 387 the first layer and 2 in the second. We will refer to the first as the “simple
 388 topology FFNN” and to the second as the “complex topology FFNN”. One
 389 characteristic of the second FFNN is that the learning process is more stable,
 390 i.e. it usually ends in the same distance from the training set. On the other
 391 hand, using simpler topologies is better especially to avoid overfitting issues.
 392 Indeed, in Section 4 we demonstrate that the simpler topology gains overall
 393 better performance. In the same way we did for MaxEnt, we calculated that
 394 for the FFNNs the best threshold to filter out too low habitat suitability was
 395 0.1. Figure 3 reports the maps associated to the two FFNN topologies when
 396 we projected the models at global scale, with 1-degree resolution.

397 *3.6. Applicability to other species*

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

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

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

414 **4. Results**

415 In this Section we describe the qualitative and quantitative approaches
416 we used to compare the trained models with existing literature data. First,
417 we report a “qualitative” comparison on coarse presence locations reported in
418 literature for *Architeuthis dux* and *Architeuthis* spp. In order to investigate
419 the differences between the models in detail, we also report the results of a
420 quantitative comparison, with respect to a map drawn by an expert (Nesis,
421 2003).

422 *4.1. Qualitative evaluation*

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

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

442 When we set a probability threshold to filter out values lower than 0.8,
 443 the maps highlight only the places with high habitat suitability. In this case,
 444 the results of the assessments by the models are reported in Table 4. We
 445 notice that the FFNN with the simple topology and AquaMaps Suitable still
 446 present high performance. In particular, the FFNN predicts species presence
 447 in Newfoundland, Norway Sea, South America, South-Eastern Africa and in
 448 the Mediterranean Sea. Conversely, the AquaMaps Suitable model covers

449 the Eastern-North Atlantic, the Kerguelen Islands, the New Zealand coasts
450 and the Tasman Sea. Using this probability threshold, the complex topology
451 FFNN and the MaxEnt model predict very few suitable areas, especially
452 for *Architeuthis* spp. This means that, overall, the FFNN with the simple
453 topology is more stable and reliable. One evident difference between the
454 FFNNs and the AquaMaps model is that, according to AquaMaps, the species
455 is not present in open ocean but only prefers coastal areas. In order to explore
456 more such difference, we used a quantitative discrepancy analysis.

457 4.2. Quantitative evaluation

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

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

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

483 Table 5 reports the performance using several thresholds: 0.8, 0.5 and
484 0.3. Furthermore, we made three comparisons with the map of Nesis using
485 presence-only, absence-only and presence-absence polygons separately. In
486 this way we observed that, even if one model can be in good agreement with
487 either presences or absences, it can be in lower agreement with respect to
488 both. The FFNN with the simple topology has lower agreement with absence
489 locations, but overall is the closest to the expert drawn map, according to
490 all the probability thresholds.

491 5. Discussion

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

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

511 The FFNN is the model that better resembles the expert's map, but more
512 information is needed to ensure its reliability: there are some discrepancy lo-
513 cations, like the South Africa coasts, that need further investigation. The

514 highest discrepancy with respect to the expert’s map is in the South-West
 515 coast of South Africa, in the Indian Ocean and in North Australia. This
 516 discrepancy could be explained by the fact that the FFNN predicts potential
 517 habitat, while the expert indicates the known (actual) habitat. On the other
 518 hand, there are studies supporting the indications by the FFNN map: *Architeuthis*
 519 specimens were captured in South-West Pacific Ocean, and around
 520 Australian coasts, especially off the West coasts (Jackson, 1991; Sweeney and
 521 Roper, 2001). As for the Indian Ocean, several studies report the presence of
 522 *Architeuthis* near the Reunion Island, the Mauritius Islands and generally in
 523 the South-Western Indian Ocean (Sweeney and Roper, 2001; Guerra et al.,
 524 2011; Cherel, 2003; Mikhalev et al., 1981). In some Indian survey works, it
 525 is reported that *Architeuthis* species are present off the west coasts of India
 526 (Silas, 1968, 1985).

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

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

536 6. Conclusions

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

556 In summary, maximising the reliability of presence, absence and environ-
557 mental parameters gives good estimate of the distribution of *A. dux*. This
558 maximisation determines reliable patterns of occurrence related to environ-

559 mental gradients, as also supported by other studies (Segurado and Araujo,
560 2004; Franklin, 2010). A large scale distribution for *A. dux* can also help
561 understanding the role of this species on a broader geographic perspective
562 (Lordan et al., 2001).

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

575 Generally speaking, the presented work can be useful in species conser-
576 vation. In fact, model-based approaches for rare species that count on data
577 quality have proved to be valuable when used in population management
578 and conservation strategies (Austin, 2007). In particular, many conserva-
579 tion projects need a complete description of species' geographical distribu-
580 tions, and modelling techniques (e.g. MaxEnt, Artificial Neural Networks
581 and AquaMaps) have already proved to reliably support this activity (Fice-

582 tola et al., 2007; Ward, 2007; Hijmans and Graham, 2006; Fitzpatrick et al.,
583 2008; Thorn et al., 2009; Wollan et al., 2008; Echarri et al., 2009; Cordellier
584 and Pfenninger, 2009). The produced maps can be also used in fisheries,
585 because producing a potential distribution for a rare species like the giant
586 squid can help locating vulnerable marine ecosystems (Auster et al., 2010;
587 Stevens et al., 2000; Tittensor et al., 2009; Stevens et al., 2000).

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

593 **Acknowledgments**

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

598 **References**

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

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

605 Auster, P. J., Gjerde, K., Heupel, E., Watling, L., Grehan, A., Rogers, A. D.,
 606 2010. Definition and detection of vulnerable marine ecosystems on the high
 607 seas: problems with the “move-on” rule. *ICES Journal of Marine Science:*
 608 *Journal du Conseil*, fsq074.

609 Austin, M., 2007. Species distribution models and ecological theory: a crit-
 610 ical assessment and some possible new approaches. *Ecological modelling*
 611 200 (1), 1–19.

612 Bahurel, P., Adragna, F., Bell, M. J., Jacq, F., Johannessen, J. A., Le Traon,
 613 P.-Y., Pinardi, N., She, J., 2010. Ocean monitoring and forecasting core
 614 services: The european myocean example. *Proceedings of OceanObs 9*,
 615 02.

616 Bebis, G., Georgiopoulos, M., 1994. Feed-forward neural networks. *Poten-*
 617 *tials, IEEE* 13 (4), 27–31.

618 Berger, A., 1996. A brief maxent tutorial. [http://www-](http://www-2.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html)
 619 [2.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html](http://www-2.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html)
 620 25.

621 Berghe, E. V., Stocks, K. I., Grassle, J. F., 2010. Data integration: The ocean

622 biogeographic information system. Life in the World's Oceans: Diversity,
623 Distribution, and Abundance, 333.

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

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

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

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

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

637 Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F., 2013. Species
638 distribution modeling in the cloud. Concurrency and Computation: Prac-
639 tice and Experience, n/a.
640 URL <http://dx.doi.org/10.1002/cpe.3030>

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

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

646 Castelli, D., Pagano, P., Candela, L., Coro, G., 2013. The imarine data
647 bonanza: Improving data discovery and management through an hybrid
648 data infrastructure. *Bollettino di Geofisica Teorica ed Applicata* 54, 105–
649 107.

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

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

656 Clarke, M., 2006. Oceanic cephalopod distribution and species diversity in
657 the eastern north atlantic malcolm r. clarke. *Life and marine Sciences* 23A,
658 27–46.

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

662 Cordellier, M., Pfenninger, M., 2009. Inferring the past to predict the future:
663 climate modelling predictions and phylogeography for the freshwater gas-

664 tropod radix balthica (pulmonata, basommatophora). Molecular Ecology
665 18 (3), 534–544.

666 Coro, G., 2014. The D4Science MaxEnt implementation. [http://wiki.i-](http://wiki.i-marine.eu/index.php/MaxEnt)
667 [marine.eu/index.php/MaxEnt](http://wiki.i-marine.eu/index.php/MaxEnt).

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

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

674 Coro, G., Pagano, P., Ellenbroek, A., 2014. Comparing heterogeneous dis-
675 tribution maps for marine species. GIScience & Remote Sensing 51 (5),
676 593–611.

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

680 Echarri, F., Tambussi, C., Hospitaleche, C. A., 2009. Predicting the distribu-
681 tion of the crested tinamous, eudromia spp.(aves, tinamiformes). Journal
682 of Ornithology 150 (1), 75–84.

683 Edwards, J. L., Lane, M. A., Nielsen, E. S., 2000. Interoperability of bio-

684 diversity databases: Biodiversity information on every desktop. *Science*
685 289 (5488), 2312–2314.

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

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

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

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

697 Ficetola, G. F., Thuiller, W., Miaud, C., 2007. Prediction and valida-
698 tion of the potential global distribution of a problematic alien invasive
699 species—the american bullfrog. *Diversity and Distributions* 13 (4), 476–
700 485.

701 Fitzpatrick, M. C., Gove, A. D., Sanders, N. J., Dunn, R. R., 2008. Cli-
702 mate change, plant migration, and range collapse in a global biodiversity
703 hotspot: the banksia (proteaceae) of western australia. *Global Change Bi-*
704 *ology* 14 (6), 1337–1352.

- 705 Förch, E., 1998. The marine fauna of New Zealand: Cephalopoda: Oegop-
 706 sida: Architeuthidae (Giant Squid). NIWA Biodiversity Memoir 110, 1–
 707 113.
- 708 Franklin, J., 2010. Mapping species distributions: spatial inference and pre-
 709 diction. Cambridge University Press.
- 710 Gibson, L., Barrett, B., Burbidge, A., 2007. Dealing with uncertain absences
 711 in habitatmodelling: a case study of a rare ground-dwelling parrot. Diver-
 712 sity and Distributions 13, 704–713.
- 713 Gonzalez, M., Fernandez-Casado, M., Rodriguez, P., Segura, A., Martin,
 714 J. J., 2000. First record of the giant squid *Architeuthis* sp . (*Architeuthi-*
 715 *dae*) in the Mediterranean Sea “. *J.Mar.Biol.Ass.U.K.* 80, 745–746.
- 716 Guerra, A., Gonzalez, A. F., Dawe, E. G., Rocha, F., 2004. Records of giant
 717 squid in the north-eastern Atlantic, and two records of male *Architeuthis*
 718 sp. of the Iberian Peninsula. *J.Mar.Biol.Ass.U.K.* 84, 426–431.
- 719 Guerra, A., González, A. F., Pascual, S., Dawe, E. G., 2011. The giant squid
 720 *Architeuthis*: An emblematic invertebrate that can represent concern for
 721 the conservation of marine biodiversity. *Biological Conservation* 144 (7),
 722 1989–1997.
- 723 Guerra, Á., Rodríguez-Navarro, A. B., González, Á. F., Romanek, C. S.,
 724 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

744 and sepioid populations in waters off townsville, north queensland, aus-
745 tralia. Ph.D. thesis, James Cook University.

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

747 Kamino, L., Stehmann, J., Amaral, S., De Marco, P., Rangel, T., de Siqueira,
748 M., De Giovanni, R., Hortal, J., 2012. Challenges and perspectives for
749 species distribution modelling in the neotropics. *Biology letters* 8 (3), 324–
750 326.

751 Kaschner, K., Ready, J., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood,
752 P., South, A., Kullander, S., Rees, T., Close, C., et al., 2008. Aquamaps:
753 Predicted range maps for aquatic species. World wide web electronic pub-
754 lication, [www. aquamaps. org](http://www.aquamaps.org), Version 8, 2010.

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

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

760 Kubodera, T., Mori, K., 2005. First-ever observations of a live giant squid
761 in the wild. *Proceedings of the Royal Society B: Biological Sciences*
762 272 (1581), 2583–2586.

763 Locarnini, R. A., Mishonov, A., Antonov, J., Boyer, T., Garcia, H., Levitus,

764 S., et al., 2006. World ocean atlas 2005 volume 1: Temperature [+ dvd].
765 Noaa atlas nesdis 61 (1).

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

770 Lordan, C., Collins, M., Perales-Raya, C., 1998. Observations on Morphology,
771 Age and Diet of Three *Architeuthis* Caught Off the West Coast of Ireland in
772 1995. Journal of the Marine Biological Association of the United Kingdom
773 78, 903–917.

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

778 Lyons, J., Navarro-Perz, S., Cochran, P. A., Santana-C., E., Guzman-Arroyo,
779 M., 1995. Index of biotic integrity based on fish assemblages for the con-
780 servation of streams and rivers in west-central Mexico. Conserv. Biol. 9,
781 569–584.

782 MacLeod, C. D., Jun. 2010. Habitat representativeness score (HRS): a novel
783 concept for objectively assessing the suitability of survey coverage for mod-

784 elling the distribution of marine species. Journal of the Marine Biological
 785 Association of the United Kingdom 90 (07), 1269–1277.

786 Márcia Barbosa, Real, R., Olivero, J., Mario Vargas, J., Dec. 2003. Ot-
 787 ter (*Lutra lutra*) distribution modeling at two resolution scales suited to
 788 conservation planning in the Iberian Peninsula. Biological Conservation
 789 114 (3), 377–387.

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

793 Nesis, K. N., 1987. Cephalopods of the World. TFH Publications, Inc, Nep-
 794 tune City, New Jersey Nesis.

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

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

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

801 Pearson, R. G., 2012. Species distribution modeling for conservation educa-
 802 tors and practitioners. Synthesis. American Museum of Natural History.
 803 Available at <http://ncep.amnh.org>.

- 804 Pearson, R. G., Dawson, T. P., Berry, P. M., Harrison, P., 2002. SPECIES:
805 A Spatial Evaluation of Climate Impact on the Envelope of Species. Eco-
806 logical Modelling 154 (3), 289–300.
- 807 Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy
808 modeling of species geographic distributions. Ecological Modelling 190 (3-
809 4), 231–259.
- 810 Phillips, S. J., Dudik, M., 2008. Modeling of species distributions with Max-
811 ent: new extensions and a comprehensive evaluation. Ecography 31, 161–
812 175.
- 813 Pulliam, H., 2000. On the relationship between niche and distribution. Ecol-
814 ogy Letters 3, 349–361.
- 815 Quantum GIS, 2011. Quantum GIS Geographic Information System. Open
816 Source Geospatial Foundation Project. URL:[<http://qgis.osgeo.org>].
- 817 Razgour, O., Hanmer, J., Jones, G., 2011. Using multi-scale modelling to
818 predict habitat suitability for species of conservation concern: the grey
819 long-eared bat as a case study. Biological Conservation 144 (12), 2922–
820 2930.
- 821 Ré, M., Baron, P., Beron, J., Gosztonyi, A., Kuba, L., Monsalve,
822 M., Sardella, N., 1998. A giant squid *Architeuthis* sp. (Mol-lusca,
823 Cephalopoda) stranded on the Patagonian shore of Argentina. Cephalo-
824 pod Biodiversity, Ecology and Evolution. S. Afr. J. Mar. Sci. 20, 109–122.

- 825 Ready, J., Kaschner, K., South, A. B., Eastwood, P. D., Rees, T., Rius, J.,
826 Agbayani, E., Kullander, S., Froese, R., 2010. Predicting the distributions
827 of marine organisms at the global scale. *Ecological Modelling* 221 (3), 467
828 – 478.
- 829 Rebelo, H., Jones, G., 2010. Ground validation of presence-only modelling
830 with rare species: a case study on *barbastelles barbastella barbastellus*
831 (chiroptera: Vespertilionidae). *Journal of Applied Ecology* 47 (2), 410–
832 420.
- 833 Robson, G., 1933. On *architeuthis clarkei*, a new species of giant squid, with
834 observations on the genus. In: *Proceedings of the Zoological Society of*
835 *London*. Vol. 103. Wiley Online Library, pp. 681–697.
- 836 Roeleveld, M., Lipinski, M., 1991. The giant squid *Architeuthis* in southern
837 African waters. *J.Zool.Lond.* 224, 431–477.
- 838 Segurado, P., Araujo, M. B., 2004. An evaluation of methods for modelling
839 species distributions. *Journal of Biogeography* 31 (10), 1555–1568.
- 840 Silas, E., 1968. Cephalopoda of the west coast of india collected during the
841 cruises of the research vessel varuna, with a catalogue of the species known
842 from the indian ocean. In: *Proceedings of the Symposium on Mollusca*.
843 Vol. 1. pp. 277–359.
- 844 Silas, E., 1985. Cephalopod fisheries of india—an introduction to the subject
845 with methodologies adopted for this study. *CMFRI Bulletin* 37, 1–4.

846 Stevens, J., Bonfil, R., Dulvy, N., Walker, P., 2000. The effects of fishing on
847 sharks, rays, and chimaeras (chondrichthyans), and the implications for
848 marine ecosystems. *ICES Journal of Marine Science: Journal du Conseil*
849 57 (3), 476–494.

850 Sweeney, M., Roper, C., 2001. Records of architeuthis specimens from pub-
851 lished reports. Available online at: www.mnh.si.edu/cephs/archirec.pdf.

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

854 Thorn, J. S., Nijman, V., Smith, D., Nekaris, K., 2009. Ecological niche mod-
855 elling as a technique for assessing threats and setting conservation priorities
856 for asian slow lorises (primates: *Nycticebus*). *Diversity and Distributions*
857 15 (2), 289–298.

858 Tittensor, D. P., Baco, A. R., Brewin, P. E., Clark, M. R., Consalvey, M.,
859 Hall-Spencer, J., Rowden, A. A., Schlacher, T., Stocks, K. I., Rogers, A. D.,
860 2009. Predicting global habitat suitability for stony corals on seamounts.
861 *Journal of Biogeography* 36 (6), 1111–1128.

862 Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F.,
863 De Clerck, O., 2012. Bio-oracle: a global environmental dataset for marine
864 species distribution modelling. *Global Ecology and Biogeography* 21 (2),
865 272–281.

- 866 Ward, D. F., 2007. Modelling the potential geographic distribution of invasive
867 ant species in new zealand. *Biological Invasions* 9 (6), 723–735.
- 868 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R.,
869 Robertson, T., Vieglais, D., 2012. Darwin core: An evolving community-
870 developed biodiversity data standard. *PLoS One* 7 (1), e29715.
- 871 Wilson, E. O., 2003. The encyclopedia of life. *Trends in Ecology & Evolution*
872 18 (2), 77–80.
- 873 Wisz, M. S., Hijmans, R., Li, J., Peterson, A. T., Graham, C., Guisan,
874 A., 2008. Effects of sample size on the performance of species distribution
875 models. *Diversity and Distributions* 14 (5), 763–773.
- 876 Wollan, A. K., Bakkestuen, V., Kauserud, H., Gulden, G., Halvorsen, R.,
877 2008. Modelling and predicting fungal distribution patterns using herbar-
878 ium 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 Zoology 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°	μ mol/l	World Ocean Atlas
Mole concentration of Silicate (avg 450-1000 m)	1°	μ 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	<i>A. duar</i>	1	0	0	1
NEW ZEALAND-TASMAN SEA	<i>A. duar</i>	1	1	1	1
BAY OF BISCAY	<i>A. duar</i>	0	1	1	1
NORTH-EAST ATLANTIC	<i>A. duar</i>	0	0	1	1
NEWFOUNDLAND	<i>A. duar</i>	1	1	1	0
NORWEGIAN SEA	<i>A. duar</i>	1	1	1	1
IRELAND COASTS	<i>Architeuthis spp.</i>	1	1	1	1
PATAGONIA	<i>Architeuthis spp.</i>	1	1	1	0
BRAZIL	<i>Architeuthis spp.</i>	1	1	1	1
JAPAN	<i>Architeuthis spp.</i>	1	1	1	1
SOUTH AFRICA-ORANGE RIVER	<i>Architeuthis spp.</i>	0	0	0	0
SOUTH AFRICA-TABLE BAY	<i>Architeuthis spp.</i>	0	0	0	0
SOUTH AFRICA-DURBAN	<i>Architeuthis spp.</i>	1	1	1	1
FUENGIROLA BEACH-MEDITERRANEAN SEA	<i>Architeuthis spp.</i>	1	1	1	1
ACCURACY		71.4%	71.4%	78.6%	71.4%

Table 3: Comparison between the predictions of *Architeuthis duar* 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	<i>A. dux</i>	0	0	0	1
NEW ZEALAND-TASMAN SEA	<i>A. dux</i>	0	0	0	1
BAY OF BISCAY	<i>A. dux</i>	0	0	0	1
NORTH-EAST ATLANTIC	<i>A. dux</i>	0	0	0	1
NEWFOUNDLAND	<i>A. dux</i>	0	1	0	0
NORWEGIAN SEA	<i>A. dux</i>	0	1	1	1
IRELAND COASTS	<i>Architeuthis spp.</i>	0	0	0	1
PATAGONIA	<i>Architeuthis spp.</i>	0	1	0	0
BRAZIL	<i>Architeuthis spp.</i>	0	1	0	1
JAPAN	<i>Architeuthis spp.</i>	1	1	0	1
SOUTH AFRICA-ORANGE RIVER	<i>Architeuthis spp.</i>	0	0	0	0
SOUTH AFRICA-TABLE BAY	<i>Architeuthis spp.</i>	0	0	0	0
SOUTH AFRICA-DURBAN	<i>Architeuthis spp.</i>	0	1	0	1
FUENGIROLA BEACH-MEDITERRANEAN SEA	<i>Architeuthis spp.</i>	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.

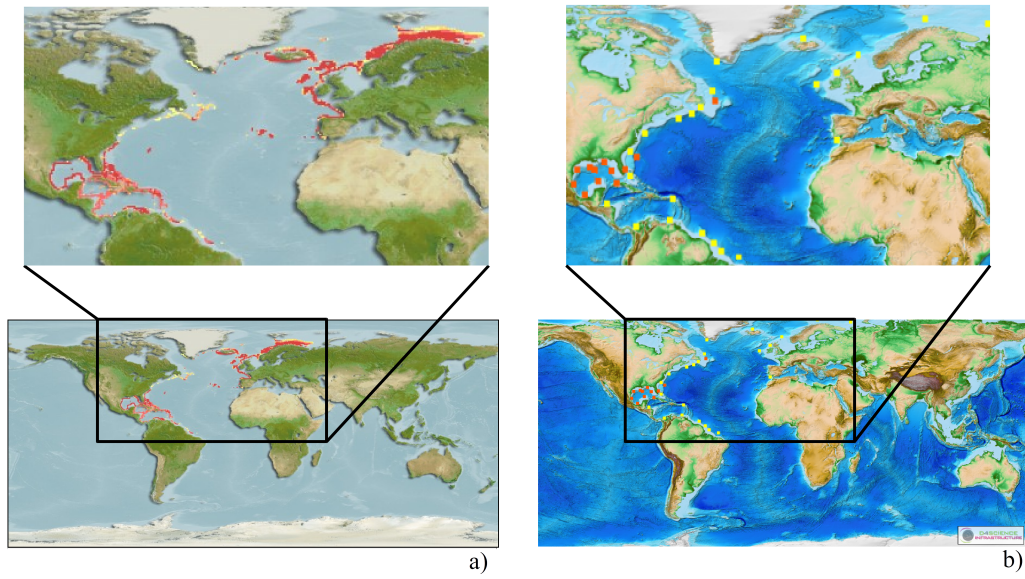


Figure 1: a. The AquaMaps Native distribution for *Architeuthis dux*. Darker colours refer to higher probability locations. b. The presences/absence points resulting from our process. Darker colours refer to presence locations.

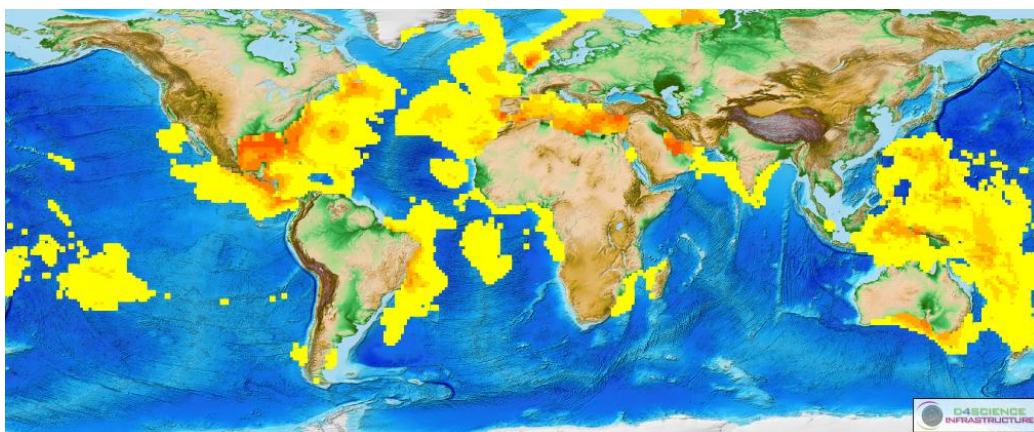


Figure 2: Distribution of *A. dux* produced with the MaxEnt model, trained using our filtered environmental features.

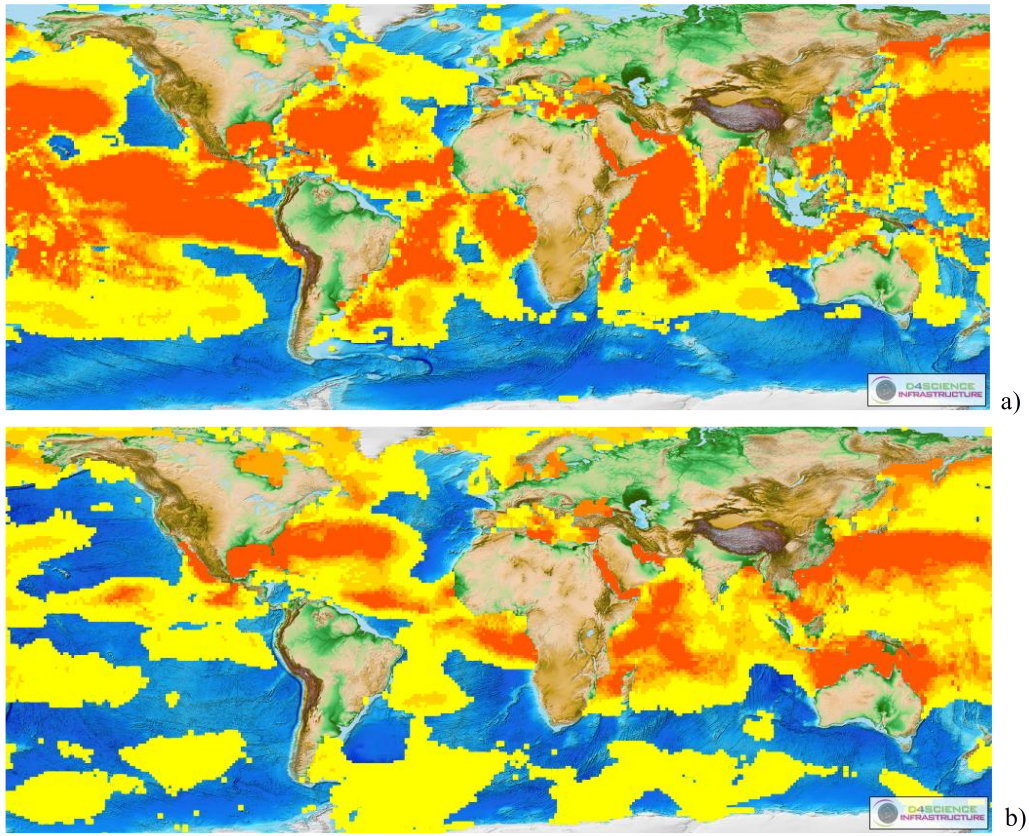


Figure 3: Distribution of *A. dux* produced by two Artificial Feed Forward Neural Networks: (a) with 2 layers, containing 10 neurons in the first layer and 2 in the second; (b) with 2 layers, containing 100 neurons in the first layer and 2 in the second.

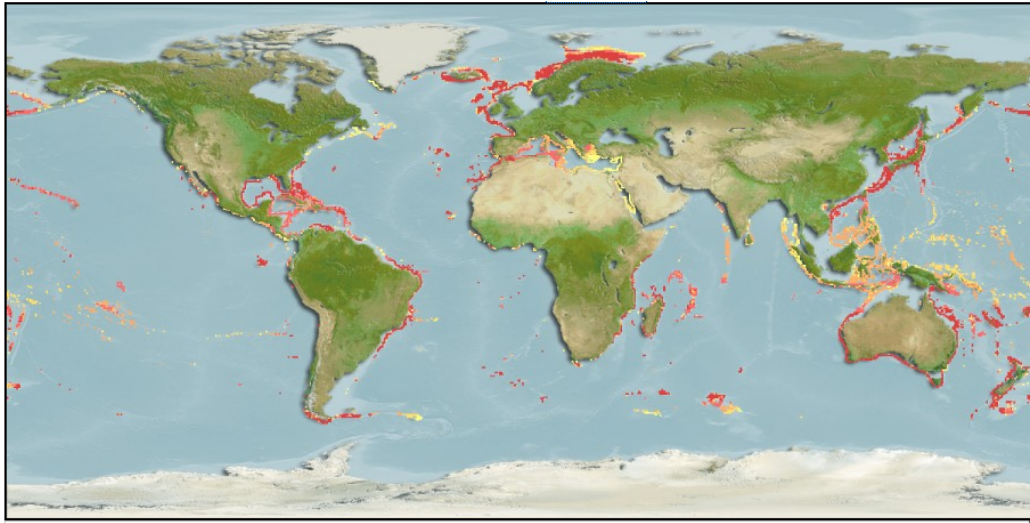


Figure 4: Distribution of *A. dux* produced with the AquaMaps Suitable model (Kaschner et al., 2008).

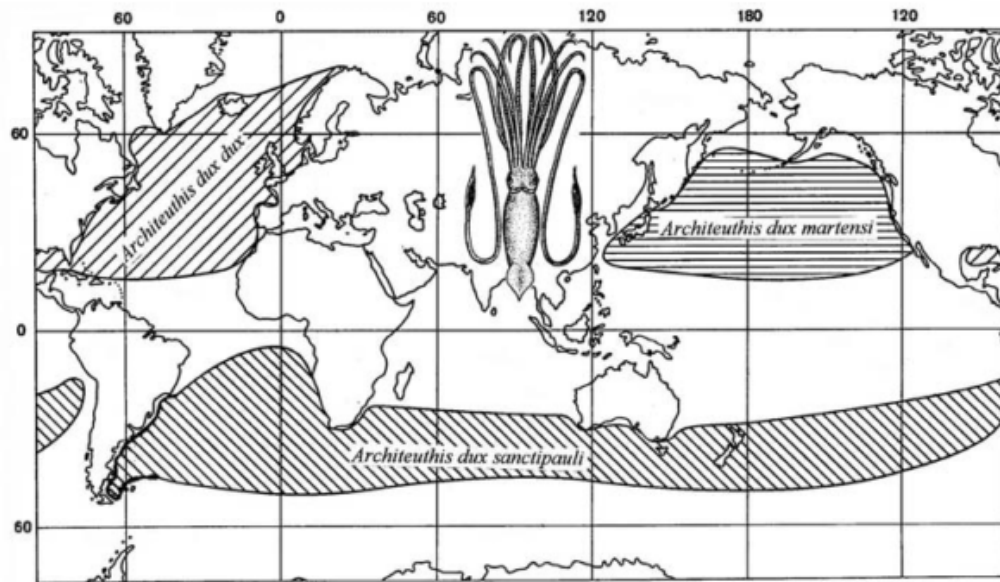


Figure 5: Distribution of *A. dux* reported by Nesis (2003).