Modelling European small pelagic fish distribution: Methodological insights

Alexandre Schickele, Boris Leroy, Gregory Beaugrand, Eric Goberville, Tarek Hattab, Patrice Francour, Virginie Raybaud

HAL Id: hal-02514810

https://hal.science/hal-02514810

Submitted on 6 Jan 2021

1 **Title: Modelling European small pelagic fish distribution: methodological insights**

2

3 **Authors:** Alexandre Schickele[1*], Boris Leroy[2], Gregory Beaugrand[3,4], Eric Goberville[2], Tarek

4 Hattab[5], Patrice Francour[1§] & Virginie Raybaud[1§]

5

6 [§]*Both authors contributed equally to this work.*

7 *Corresponding author*

8 [1]*Université Côte d'Azur, CNRS, UMR 7035 ECOSEAS, Nice, France.*

9 [2]*Unité Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), Muséum National*

10 *d'Histoire Naturelle, Sorbonne Université, Université de Caen Normandie, Université des*

11 *Antilles, CNRS, IRD, Paris, France.*

12 [3]*CNRS, Univ. Lille, Univ. Littoral Côte d'Opale, UMR 8187, LOG, Laboratoire d'Océanologie*

13 *et de Géosciences, Wimereux, France.*

14 [4]*Sir Alister Hardy Foundation for Ocean Science, The Laboratory, Citadel Hill, Plymouth,*

15 *United Kingdom.*

16 [5]*MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Avenue Jean Monnet, Sète, France*

17

18 **Authors' additional information:**

19 Alexandre Schickele: alex.schickele@gmail.com (ORCID ID: 0000-0002-4842-6890)

20 Boris Leroy: boris.leroy@mnhn.fr (ORCID ID : 0000-0002-7686-4302)

21 Gregory Beaugrand: gregory.beaugrand@univ-lille1.fr (ORCID ID: 0000-0002-0712-5223)

22 Eric Goberville: eric.goberville@upmc.fr (ORCID ID: 0000-0002-1843-7855)

23 Tarek Hattab: tarek.hattab@ifremer.fr (ORCID ID: 0000-0002-1420-5758)

24 Patrice Francour: patrice.francour@univ-cotedazur.fr (ORCID ID: 0000-0002-8489-1319)

25 Virginie Raybaud: virginie.raybaud@univ-cotedazur.fr (ORCID ID: 0000-0001-6426-3326)

ABSTRACT

The distribution of marine organisms is strongly influenced by climatic gradients worldwide. The ecological niche (*sensu* Hutchinson) of a species, i.e. the combination of environmental tolerances and resources required by an organism, interacts with the environment to determine its geographical range. This duality between niche and distribution allows climate change biologists to model potential species' distributions from past to future conditions. While species distribution models (SDMs) have been intensively used over the last years, no consensual framework to parametrise, calibrate and evaluate models has emerged. Here, to model the contemporary (1990-2017) spatial distribution of seven highly harvested European small pelagic fish species, we implemented a comprehensive and replicable numerical procedure based on 8 SDMs (7 from the Biomod2 framework plus the NPPEN model). This procedure considers critical issues in species distribution modelling such as sampling bias, pseudo-absence selection, model evaluation and uncertainty quantification respectively through (i) an environmental filtration of observation data, (ii) a convex hull based pseudo-absence selection, (iii) a multi-criteria evaluation of model outputs and (iv) an ensemble modelling approach. By mitigating environmental sampling bias in observation data and by identifying the most ecologically relevant predictors, our framework helps to improve the modelling of fish species' environmental suitability. Not only average temperature, but also temperature variability appears as major factors driving small pelagic fish distribution, and areas of highest environmental suitability were found along the north-western Mediterranean coasts, the Bay of Biscay and the North Sea. We demonstrate in this study that the use of appropriate data pre-processing techniques, an often-overlooked step in modelling, increase model predictive performance, strengthening our confidence in the reliability of predictions.

52

53   **Highlights:**

54   • Temperature (mean and variability) are major species distribution drivers

55   • Small pelagic species have high environmental suitability along western Europe

56   • Environmental filtering compensates spatial sampling bias

57   • The convex hull method is a robust pseudo-absence selection strategy

58

1. INTRODUCTION

Fish species distribution and assemblages are strongly influenced by both climatic and physical gradients (Ben Rais Lasram et al. 2010, Beaugrand et al. 2011, Raybaud et al. 2017). Temperature is known as a master parameter driving fish distribution at a macroecological level (Lenoir et al. 2011, Beaugrand et al. 2018). This parameter influences a large range of biological processes such as growth, reproduction, larval development, recruitment, and act as a major stressing factor depending on species thermal tolerance (psychrophile or thermophile species; Angilletta 2011, Beaugrand and Kirby 2018). Salinity, oxygen concentration, primary production (that are indirectly influenced by changes in temperature; e.g. Kirby and Beaugrand 2009) or the physical habitat (e.g. sediment type; Poloczanska et al. 2013) may also highly influence marine fish species at different spatial scales.

Hutchinson (1957) conceptualised the ecological niche as the "n-dimensional ensemble of environmental conditions that enable a species to live and reproduce" and subsequently made a distinction between the fundamental and the realised niche (Hutchinson 1978). Due to biotic interactions, dispersal limitation and/or historical factors (Soberon and Peterson 2005), species generally occupy only their realised niche, i.e. the subset of their fundamental niche that represents the response of all physiological processes of a species to the synergistic effects of environmental conditions (Helaouet and Beaugrand 2009, Beaugrand et al. 2013). By defining the niche as an attribute of species instead of a portion of the environment, the Hutchinson's concept enables duality between niche and distribution (Pulliam 2000, Colwell and Rangel 2009). Such a relationship is of major interest in biogeography as each georeferenced species occurrence, i.e. where a given species has been observed, can be related to several environmental parameters such as temperature, salinity and primary production. When species are in equilibrium with their environment, associating environmental conditions and observed

84   distributions permits climate change biologists to estimate species' potential niche (Jiménez-

85   Valverde et al. 2008).

86

87   The relationship between species occurrences, environmental conditions and species'

88   potential niche has become intensively studied over the last two decades, using a wide range of

89   modelling techniques - hereafter referred to as Species Distribution Models (SDMs) to assess

90   past, contemporary and future species distribution in both marine and terrestrial ecosystems

91   (e.g. Cheung et al. 2009, Bellard et al. 2016, Cristofari et al. 2018). SDMs rely on several

92   ecological assumptions, such as species distribution in equilibrium or habitat saturation

93   (Soberon and Peterson 2005), niche conservatism (Crisp et al. 2009), unlimited dispersal

94   abilities (Wiens et al. 2009) or the non-influential role of biotic interactions in shaping large-

95   scale distributions (i.e. the Gleasonian vision of biotic communities; Gleason 1926, Guisan and

96   Thuiller 2005, Wiens et al. 2009). Superimposed to these assumptions, several sources of errors

97   and uncertainties may lead to variation – sometimes conflicting – in the outputs of SDMs for a

98   given species (Beaumont et al. 2008): (i) accuracy of observation data and (ii) lack of true

99   absences (Proosdij et al. 2016), (iii) identification of ecologically meaningful environmental

100   predictors with high explanatory power (Guisan and Thuiller 2005), (iv) choice of the modelling

101   algorithm (Buisson et al. 2010) and (v) SDMs' evaluation processes (Leroy et al. 2018). While

102   tremendous progresses have been made on both the building and evaluation of SDMs in recent

103   years with a plethora of new methods for modelling species' distribution (Araújo and Guisan

104   2006, Leroy et al. 2018, Støa et al. 2018), the development of further procedures is still required

105   for improving the quality of SDMs.

106

107   Species distribution models are known to be very sensitive to different sources of

108   uncertainties and sustained attention should be devoted to each step of the modelling procedure,

from the pre-processing of species occurrences data to model evaluation. Such an approach is

essential to increase confidence in model outputs (Porfirio et al. 2014): for most areas of the

world and species, survey effort often exhibits strong spatial and temporal bias, occurrence

records being frequently too scarce, constrained to presence-only data or both. Working with

biased observation datasets may result in under- or over-estimated species distributional ranges

(Araújo and Guisan 2006, Dormann et al. 2007), leading therefore to inaccurate modelled

contemporary distributions, which are inadequate for assessing potential future range shifts or

for defining conservation measures. Similarly, biased pseudo-absence datasets (e.g. multiple

pseudo-absences selected in the same environmental conditions or coinciding with

environmental conditions where the species is observed) may lead to a distorted estimation of

species distributional ranges (e.g. Wisz and Guisan 2009, Lobo and Tognelli 2011). A

modelling framework that includes a preliminary stage devoted to the construction of a

representative calibration dataset – as well as its associated level of uncertainty assessment – is

therefore essential (e.g. Varela et al. 2014).

Here, we developed a framework that encompasses recent advances on the building,

calibration and evaluation of SDMs with the aim of (i) selecting the most relevant

environmental parameters, (ii) generating consistent pseudo-absence data and (iii) validating

representative model outputs (Cornwell et al. 2004, Varela et al. 2014, Leroy et al. 2018).

We applied this framework on seven economically important European Small Pelagic

Fish (SPF) species (Mediterranean horse mackerel *Trachurus mediterraneus*, Atlantic horse

mackerel *Trachurus trachurus*, European pilchard *Sardina pilchardus*, round sardinella

Sardinella aurita, European sprat *Sprattus sprattus*, European Anchovy *Engraulis encrasicolus*

and bogue *Boops boops*). These seven SPF species are widely distributed planktonic feeders

134   known for their central role in marine food webs (Cury 2000, Checkley et al. 2009). Moreover,

135   they are of major economic importance and represent a large part of the Mediterranean and

136   Black Sea commercial landings (more than 50% between 2000 and 2013; FAO 2016).

137   However, while SPFs are ideal candidates for SDMs because of their sensitivity to

138   environmental factors (Perry et al. 2005), their European distribution is far from being

139   exhaustively documented and available records originated from diverse and/or non-

140   standardised monitoring surveys (FAO 2016).

141

142   **2. MATERIAL AND METHODS**

143   2.1. Biological and environmental data

144   *2.1.1. Small pelagic fish occurrence data*

145   Occurrence records (e.g. fisheries independent trawl surveys, discrete research

146   samplings) for the seven SPF species (Mediterranean horse mackerel, Atlantic horse mackerel,

147   European pilchard, round sardinella, European sprat, European Anchovy and Bogue) were

148   compiled from three available public databases: the Ocean Biogeographic Information System

149   Mapper (OBIS, http://www.iobis.org/mapper/), the Global Biodiversity Information Facility

150   (GBIF, https://www.gbif.org/) and Fishbase (http://www.fishbase.org/). When possible, we

151   included observations retrieved from the literature to construct the most up-to-date datasets

152   encompassing their entire distribution range (see Supplementary material 1). Biological data

153   retrieved for our study ranged from 1950 to 2017, recent records (since 1990) prevailing

154   (83.2±6.7 %) over both past (1950-1990; 12.2±8.7 %) and undated observations (4.6±3.6 %).

155   Past or undated records were only considered along the distribution edge when the species

156   presence was confirmed by recent records. This precautionary approach avoided over- or under-

157   predictions of the model due to low quality presence data (Kramer-Schadt et al. 2013). The

158   observation records pre-processing consisted in a data cleaning procedure applied on each

7

159   species dataset to (i) remove unreliable observations (e.g. preserved specimen; Newbold 2010)

160   and false identifications (e.g. taxonomic confusion), (ii) discard duplicate occurrences and (iii)

161   ensure the temporal and locational reliability at the edge of the observed distribution (e.g. data

162   on land, longitudinal and/or latitudinal inversion, historical or undated data). According to the

163   ecology of SPFs – species cannot be observed below 300 m depth (Checkley et al. 2009) –

164   while remaining permissive, a precautionary bathymetry threshold (-1000 m) was applied to

165   remove inconsistent occurrences. Following this pre-processing, we obtained seven clean

166   datasets, with a number of observations ranging from 1314 (for Mediterranean horse mackerel)

167   to 24806 (for European sprat). For the seven SPFs, occurrences were aggregated on a 0.1° x

168   0.1° spatial grid (from 70°N to 70°S and 180°E to 180°W) that corresponds to that of

169   environmental parameters.

170

171   *2.1.2 Environmental data*

172   To calculate the ecological niche (*sensu* Hutchinson, 1957) of each SPF, we collected

173   environmental parameters from different databases (see **Table 1** for details). Environmental

174   parameters values for each spatial grid cell were first calculated for each year and then averaged

175   on the 1990-2017 period. The environmental parameters presented in **Table 1** were retrieved in

176   different spatial resolutions ranging from 0.1° to 0.5°. For modelling purpose, all variables were

177   therefore interpolated to a 0.1° × 0.1° grid using a bilinear interpolation in the geographical

178   domain available for all environmental parameters, ranging from 70°N to 70°S and 180°E to

179   180°W.

180

181   2.2. Description of the models

182   We used two approaches to model the potential environmental suitability (i.e. spatialised

183   index between 0 and 1, defined as a probability of presence based on environmental parameters)

184 of each SPF species over the 1990-2017 period: (i) the Non-Parametric Probabilistic Ecological

185 Niche (NPPEN; Beaugrand et al. 2011) model and (ii) seven modelling algorithms available

186 within the BIOMOD2 package (Thuiller et al. 2016). The NPPEN model is a presence only

187 model based on the Mahalanobis generalised distance (Mahalanobis 1936) and on a modified

188 version of the Multiple Response Permutation Procedure (MRPP; Mielke et al. 1981). The

189 BIOMOD2 framework allows ensemble modelling of species distribution (i.e. an average

190 model of a wide range of algorithms; Thuiller et al. 2009). Here, seven algorithms were

191 considered: (i) Generalised Linear Model (GLM), (ii) Generalised Additive Model (GAM), (iii)

192 Generalised Boosting Model (GBM), (iv) Artificial Neural Network (ANN), (v) Flexible

193 Discriminant Analysis (FDA), (vi) Multiple Adaptive Regression Splines (MARS) and (vii)

194 Random Forest (RF). Because the models used in this study have been already described and

195 discussed in several publications (e.g. Beaugrand et al. 2011, Lenoir et al. 2011, Raybaud et al.

196 2015 for NPPEN, e.g. Thuiller et al. 2009, Albouy et al. 2012, Bellard et al. 2013 for

197 BIOMOD2), we refer the reader to this literature for further information. The algorithms were

198 calibrated using the default parameters in BIOMOD2, optimised for species distribution

199 modelling (details in Thuiller et al. 2016). By including this large range of algorithms within

200 an ensemble model approach, we quantified the uncertainty related to the selection of SDMs

201 (Pearson et al. 2006, Buisson et al. 2010) by calculating the standard deviation (SD) and the

202 coefficient of variation (CV) among SDM outputs.

203

204 2.3. Data preparation and ensemble model selection

205 *2.3.1. Pre-selection of the environmental parameters and assessment of multicollinearity*

206 To model the ecological niche of the seven SPFs, we first constructed the full dataset of

207 environmental parameters based on our knowledge of the ecology of SPFs. A variable selection

208 process (**Figure 1**, step 1) was then applied to identify, at the species level, the most

parsimonious dataset that explained each species distribution. This process follows the

procedure described in Leroy et al. (2014) and Bellard et al. (2016). Because most of the

algorithms (especially regression-based models such as GLM) are sensitive to multicollinearity

– that may distort model estimation (Dormann et al. 2013) – relations among environmental

parameters were assessed by means of the Pearson correlation coefficient, using a threshold $r >$

0.7 to reduce the initial environmental matrix. When two or more environmental parameters

showed correlation values above this threshold, only one variable was retained (details in

Supplementary material 2).

We subsequently assessed the relative importance of each environmental parameter by

sequentially randomising each variable and by calculating the resulting current distribution

(Leroy et al. 2014). The variables that best predicted SPF distribution were sea surface

temperature annual mean (SST), temperature variability (sea surface temperature annual range

or monthly variance, depending on the targeted species), bathymetry and distance to coast (see

Supplementary material 2). In order to avoid model over-parameterisation (that affects model

performance, model transferability and assessment of variable importance), we chose not to

include bathymetry and distance to coast directly in the models, but in a hierarchical filtering

approach (Hattab et al. 2014): for a given geographical cell, environmental conditions were

considered as suitable for a marine species only if a probability of occurrence coincided with a

distance to coast less than 50km or up to a 300m depth for oceanic cells, i.e. outside the 50km

wide coastal area. Concerning environmental predictors, we systematically considered

temperature (mean and variability) in our models. Finally, we tested the relevance of including

sea surface salinity (SSS) and/or primary production (log_PP) as a potential third explanatory

environmental parameter in the models. Each run is detailed in Supplementary material 3.

2.3.2. Environmental filtration and pseudo-absence selection

Because sampling effort is neither homogeneous and nor standardised over marine

regions, occurrence data may not be representative of the whole populations, a requirement to

increase the reliability of SDMs (Lobo and Tognelli 2011). While under-sampling is commonly

observed at the edge of species range (Varela et al. 2014), observation datasets can also be

biased toward regions more comprehensively investigated due to an easy access or a long

tradition of monitoring (Fithian et al. 2015).

To consider the risk of over-sampling, and the ensuing over-representation of

environmental features (Kramer-Schadt et al. 2013), we first homogenised species datasets to

assign the same weight to over- and under-sampled regions (**Figure 1**, step 2). A

multidimensional matrix was designed for each species and each combination of environmental

parameters, a dimension reflecting an environmental factor. Each cell of the homogenised

matrix was considered as an environmental stratum, i.e. a combination of a set of parameters,

with the following resolution: 0.5°C for temperature-related parameters, 0.5 for SSS and 0.5

mol.m^{-2}.s^{-1} (in log) for primary production. In case an environmental stratum contained multiple

occurrences, only one occurrence (i.e. one 0.1° x 0.1° geographical cell with the corresponding

environmental conditions) was kept in the homogenised dataset.

We also considered the lack of absence data. To assess this gap, we generated pseudo-

absences using the convex hull method (Cornwell et al. 2004, Getz and Wilmers 2006). The

convex hull was defined here as the smallest convex hyper-volume in the environmental space

containing all species observation records. A restricted convex hull (see **Figure 2**) has been

defined as a convex hull excluding occurrence points within the 2.5% and 97.5% percentiles

for each environmental parameter (*i.e.* excluding observations in the most extreme

environmental conditions). This restricted convex hull is considered as a proxy of the suitable

environmental conditions outside which, pseudo-absences were randomly generated in equal

259 number to the filtered occurrences as recommended by the "D-designs" theory (Montgomery

260 2005): the optimal design to minimise prediction variance is when an equal number of

261 observations are at opposite value extremes (Montgomery 2005, Hengl et al. 2009) and when

262 there is a high spreading in the feature space. Finally, for each species, pseudo-absence were

263 projected back in geographical cells showing environmental conditions outside SPF species'

264 environmentally favourable areas (**Figure 2**; Varela et al. 2014). Finally, model outputs

265 obtained from our environmental filtration approach were compared with outputs for which

266 neither environmental filtration nor the convex hull pseudo-absence selection method was

267 applied (**Figure 3**).

268

269 *2.3.3. Validation and selection of the best models*

270 We then quantified the performance of our models using five commonly used evaluation

271 metrics: (i) the Continuous Boyce Index (CBI; Hirzel et al. 2006), a metric specifically designed

272 for presence-only models and insensitive to pseudo-absences, (ii) the Area Under the Curve

273 (AUC; Swets 1988, Fielding and Bell 1997), (iii) the True Skill Statistic (TSS; Allouche et al.

274 2006), (iv) the Jaccard and (v) the Sørensen similarity indices (Jaccard 1908, Sørensen 1948).

275 However, because all evaluation metrics – except the CBI – require both presence and absence

276 data (see discussion in Leroy et al. 2018 about the use of pseudo-absence to evaluate the

277 performance of models) and because some may be affected by prevalence (i.e. the ratio between

278 the number of observed presence and generated pseudo-absence; Leroy et al. 2018) we based

279 our selection process of the best models on CBI values only. We considered models to be wrong

280 when CBI values were below -0.5, "average to random" for values ranging from -0.5 to 0.5,

281 and good for values above 0.5 (Faillettaz et al. 2019).

282 For each model, we computed evaluation metrics by performing a cross-validation

283 procedure with 10 repetitions. We randomly sampled 70% of the occurrence data to calibrate

284   the model and kept the remaining 30% for model validation (Merow et al. 2013). Following the

285   "evaluation strip method" detailed by Elith et al. (2005), the adequacy between observed and

286   modelled spatial distributions was also assessed by means of response curves. For a given

287   environmental parameter, the corresponding response curve was calculated, while keeping the

288   other parameters constant (i.e. at the mean value corresponding to their occurrence points). By

289   doing this, we identified spurious results (e.g. we do not expect bimodal responses to

290   temperature) and/or unexpected distribution ranges (e.g. large portions of predicted range in

291   regions where the species has never been observed, and vice-versa; Supplementary material 4).

292

293

294   **3. RESULTS**

295   <u>3.1. SDMs and parameters selected in the ensemble models</u>

296   Based on the calculation of the CBI values and the examination of species response

297   curves (Supplementary material 3 and 4), we identified the best models for each SPF species.

298   Our results showed that both GLM and NPPEN models were almost always selected in the

299   ensemble model, except for the European anchovy.

300   Ensemble models showed that temperature-related variables were essential to assess the

301   spatial distribution of SPFs'. For virtually all species, the models that considered mean

302   temperature and variability showed high ability to reproduce the overall SPFs distributions

303   (**Table 2**, Supplementary material 3) with CBI values always above 0.5 (Faillettaz et al. 2019).

304   However, some discrepancies were observed among species. While Mediterranean horse

305   mackerel, Atlantic horse mackerel and European anchovy distributions were more related to

306   mean monthly temperature variance (SSTvar), European pilchard, round sardinella, European

307   Sprat and bogue distributions were better reproduced when mean annual temperature range

308   (SSTr) was considered. Despite the high correlation between SSTr and SSTvar (r=0.80,

309　Supplementary material 2), both variables have dissimilar ecological influences (seasonality

310　versus short-term climatic variability respectively). Primary production also emerged as

311　important to model species' spatial distribution. Finally, we highlighted the important role of

312　sea surface salinity (SSS) for both European pilchard and European anchovy, by discriminating

313　both the Baltic and the Black seas from other regions (**Table 2**).

314

315　　　　By applying our environmental filtration framework, we substantially improved the

316　modelling of most of the SPFs spatial distributions (**Figure 3,** individual contributions of the

317　filtration process and the convexhull are presented in Supplementary material 5), except for the

318　European pilchard (**Figure 3b**). Specifically, we observed an increase in mean CBI values that

319　ranged from +0.05 to +0.23 (**Figure 3**). For most SPFs, lower Environmental Suitability Index

320　(ESI) values were obtained (-0.2 without filtration to -0.6 with filtration), suggesting that our

321　procedure alleviated the risk of over-prediction, especially in the Black and Baltic seas, and

322　beyond 60°N where species have never been observed (**Figure 4**, left panels). By increasing

323　ESI values from +0.4 to +0.6, environmental filtration also emphasised regions known to be

324　highly suitable for SPF species, but in which occurrences were only scarcely available (e.g. in

325　the eastern Mediterranean Sea for Atlantic horse mackerel, round sardinella and bogue; **Figure**

326　**4a, f and g**).

327　3.2. Contemporary (1990-2017) environmental suitability of small pelagic fishes

328　　　　We then represented the contemporary (1990-2017) spatial distribution of the seven

329　SPFs in the spatial domain ranging from 10 to 70°N and from 30°W to 45°E (**Figure 4**, middle

330　panel) Environmental suitabilities at the calibration range (i.e. the entire distribution range) are

331　provided in Supplementary material 6.

332

14

333   According to the observed and modelled distributions (**Figure 4**, left and middle

334   panels), two species groups were identified with respect to their environmental suitability along

335   the European coasts. The first group encompassed temperate-to-cold water species (hereafter

336   "temperate-cold" species; i.e. Atlantic horse mackerel, European pilchard, European sprat and

337   European anchovy; **Figure 4a-d**) that were more likely to be present in northern Europe. The

338   second grouped temperate-to-warm water species (hereafter "temperate-warm" species; i.e.

339   Mediterranean horse mackerel, round sardinella and bogue; **Figure 4e-g**) located along the

340   Mediterranean coasts down, to North Africa.

341   The four temperate-cold species showed the highest ESI values in the North Sea, in the

342   Celtic Sea, in the Bay of Biscay (ESI values > 0.8) and to a lesser extent along Norwegian

343   coasts (ESI values ranging from 0.2 to 0.8). For all temperate-cold species, but European

344   pilchard, high ESI values (from 0.4 to 0.8) were expected in the western and central regions of

345   the Baltic Sea (**Figure 4**), suggesting that these species can tolerate a wide salinity range (from

346   8 to 38) and a high thermal variability (up to 20°C annual range). All temperate-cold species,

347   but European sprat, showed high ESI values (from 0.6 to 0.8) in the north-western part of the

348   Mediterranean basin (**Figure 4**). For all temperate-cold species, the modelled ESIs are in

349   accordance with the observation data except in southern Iceland, western Norway and to a lesser

350   extent in the eastern Black Sea where positive ESI values (between 0.05 to 0.6) are predicted

351   while no observed distribution is available.

352   The three temperate-warm species showed the highest ESI values (from 0.4 to 0.8) in

353   nearly all the regions of the Mediterranean Sea and medium to low ESI values (from 0.2 to 0.7)

354   in the Black Sea and along the north-western African coasts. However, some discrepancies

355   among species were detected (**Figure 4**). Round sardinella appears as the most southern SPF

356   species with no suitable conditions north of the Portuguese coast. On the contrary,

357   Mediterranean horse mackerel and bogue showed high ESI values (from 0.6 to 0.8) along the

15

358 Atlantic coasts from the Celtic sea down to northern Africa, up to 0.8 in the Bay of Biscay.

359 While bogue showed maximum ESI values (> 0.8) in the whole Mediterranean Sea, only the

360 north-western regions of the Mediterranean Sea were highly suitable for Mediterranean horse

361 mackerel and round sardinella. The modelled ESIs are in accordance with the observation data

362 except in the North Sea for Mediterranean horse mackerel and Bogue and to a lesser extent in

363 the eastern Black Sea for all temperate-warm species. These regions highlight positive ESI

364 values (between 0.05 and 0.6) while no observations are available. These discrepancies may

365 result from an absence of sampling in these regions or external factors hindering species

366 establishment despite suitable environmental conditions.

367

368 3.3. Model uncertainties

369   Two main sources of uncertainties in projected species distributions were considered in

370 our study: (i) biological uncertainties, related to the quality of occurrence datasets and (ii)

371 numerical uncertainties, inherent to the selection of different modelling algorithms (Pearson et

372 al. 2006, Buisson et al. 2010). Standard deviations (SD) – computed, for each species, from

373 outputs that originated from both selected algorithms and cross-validation runs – ranged from

374 0.1 to 0.4, indicating a convergence between models (**Figure 4**, right panels). The lowest SD

375 values (close to 0.2) were found in the north-western Mediterranean Sea for virtually all SPFs,

376 and in the Bay of Biscay and in the North Sea when temperate-cold species were studied

377 (**Figure 4, a-d**). The highest SD values (close to 0.4) were observed in the Mediterranean Sea

378 for Mediterranean horse mackerel, European pilchard and round sardinella (**Figure 4, e-g**). For

379 all species, the coefficient of variation (CV; Supplementary material 7) highlighted very low

380 CV variations (< 20%) towards their centre of distribution (in the Mediterranean Sea for all

381 species and North Sea for temperate-cold species) while showing high variations at the leading

382   or the trailing edge of their distribution (up to 100% in the Black, Baltic and the Norwegian

383   seas).

384

385   **4. DISCUSSION**

386   By combining several numerical techniques such as the convex hull method, the

387   ensemble models approach and an examination of species response curves in a comprehensive

388   modelling framework, we modelled the contemporary (1990-2017) environmental suitability

389   of seven of the most commercially and ecologically important European small pelagic fish. By

390   relying on both an understanding of the ecological requirements of species and on the use of

391   innovative statistical tools, our framework allowed us to focus only on the best models, to

392   improve the way species distribution modelling is carried out, and therefore to produce more

393   robust ecological scenarios.

394

395   At a macroecological level, thermal-induced effects have been frequently related to

396   latitudinal mean temperature gradients (Angilletta 2011). While our analysis showed that mean

397   temperature (SST) had a major influence on species distributions, we also revealed the key role

398   of temperature seasonality (SSTr) and short-term temperature variations (SSTvar) in shaping

399   distributional ranges (**Table 2**). Small pelagic fishes are marine ectotherms, that mainly depend

400   on external heat sources, their body temperature being directly controlled by environmental

401   conditions directly (Checkley et al. 2009). Changes in temperature may therefore affect SPFs'

402   physiological performances (i.e. their fitness; Perry et al. 2005, Payne et al. 2016). Because the

403   relationship between temperature and fitness occurred through species' thermal optimum and

404   range, and because SPFs are short lifespan species (Checkley et al. 2009), annual temperature

405   changes may affect several life stages (especially reproduction and larval development; e.g.

406   Peck et al. 2013), with long-term consequences on population dynamics (Fréon et al. 2005).

407 Small pelagic fishes may also experience ontogenetic shifts in thermal tolerance during their

408 development (Peck et al. 2013) and temperature seasonality (here SSTr) may either favour or

409 perturb species development, with potential consequences on distributional patterns (Figure 4,

410 middle panels; Peck et al. 2013). This is especially noticeable in regions characterised by an

411 important thermal variability, such as in the Black and Azov seas, in the Northern Adriatic Sea,

412 in the Baltic Sea and to a lesser extent in the eastern part of the North Sea. Considering thermal

413 variability in SDMs (e.g. the monthly SST variance) may therefore help to better define species

414 environmental suitability and to minimise the risk of over-prediction at the leading and the

415 trailing edges of their distributions (Lenoir et al. 2011).

416

417   When used in distribution modelling, regression-based algorithms such as GLM, are

418 known to be rather sensitive to environmental sampling bias, which may induce type I errors

419 (i.e. false positive), with consequences on projected species environmental suitability (Araújo

420 and Guisan 2006, Dormann et al. 2007). However, as for many other species (e.g. Boakes et al.

421 2010), commonly available databases of SPFs provide a distorted view of their actual

422 distribution because of spatial and temporal bias in species observations (e.g. Beck et al. 2014).

423 When the time comes to evaluate the quality of biodiversity datasets, three major issues have

424 been raised in the literature (e.g. Kramer-Schadt et al. 2013, Guillera-Arroita et al. 2015): the

425 influence of (i) prevalence, i.e. the proportion of sites in which the species was recorded as

426 present, (ii) imperfect species detection and (iii) sampling bias. Despite an increasing

427 availability of information, the biogeographic distribution of most species remain still

428 frequently incomplete (Bini et al. 2006); a shortcoming explained, inter alia, by heterogeneous

429 sampling effort among surveys, or the inaccessibility of some areas. For all SPF datasets, this

430 effect is undeniable when comparing the north-western Mediterranean Sea, the Bay of Biscay,

431 the North Sea with other European regions. (**Figure 4**, left panels). To lower this issue, a

432   plethora of data sources (e.g. standardised scientific surveys, biodiversity portals) are now

433   available in collaborative databases (e.g. GBIF), offering more cohesive summaries of species'

434   distributions although leading – sometimes – to enhanced spatial and environmental biases

435   (Kramer-Schadt et al. 2013, Beck et al. 2014). Considering independent distributional data (i.e.

436   from private collections or from the literature; Beck et al. 2013) along with the associated pre-

437   processing (e.g. Kramer-Schadt et al. 2013, Varela et al. 2014, Aiello-Lammens et al. 2015,

438   Fithian et al. 2015), can contribute to cover the ecological niches of species more

439   comprehensively and to improve model accuracy. By coupling these procedures with our

440   restricting convex hull pseudo-absence selection, we (i) assigned the same weight to

441   environmental conditions independently of the observation density (i.e. alleviating observation

442   sampling bias), (ii) lowered the weight of presence records at the distribution edge (i.e. avoiding

443   the risk of over-prediction) and (iii) selected unbiased pseudo-absence (i.e. independent of the

444   observation bias).

445

446   Applying environmental filtering and the restricted convexhull pseudo-absence

447   selection method resulted in ensemble models characterised by a reduced ESI in over-sampled

448   areas and an increased ESI in undersampled areas. Our results are consistent with our

449   expectations and in line with previous studies that suggested that random generation of pseudo-

450   absences and/or a selection process based on geographical criterion may lead to lower

451   predictability (e.g. Wisz and Guisan 2009, Hattab et al. 2014). Although real absences lead to

452   higher model accuracy (Wisz and Guisan 2009), they are rarely available (Boakes et al. 2010)

453   and determining the location of pseudo-absences on the basis of a statistical analysis such as

454   the convex hull is a reliable alternative (Hattab et al. 2013). Finally, our approach limits

455   spurious species response curves (i.e. overfitted or bimodal curves; Supplementary material 4)

456   and decreases the risk of over-predictions towards the edge of the species range. We

457   acknowledge that we may have slightly underpredicted the European pilchard distribution in

458   Kattegat (i.e. strait between Denmark and Sweden); the high amount of occurrence records

459   slightly outside the modelled distribution in this region may have biased the calculation of the

460   CBI. Despite the well-known robustness of this index (Breiner et al. 2015, Faillettaz et al. 2019),

461   our result highlight that no evaluation metric is optimal and that both comparison between

462   observed and modelled distributions and examination of species responses curves are essential

463   for assessing the reliability of model outputs.

464

465   While the assessment of the environmental suitability for a given species may differ –

466   slightly or markedly – from one SDM to another (Pearson et al. 2006, Buisson et al. 2010), it is

467   still challenging to identify the most appropriate model (see discussion in Buisson et al. 2010).

468   Even if several methods have been recently proposed, no consensus has emerged (see

469   discussion in Leroy et al. 2018). and the use of different – well-fitted and evaluated – SDMs

470   may help to better simulate potential species distributions, for past, contemporary and future

471   environmental conditions (Araújo and New 2007). In complementarity with a multi-SDM

472   approach, we think that researchers should examine species response curves during the

473   evaluation process (e.g. Elith and Leathwick 2009, Jarnevich et al. 2018, Erauskin-Extramiana

474   et al. 2019). As observed for Mediterranean horse mackerel (see details in Supplementary

475   material 4), we invalidated response curves that were statistically significant but not in

476   agreement with the ecological niche theory. Without this complementary evaluation method,

477   the corresponding algorithms would have been considered in the ensemble model, therefore

478   potentially resulting in spurious patterns of ESIs. Therefore, this multi-criteria evaluation

479   procedure is of great interest from a (i) numerical (i.e. metric adapted to presence-only datasets)

480   and an ecological (i.e. validation of the species-environment relationships) perspective. Note

481   that the seven SPFs we chose are representative of a large spectrum of environmental

482 conditions, from temperate-to-cold waters (e.g. European sprat) to temperate-to-warm waters

483 (e.g. bogue and round sardinella).To conclude, our framework has been faced with a wide range

484 of environmental conditions, allowing us to better evaluate its robustness, sensitivity and

485 possible transferability to other species and ecosystems.

486

487   In this work, we have estimated species' potential niche and not the realised niche

488 (Soberón and Nakamura 2009). We caution that additional environmental parameters,

489 biological interactions and species life traits (e.g. dispersal abilities) may explain why we

490 detected environmentally suitable conditions in regions where SPFs were not observed (e.g. the

491 Norwegian Sea; Pulliam 2000, Pearman et al. 2008). Considering the role of biotic interactions

492 in shaping species distributions (Chaalali et al. 2016) would improve the reliability of SDMs

493 outputs by better estimating and simulating the realised niche of species (Wisz et al. 2013,

494 Louthan et al. 2015). Including dispersal mechanisms while accounting for oceanic currents

495 and physical barriers after the potential distribution modelling step may help to refine the

496 distributional range of species (Engler and Guisan 2009). These approaches require an

497 exhaustive ecological understanding of the interaction process at a macroecological scale and a

498 deep knowledge of species life traits to implement metrics that simulate the ability of species

499 to disperse (e.g. Petitgas et al. 2012). Moreover, it is important to notice that no direct

500 correlations between ESI (potential or realised) and spatialised biomass or official catches have

501 been established in the literature although temporal correlations have been suggested however

502 (e.g. Chaalali et al. 2016). Therefore, discrepancies between SPF's ESI, biomass and official

503 catches (e.g. FAO 2016) may be explained by population-related parameters (e.g. recruitment,

504 growth, biotic interaction) or management policies and stock status (e.g. under or over-fishing),

505 respectively. Finally, inter-specific absolute ESI comparison is challenging because of the

506 monospecific nature of SDMs.

507

508   Our study presents a detailed environmental suitability assessment of seven of the most

509 heavily harvested European SPFs. By focusing on the most common sources of errors and

510 uncertainties in SDMs, we designed a comprehensive - fully transferable to other species and

511 ecosystems - modelling framework which is intended to elaborate more robust ecological

512 scenarios. Our framework addressed several critical steps in SDMs, i.e. the treatment of

513 sampling biases in observation records, the generation relevant pseudo-absences and a dual

514 assessment of model outputs that proposes to evaluate models from both a numerical and an

515 ecological perspective. In a conservation decision-making perspective, these different steps are

516 essential to increase confidence in SDMs, a prerequisite to propose effective resource

517 management measures (e.g. accounting for environmental stress) or to measure the

518 effectiveness of protected areas (e.g. regarding environmental resilience). Moreover, when used

519 in combination with scenarios of future environmental conditions (i.e. IPCC climate scenarios),

520 this framework provides robust contemporary predictions to assess possible changes in species

521 distribution in the context of global climate change. Despite the growing literature on the

522 development and testing of new modelling and evaluation processes, the use of SDMs in

523 quantitative resource management and scientific surveys is still a great challenge.

524

525 **DECLARATIONS**

532   Intercomparison Project (CMIP) and the climate modelling groups for making available their

533   model output.

534

539

540   *Author contribution* - VR and PF conceived and supervised the study. AS, VR and EG collected

541   the data. AS performed the numerical analysis. BL, GB, TH, EG and VR helped in the

542   modelling process. AS and EG wrote the first draft. BL provided the code to use BIOMOD2.

543   All authors made substantial contributions in the successive versions of the manuscript.

544

545   *Conflicts of interest* – All authors have no conflict of interest to declare

546

547   **References**
548   Aiello-Lammens, M. E. et al. 2015. spThin: an R package for spatial thinning of species
549   occurrence records for use in ecological niche models. - Ecography 38: 541–545.

550   Albouy, C. et al. 2012. Combining projected changes in species richness and composition
551   reveals climate change impacts on coastal Mediterranean fish assemblages. - Global
552   Change Biology 18: 2995–3003.

553   Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence,
554   kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models.
555   - Journal of Applied Ecology 43: 1223–1232.

556   Angilletta, M. J. 2011. Thermal adaptation: a theoretical and empirical synthesis. - Oxford
557   University Press.

558   Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling.
559   - Journal of Biogeography 33: 1677–1688.

560   Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. - Trends in
561   Ecology & Evolution 22: 42–47.

562 Beaugrand, G. and Kirby, R. R. 2018. How Do Marine Pelagic Species Respond to Climate
563   Change? Theories and Observations. - Annual Review of Marine Science 10: 169–
564   197.

565 Beaugrand, G. et al. 2011. A new model to assess the probability of occurrence of a species,
566   based on presence-only data. - Marine Ecology Progress Series 424: 175–190.

567 Beaugrand, G. et al. 2013. Applying the concept of the ecological niche and a
568   macroecological approach to understand how climate influences zooplankton:
569   Advantages, assumptions, limitations and requirements. - Progress in Oceanography
570   111: 75–90.

571 Beaugrand, G. et al. 2015. Future vulnerability of marine biodiversity compared with
572   contemporary and past changes. - Nature Climate Change 5: 695–701.

573 Beaugrand, G. et al. 2018. Marine biodiversity and the chessboard of life. - PLOS ONE 13:
574   e0194006.

575 Beaumont, L. J. et al. 2008. Why is the choice of future climate scenarios for species
576   distribution modelling important? - Ecology Letters 11: 1135–1146.

577 Beck, J. et al. 2013. Online solutions and the 'Wallacean shortfall': what does GBIF
578   contribute to our knowledge of species' ranges? - Diversity and Distributions 19:
579   1043–1050.

580 Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species'
581   geographic distributions. - Ecological Informatics 19: 10–15.

582 Bellard, C. et al. 2013. Will climate change promote future invasions? - Global Change
583   Biology 19: 3740–3748.

584 Bellard, C. et al. 2016. Major drivers of invasion risks throughout the world. - Ecosphere 7:
585   e01241.

586 Ben Rais Lasram, F. et al. 2010. The Mediterranean Sea as a 'cul-de-sac' for endemic fishes
587   facing climate change: A marine endemic hotspot under threat. - Global Change
588   Biology 16: 3233–3245.

589 Bini, L. M. et al. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients
590   and conservation planning in a biodiversity hotspot. - Diversity and Distributions 12:
591   475–482.

592 Boakes, E. H. et al. 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in
593   Species Occurrence Data. - PLOS Biology 8: e1000385.

594 Breiner, F. T. et al. 2015. Overcoming limitations of modelling rare species by using
595   ensembles of small models (B Anderson, Ed.). - Methods in Ecology and Evolution 6:
596   1210–1218.

597 Buisson, L. et al. 2010. Uncertainty in ensemble forecasting of species distribution. - Global
598   Change Biology 16: 1145–1157.

599 Casey, K. S. et al. 2010. The Past, Present, and Future of the AVHRR Pathfinder SST
600   Program. - In: Oceanography from Space. Springer, Dordrecht, pp. 273–287.

601 Chaalali, A. et al. 2016. From species distributions to ecosystem structure and function: A
602   methodological perspective. - Ecological Modelling 334: 78–90.

603 Checkley, D. et al. 2009. Climate change and small pelagic fish. - Cambridge University
604   Press.

605 Cheung, W. W. L. et al. 2009. Projecting global marine biodiversity impacts under climate
606   change scenarios. - Fish and Fisheries 10: 235–251.

607 Colwell, R. K. and Rangel, T. F. 2009. Hutchinson's duality: The once and future niche. -
608   PNAS 106: 19651–19658.

609 Cornwell, W. K. et al. 2004. A Trait-Based Test for Habitat Filtering: Convex Hull Volume. -
610   Ecology 87: 1465–1471.

611 Crisp, M. D. et al. 2009. Phylogenetic biome conservatism on a global scale. - Nature 458:
612   754–756.

613 Cristofari, R. et al. 2018. Climate-driven range shifts of the king penguin in a fragmented
614   ecosystem. - Nature Climate Change 8: 245–251.

615 Cury, P. 2000. Small pelagics in upwelling systems: patterns of interaction and structural
616   changes in "wasp-waist" ecosystems. - ICES Journal of Marine Science 57: 603–618.

617 Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of
618   species distributional data: a review. - Ecography 30: 609–628.

619 Dormann, C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation
620   study evaluating their performance. - Ecography 36: 27–46.

621 Dufresne, J.-L. et al. 2013. Climate change projections using the IPSL-CM5 Earth System
622   Model: from CMIP3 to CMIP5. - Climate Dynamics 40: 2123–2165.

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

626 Elith, J. et al. 2005. The evaluation strip: A new and robust method for plotting predicted
627   responses from species distribution models. - Ecological Modelling 186: 280–289.

628 Engler, R. and Guisan, A. 2009. MigClim: Predicting plant distribution and dispersal in a
629   changing climate. - Diversity and Distributions 15: 590–601.

630 Erauskin-Extramiana, M. et al. 2019. Historical trends and future distribution of anchovy
631   spawning in the Bay of Biscay. - Deep Sea Research Part II: Topical Studies in
632   Oceanography 159: 169–182.

633 Faillettaz, R. et al. 2019. Atlantic Multidecadal Oscillations drive the basin-scale distribution
634   of Atlantic bluefin tuna. - Sci Adv 5: eaar6993.

635 FAO 2016. The State of Mediterranean and Black Sea Fisheries. - General Fisheries
636   Commission for the Mediterranean.

637 Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction
638   errors in conservation presence/absence models. - Environmental Conservation 24:
639   38–49.

640 Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and
641   collection data for multiple species. - Methods in Ecology and Evolution 6: 424–438.

642 Fréon, P. et al. 2005. Sustainable exploitation of small pelagic fish stocks challenged by
643   environmental and ecosystem changes: A review. - Bulletin of Marine Science 76: 79.

644 Getz, W. M. and Wilmers, C. C. 2006. A local nearest-neighbor convex-hull construction of
645   home ranges and utilization distributions. - Ecography 27: 489–505.

646 Giorgetta, M. A. et al. 2013. Climate and carbon cycle changes from 1850 to 2100 in MPI-
647   ESM simulations for the Coupled Model Intercomparison Project phase 5: Climate
648   Changes in MPI-ESM. - Journal of Advances in Modeling Earth Systems 5: 572–597.

649 Gleason, H. A. 1926. The Individualistic Concept of the Plant Association. - Bulletin of the
650   Torrey Botanical Club 53: 7–26.

651 Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching
652   data and models to applications. - Global Ecology and Biogeography 24: 276–292.

653 Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple
654   habitat models. - Ecology Letters 8: 993–1009.

655 Hattab, T. et al. 2013. The Use of a Predictive Habitat Model and a Fuzzy Logic Approach for
656   Marine Management and Planning. - PLOS ONE 8: e76430.

657 Hattab, T. et al. 2014. Towards a better understanding of potential impacts of climate change
658   on marine species distribution: a multiscale modelling approach. - Global Ecology and
659   Biogeography 23: 1417–1429.

660 Helaouet, P. and Beaugrand, G. 2009. Physiology, Ecological Niches and Species
661   Distribution. - Ecosystems 12: 1235–1245.

662 Hengl, T. et al. 2009. Spatial prediction of species' distributions from occurrence-only
663   records: combining point pattern analysis, ENFA and regression-kriging. - Ecological
664   Modelling 220: 3499–3511.

665 Hirzel, A. H. et al. 2006. Evaluating the ability of habitat suitability models to predict species
666   presences. - Ecological Modelling 199: 142–152.

667 Hourdin, F. et al. 2013. Impact of the LMDZ atmospheric grid configuration on the climate
668   and sensitivity of the IPSL-CM5A coupled model. - Climate Dynamics 40: 2167–
669   2192.

670 Hutchinson, G. E. 1957. Concluding Remarks. - Cold Spring Harb Symp Quant Biol 22: 415–
671   427.

672   Hutchinson, G. E. 1978. An introduction to population ecology. - New Haven : Yale
673   University Press.

674   Jaccard, P. 1908. Nouvelles Recherches Sur la Distribution Florale. - Bulletin de la Societe
675   Vaudoise des Sciences Naturelles 44: 223–70.

676   Jarnevich, C. S. et al. 2018. Forecasting an invasive species' distribution with global
677   distribution data, local data, and physiological information. - Ecosphere 9: e02279.

678   Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concepts in
679   species distribution modelling. - Diversity and Distributions 14: 885–890.

680   Jones, C. D. et al. 2011. The HadGEM2-ES implementation of CMIP5 centennial simulations.
681   - Geoscientific Model Development Discussions 4: 689–763.

682   Kirby, R. R. and Beaugrand, G. 2009. Trophic amplification of climate warming. -
683   Proceedings of the Royal Society B: Biological Sciences 276: 4095–4103.

684   Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt
685   species distribution models. - Diversity and Distributions 19: 1366–1379.

686   Lenoir, S. et al. 2011. Modelled spatial distribution of marine fish and projected modifications
687   in the North Atlantic Ocean. - Global Change Biology 17: 115–129.

688   Leroy, B. et al. 2014. Forecasted climate and land use changes, and protected areas: the
689   contrasting case of spiders. - Diversity and Distributions 20: 686–697.

690   Leroy, B. et al. 2018. Without quality presence–absence data, discrimination metrics such as
691   TSS can be misleading measures of model performance. - Journal of Biogeography
692   45: 1994–2002.

693   Levitus, S. 2011. Climatological Atlas of the World Ocean. - Eos, Transactions American
694   Geophysical Union 64: 962–963.

695   Lobo, J. M. and Tognelli, M. F. 2011. Exploring the effects of quantity and location of
696   pseudo-absences and sampling biases on the performance of distribution models with
697   limited point occurrence data. - Journal for Nature Conservation 1: 1–7.

698   Louthan, A. M. et al. 2015. Where and When do Species Interactions Set Range Limits? -
699   Trends in Ecology & Evolution 30: 780–792.

700   Mahalanobis, P. 1936. On the generalised distance in statistics. - Proceedings of the National
701   Institute of Science, India 2: 49–55.

702   Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what
703   it does, and why inputs and settings matter. - Ecography 36: 1058–1069.

704   Mielke, P. W. et al. 1981. Application of Multi-Response Permutation Procedures for
705   Examining Seasonal Changes in Monthly Mean Sea-Level Pressure Patterns. - Mon.
706   Wea. Rev. 109: 120–126.

707   Montgomery, D. C. 2005. Design and analysis of experiments. - Wiley.

708 Newbold, T. 2010. Applications and limitations of museum data for conservation and
709   ecology, with particular attention to species distribution models. - Progress in Physical
710   Geography: Earth and Environment 34: 3–22.

711 Payne, N. L. et al. 2016. Temperature dependence of fish performance in the wild: links with
712   species biogeography and physiological thermal tolerance. - Functional Ecology 30:
713   903–912.

714 Pearman, P. B. et al. 2008. Niche dynamics in space and time. - Trends in Ecology &
715   Evolution 23: 149–158.

716 Pearson, R. G. et al. 2006. Model-based uncertainty in species range prediction. - Journal of
717   Biogeography 33: 1704–1711.

718 Peck, M. A. et al. 2013. Life cycle ecophysiology of small pelagic fish and climate-driven
719   changes in populations. - Progress in Oceanography 116: 220–245.

720 Perry, A. L. et al. 2005. Climate change and distribution shifts in marine fishes. - Science 308:
721   1912–1915.

722 Petitgas, P. et al. 2012. Anchovy population expansion in the North Sea. - Marine Ecology
723   Progress Series 444: 1–13.

724 Poloczanska, E. S. et al. 2013. Global imprint of climate change on marine life. - Nature
725   Climate Change 3: 919–925.

726 Porfirio, L. L. et al. 2014. Improving the Use of Species Distribution Models in Conservation
727   Planning and Management under Climate Change (L Kumar, Ed.). - PLoS ONE 9:
728   e113749.

729 Proosdij, A. S. J. van et al. 2016. Minimum required number of specimen records to develop
730   accurate species distribution models. - Ecography 39: 542–552.

731 Pulliam, H. R. 2000. On the relationship between niche and distribution. - Ecology Letters 3:
732   349–361.

733 Raybaud, V. et al. 2015. Climate-induced range shifts of the American jackknife clam Ensis
734   directus in Europe. - Biological Invasions 17: 725–741.

735 Raybaud, V. et al. 2017. Forecasting climate-driven changes in the geographical range of the
736   European anchovy (Engraulis encrasicolus). - ICES Journal of Marine Science 74:
737   1288–1299.

738 Schmidt, G. A. et al. 2014. Configuration and assessment of the GISS ModelE2 contributions
739   to the CMIP5 archive: GISS MODEL-E2 CMIP5 SIMULATIONS. - Journal of
740   Advances in Modeling Earth Systems 6: 141–184.

741 Smith, W. H. F. and Sandwell, D. T. 1997. Global Sea Floor Topography from Satellite
742   Altimetry and Ship Depth Soundings. - Science 277: 1956–1962.

743 Soberon, J. and Peterson, A. T. 2005. Interpretation of Models of Fundamental Ecological
744   Niches and Species' Distributional Areas. - Biodiversity Informatics in press.

745   Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: Concepts, methods, and
746   assumptions. - Proceedings of the National Academy of Sciences of the United States
747   of America 106 Suppl 2: 19644–50.

748   Sørensen, T. 1948. {A method of establishing groups of equal amplitude in plant sociology
749   based on similarity of species and its application to analyses of the vegetation on
750   Danish commons}. - Biol. Skr. 5: 1–34.

751   Stevens, B. et al. 2013. Atmospheric component of the MPI-M Earth System Model:
752   ECHAM6. - Journal of Advances in Modeling Earth Systems 5: 146–172.

753   Støa, B. et al. 2018. Sampling bias in presence-only data used for species distribution
754   modelling: theory and methods for detecting sample bias and its effects on models. -
755   Sommerfeltia 38: 1–53.

756   Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. - Science 240: 1285–1293.

757   Thuiller, W. et al. 2009. BIOMOD - a platform for ensemble forecasting of species
758   distributions. - Ecography 32: 369–373.

759   Thuiller, W. et al. 2016. Ensemble Platform for Species Distribution Modelling. in press.

760   Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve
761   predictions of ecological niche models. - Ecography 37: 1084–1091.

762   Voldoire, A. et al. 2013. The CNRM-CM5.1 global climate model: description and basic
763   evaluation. - Climate Dynamics 40: 2091–2121.

764   Wiens, J. A. et al. 2009. Niches, models, and climate change: Assessing the assumptions and
765   uncertainties. - PNAS 106: 19729–19736.

766   Wisz, M. S. and Guisan, A. 2009. Do pseudo-absence selection strategies influence species
767   distribution models and their predictions? An information-theoretic approach based on
768   simulated data. - BMC Ecology 9: 8.

769   Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised
770   assemblages of species: implications for species distribution modelling. - Biological
771   Reviews 88: 15–30.

772
773

774   **TABLES AND FIGURES**

775

Name (Period)	Description	Reference
Bathymetry	Spatial seafloor depth (m)	Global seafloor topography (Smith and Sandwell 1997)
Distance to coast	Distance to the nearest coast (km)	NASA Goddard Space Flight Center (2009) (https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/)
SSS (1990-2017)	Sea Surface Salinity	Levitus' climatology (Levitus 2011) completed with (http://www.ices.dk/)
PP (1990-2017)	Sea Surface Primary Production (mol.m^{-2}.s^{-1}). Averaged from five general circulation models (IPSL, MPI, CNRM, HadGEM and GISS).	IPSL (Dufresne et al. 2013, Hourdin et al. 2013), MPI (Stevens et al. 2013, Giorgetta et al. 2013), CNRM (Voldoire et al. 2013), HadGEM (Jones et al. 2011) and GISS (Schmidt et al. 2014) models.
Log_PP (1990-2017)	Log10-transformed Sea Surface Primary Production	
SST (1990-2017)	Mean annual Sea Surface Temperature (°C)	AVHRR Very High Resolution Radiometer (Casey et al. 2010)
SSTmax (1990-2017)	Mean sea surface temperature of the hottest month (°C)	
SSTmin (1990-2017)	Mean sea surface temperature of the coldest month (°C)	
SSTr (1990-2017)	Mean annual sea surface temperature range (°C). Difference between SSTmax and SSTmin.	
SSTvar (1990-2017)	Mean monthly sea surface temperature variance (°C). Calculated using monthly SST data.	

776

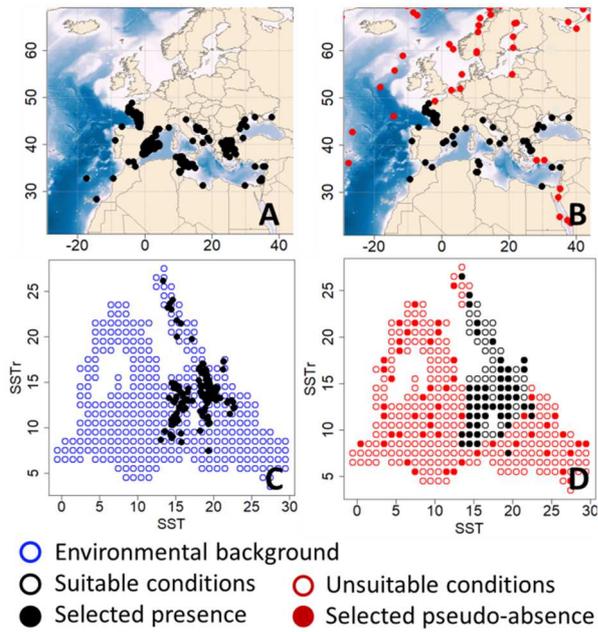777   **Table 1:** Environmental parameters used to model SPF distribution.

778

779
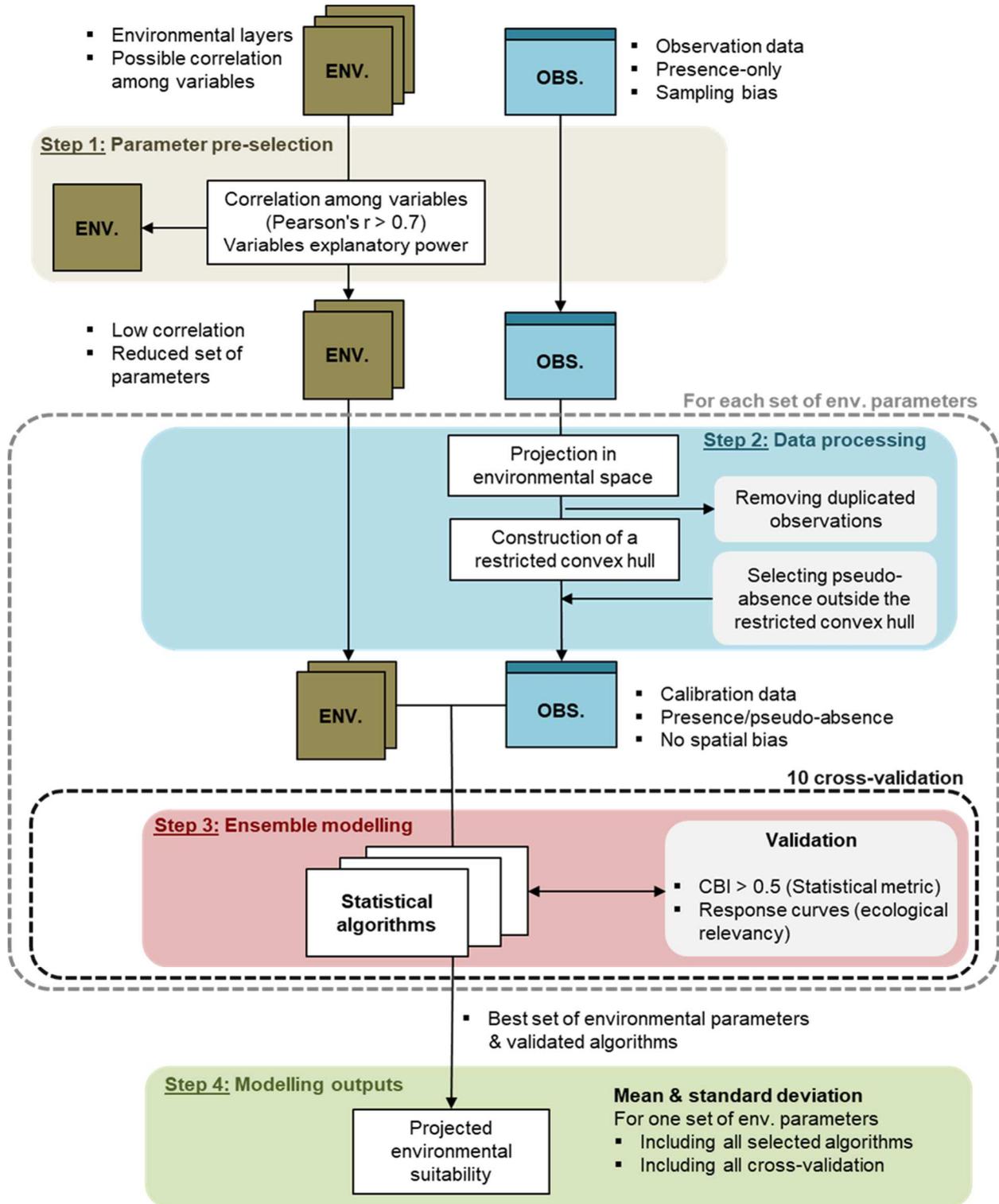
Figure 1: Sketch diagram of the modelling framework applied to model SPFs species. "ENV." = environmental parameters and "OBS." = georeferenced presence data.

785

786

Figure 2: Example of pseudo-absences generation for the Mediterranean horse mackerel (environmental

parameters = SST + SSTr, 1°C resolution). **A-C:** Species occurrences (black dots) in (**A**) the geographical domain

and (**C**) the environmental space. **B-D:** Species occurrences (black dots) and pseudo-absences (red dots) generated

from the restricted convex hull method in (**B**) the geographical domain and (**D**) the environmental space.

791

792
793

794

| Mediterranean horse mackerel | Parameters: | SST, SSTvar, log_PP |
	Models:	GLM, RF, NPPEN
	CBI (mean):	**0.71**
Atlantic horse mackerel	Parameters:	SST, SSTvar, log_PP
	Models:	GLM, NPPEN
	CBI (mean):	**0.95**
European pilchard	Parameters:	SST, SSTr, SSS
	Models:	GLM, GAM, NPPEN
	CBI (mean):	**0.75**
Round sardinella	Parameters:	SST, SSTr, log_PP
	Models:	GLM, RF, FDA, NPPEN
	CBI (mean):	**0.88**
European sprat	Parameters:	SST, SSTr, log_PP
	Models:	GLM, MARS, NPPEN
	CBI (mean):	**0.92**
European anchovy	Parameters:	SST, SSTvar, SSS
	Models:	GLM, FDA, MARS
	CBI (mean):	**0.88**
Bogue	Parameters:	SST, SSTr
	Models:	GLM, ANN, NPPEN
	CBI (mean):	**0.65**

795

796 **Table 2:** Environmental parameters and SDMs selected by our procedure.

797 The selected SDMs had a CBI>0.5 and satisfying response curves**. Parameters:** (**SST**) Sea Surface Temperature,

798 (**SSTr**) annual range of Sea Surface Temperature, (**SSTvar**) monthly variance of Sea Surface Temperature,

799 (**log_PP**) log-transformed Primary Production and (**SSS**) Sea Surface Salinity. **Models**: (**GLM**) Generalised

800 Linear Model, (**GAM**) Generalised Additive Model, (**GBM**) Generalised Boosting Model, (**ANN**) Artificial

801 Neural Network, (**FDA**) Flexible Discriminant Analysis, (**MARS**) Multiple Adaptive Regression Splines, (**RF**)

802 Random Forest and (**NPPEN**) Non Parametric Probabilistic Ecological Niche model.

803

804
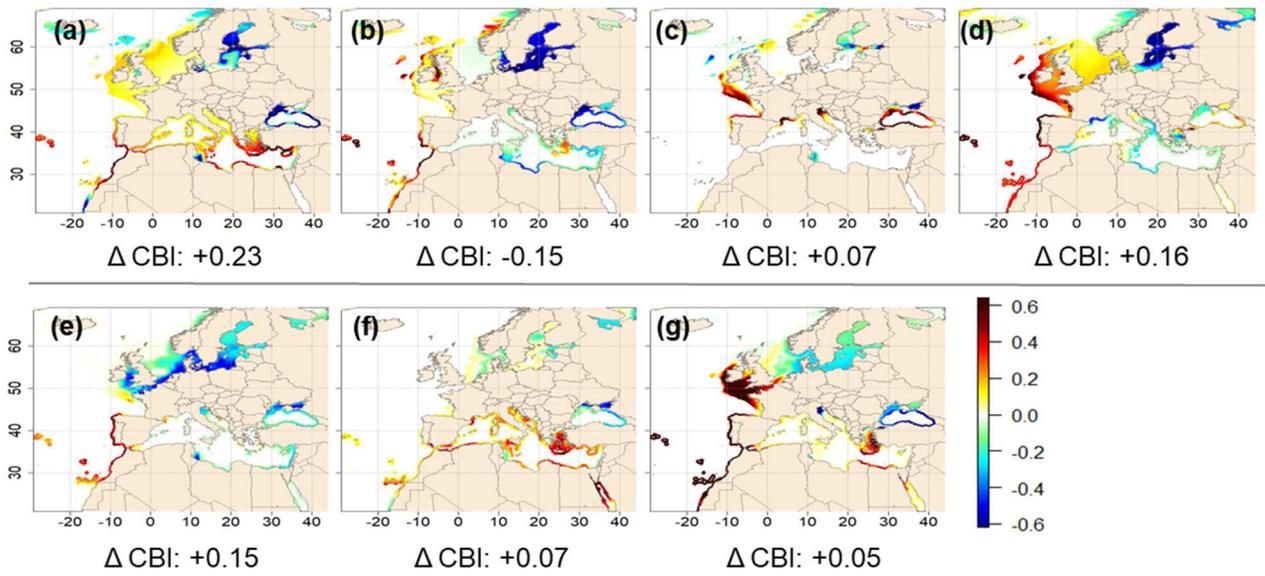
Figure 3: Environmental suitability index and CBI differences between ensemble models originating from our modelling framework and ensemble models constructed without data filtration and random pseudo-absence selection for (a) Atlantic horse mackerel, (b) European pilchard, (c) European sprat, (d) European anchovy, (e) Mediterranean horse mackerel, (f) round sardinella and (g) bogue.

807

808

809

810

811

812

813
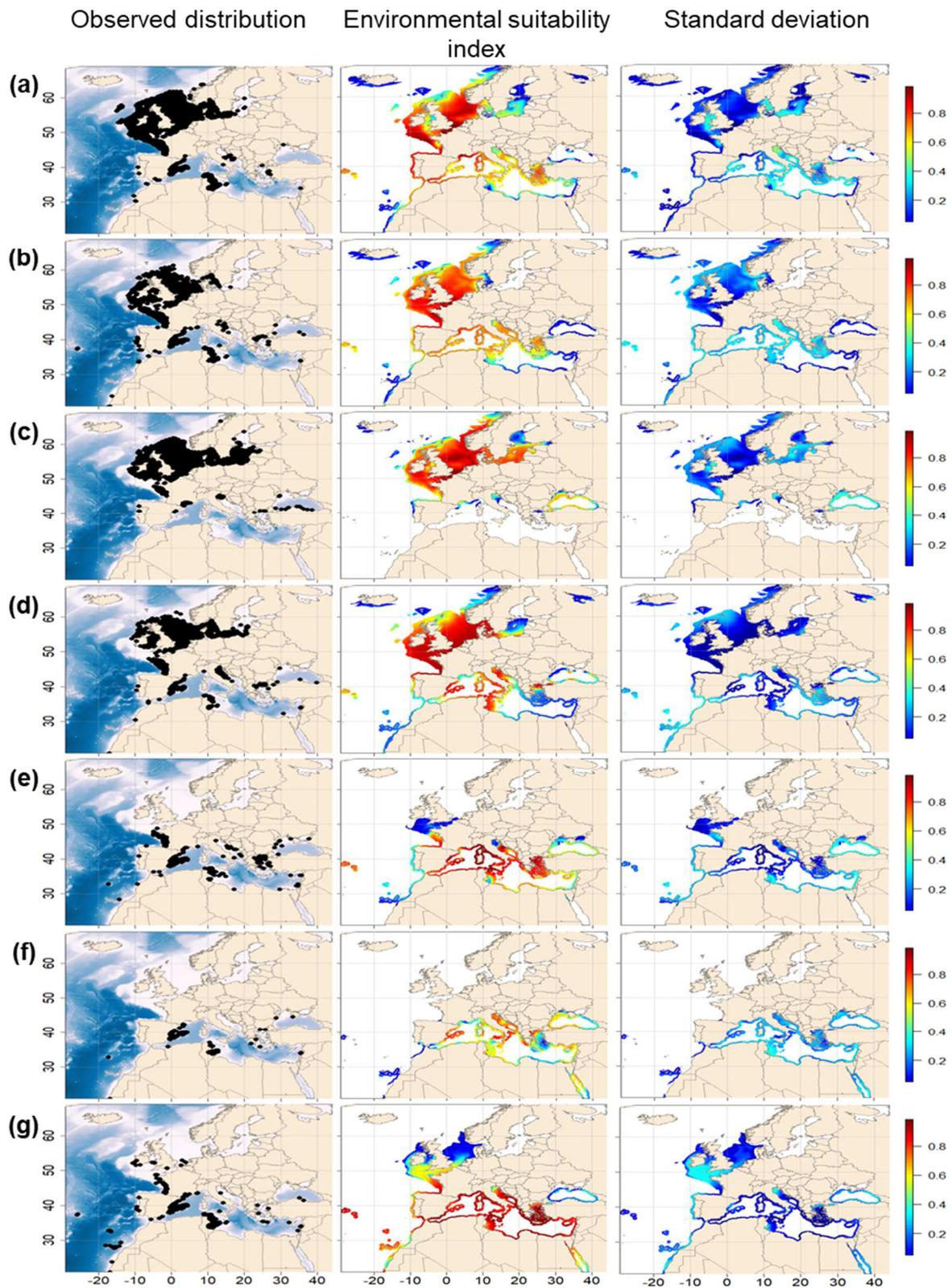
Figure 4: Contemporary (1990-2017) observed distribution (left panels), modelled environmental suitability index (0 to 1, middle panels) and its associated standard deviation (0 to 1, based on all validated SDMs and cross-validation runs, right panels) for (a) Atlantic horse mackerel, (b) European pilchard, (c) European sprat, (d) European anchovy, (e) Mediterranean horse mackerel, (f) round sardinella and (g) bogue.