Chefaoui, R. M., & Serrão, E. A. (2017).

Accounting for uncertainty in predictions of a marine species: Integrating population genetics to verify past distributions.

***Ecological Modelling*, *359*, 229-239.**

https://doi.org/10.1016/j.ecolmodel.2017.06.006

ARTICLE TYPE: Original research paper

TITLE

Accounting for uncertainty in predictions of a marine species: integrating population genetics to verify past distributions.

AUTHORS

Rosa M. Chefaoui[a]*, Ester A. Serrão[a]

[a] Centro de Ciências do Mar (CCMAR), CIMAR Laboratorio Associado, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal

Rosa M. Chefaoui email: rosa.chef@gmail.com

Ester A. Serrão email: eserrao@ualg.pt

*Author for correspondence: Rosa M. Chefaoui; Centro de Ciências do Mar (CCMAR), Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal;

email: rosa.chef@gmail.com

SHORT RUNNING HEAD: Uncertainty in predictions of a marine species

ABSTRACT

We develop a new perspective on the uncertainties affecting the predictions of coastal species distributions using patterns of genetic diversity to assess the congruence of hindcasted distribution models. We model the niche of the subtidal seagrass *Cymodocea nodosa*, for which previous phylogeographic findings are used to contrast hypotheses for the Last Glacial Maximum (LGM) in the Mediterranean and adjacent Atlantic coastal regions. We focus on amelioration of sampling bias, and explore the influence of other sources of uncertainty such as the number of variables, Ocean General Circulation Models (OGCMs), and thresholds used. To do that, we test geographical and environmental filtering of presences, and a species-specific weighted filter related to political boundaries for background data. Contrary to our initial hypothesis that reducing sampling bias by means of geographical, environmental or background filtering would enhance predictive power and reliability of the models, none of these approaches consistently improved performance. These counter-intuitive results might be explained by the higher relative occurrence area (ROA) inherent to linear coastal study areas in relation to terrestrial regions, which may cause worse predictions and, thus, higher variability among models. We found that the Ocean General Circulation Models (OGCMs), the threshold and, to a smaller extent, the number of variables used, conditioned greatly the variability of the predictions in both accuracy and geographic range. Despite these uncertainties, all models achieved the goal of identifying long-term persistence regions (glacial refugia) where the highest genetic diversity for *Cymodocea nodosa* is found nowadays. However, only the CCSM corroborated the hypothesis, raised in previous studies, of a vicariant process in shaping the species' genetic structure.

KEYWORDS

Ecological niche modelling; genetic diversity; Last Glacial Maximum; Ocean General Circulation Models; sampling bias; threshold.

1. INTRODUCTION

Although modelling the niche of species has proven to be an efficient approach to ecological, conservational and biogeographical questions during the last decades (see Guisan & Thuiller, 2005; Araújo & Peterson, 2012), a growing number of studies focused on the terrestrial realm highlights the importance of assessing the sources of uncertainty affecting species distribution models (SDMs) in order to obtain more reliable predictions. The potential sources of uncertainty may be related to a large list of factors ranging from the data included in the models to the algorithm used (see e.g. Rocchini et al., 2011; Beale & Lennon, 2012; Gould et al., 2014 for a review). Many of the studies on uncertainty in ecological niche modelling pay particular attention to sampling bias in occurrence data caused by proximity to cities, rivers, roads or conservation reserves (e.g. Reddy & Dávalos, 2003; Kadmon et al., 2004) or historical bias (Hortal et al., 2008), as they may result in inaccurate estimations of niche. More recent attention has focused on the provision of methodologies to ameliorate geographic bias by filtering (or "thinning") occurrences to avoid clumping and autocorrelation (Veloz, 2009; Beck et al., 2014; Aiello-Lammens et al., 2015), or to correct the derived bias in the environmental space which improved performance of the models in comparison to the use of geographic filters (Varela et al., 2014; de Oliveira et al., 2014). Finally, filters for background data have also been applied (Phillips et al., 2009; Kramer-Schadt et al., 2013), as the location of pseudo-absences or background data also affects predictions (Zaniewski et al., 2002; Chefaoui & Lobo, 2008; Merow et al., 2013).

Despite previous efforts to reduce uncertainty, much of the research up to now referred to biases shown by terrestrial species, while uncertainties related to the study of marine taxa have been poorly investigated. The relatively recent availability of satellite data on marine environmental variables has driven an increase in the use of SDMs in this realm and, consequently, studies suited to its peculiarities are needed. One particularly relevant distinction is that most important marine species are coastal and therefore have a distribution that is more linear than bidimensional. Regarding sampling bias, coastal studies are not biased by the same geographic elements as

terrestrial do (e.g. distance to roads). The proximity to cities or research institutions with diving centers may affect sampling effort, and at a large scale, the observed under-surveyed regions are coincident with countries with limited investment in research and political instability such as those located in coastal regions of North Africa (Chefaoui et al., 2016). In the Global Biodiversity Information Facility (GBIF, www.gbif.org/), it has been shown that more records of the common Eurasian butterfly are available from developed countries though the real occurrence of the species is higher in less developed ones (Beck et al., 2014). These differences among countries might be even more prevalent concerning subtidal marine habitats, where sampling marine campaigns are more expensive and require more infrastructure and technical expertise.

The unavailability of predictors is another source of uncertainty affecting niche projections to the past for marine species in comparison with terrestrial ones. Uncertainties affecting the General Circulation Models (GCMs) pertaining to the Coupled Model Intercomparison Project (CMIP) have been reported both for oceanic variables (Wang et al., 2013) and for atmospheric ones (Svenning et al., 2008; Braconnot et al., 2012; Varela et al., 2015). However, although those uncertainties are common to both marine and terrestrial realms and despite the improved reconstructions of climate during the LGM (Last Glacial Maximum; 23 000–18 000 years BP), there is a scarcity of ocean paleoclimatic variables corresponding to the Ocean GCMs (OGCMs) for the LGM in comparison with the models of atmospheric variables used in terrestrial studies. This limits the possibilities to use a wider set of oceanic predictors for the LGM to hindcast marine niches (see e.g. Chefaoui et al 2017). There is also an observable incompleteness of data for all seas and oceans, as important gaps are found in some models (e.g. Black Sea), and some OGCMs are available just at a coarse resolution (~2º).

To provide a new perspective for marine coastal species, we explore these uncertainties in SDMs for the subtidal seagrass *Cymodocea nodosa* (Ucria) Ascherson. Although *C. nodosa* is the most common seagrass in the eastern Mediterranean, no exact mapping has been carried out and scanty reports of its distribution exist from some countries (e.g., Egypt, Lebanon…) in comparison

105 with the western basin (Green & Short, 2003). We will estimate if sampling bias linked to
106 geographical boundaries is affecting data input of *C. nodosa* to calibrate niche models and,
107 consequently, derived past projections. If so, we will try to ameliorate sampling bias under the
108 expectation that the accuracy of niche models for coastal species may be improved similarly as
109 terrestrial ones by the use of filters for species data. Thus, in addition to testing geographical and
110 environmental filters for presences, we will seek to ameliorate a possible political bias by creating a
111 weighted probability filter for background data in accordance to political boundaries. This "weight"
112 filter will combine an estimation of sampling effort for *C. nodosa* in relation to the available
113 distribution records of similar marine plants from GBIF, with the developmental level of each
114 country.

115 We will also test if the lack of oceanic predictors pertaining to climate simulations for the
116 LGM might be hindering the reliability of SDM-based hindcasting marine studies. Under the
117 assumption that differences found using current climatic predictors may prevail through niche
118 hindcasting, we give an account of the uncertainty caused by the reduced availability of oceanic
119 LGM variables examining performance differences among current climate models obtained using
120 nested groups of predictors varying in number. Finally, we will also take into consideration the
121 effect on our predictions of the threshold used to validate and transform probabilities into binary
122 outputs, which has been identified as a yet unexplored source of uncertainty modelling species
123 range shifts (Nenzén & Araújo, 2011).

124 To assess these uncertainties on our current and LGM models for *C. nodosa*, we will take
125 advantage of previous ecological (Chefaoui et al., 2015) and genetic (Alberto et al., 2008; Masucci
126 et al., 2012) findings allowing us to assess congruence between the hypotheses supported by
127 different approaches. Species distribution shifts throughout the Quaternary leave genetic signatures
128 on populations (Hewitt 2004). High congruence between the distribution of present intraspecific
129 genetic diversity and regions of long-term population persistence since the LGM have been found
130 for several marine species (Assis et al. 2014; Neiva et al. 2014; Assis et al. 2016; Chefaoui et al

2017). Thus, finding a congruence between genetic information and SDMs for the LGM could help us identify possible uncertainties in the SDMs. We will contrast our hindcast results with the predictions made based on the present genetic structure of the species along its distribution, namely environmental barriers limiting gene flow inferred by Alberto et al. (2008) and Masucci et al. (2012) for the species across the Mediterranean and Atlantic. Alberto et al. (2008) identified four regions with a strong genetic structure, and two potential imprints of vicariance located in the range areas with presumably higher sea surface temperature (SST), namely the low-latitude Atlantic (AL) and the Eastern Mediterranean (EM) regions (Fig. 1). Masucci et al. (2012) inferred the gene flow directionality among *C. nodosa* populations. We will examine if these hypothesized glacial refugia (persistence regions with high genetic diversity) are supported by the LGM projections obtained by all models to evaluate the most accurate prediction on the basis of independent genetic data.

Our goals are to examine for the first time how sampling bias, and other sources of uncertainty such as the availability of variables, the Ocean General Circulation Model (OGCM), and the threshold used, affect the predictions of a coastal marine species, in this case a subtidal seagrass, comparing predictions derived from our models with previous predictions derived from data on genetic diversity throughout the species range. This paper investigates those uncertainties for current and Last Glacial Maximum (LGM) predictions, using the observed geographical distribution of extant genetic diversity for assessment of congruence of the predictions.

2. METHODS

2.1. Effect of the number of variables on performance of models under current conditions

We used 299 records of *Cymodocea nodosa* compiled from the literature, the Global Biodiversity Information Facility (GBIF, www.gbif.org/), and Algaebase (Guiry & Guiry, 2014), covering the entire species range (Mediterranean Sea, North-East Atlantic coasts and the Black Sea) at any depth along its known distribution (0 to 35 m depth). A total of 210 presence cells remained after georeferencing data to a 0.083°×0.083°(~9.2 km) grid resolution (Fig. 1). To assess the influence of a reduction in the number of variables on performance, we compared different sets of nested

158 predictors under current conditions, using a specific SDM for this purpose (see Fig. 2). The
159 complete set of 18 variables was comprised of the environmental variables and landscape metrics
160 previously found to be the best for modelling the niche of *C. nodosa* in Chefaoui et al. (2016).
161 Environmental predictors were: sea surface temperature (SST) of winter and summer, diffuse
162 attenuation coefficient (Kd), wave height, nitrate, phosphate, pH, photosynthetically available
163 radiation (PAR), and salinity (Appendix A in the Supplementary material, Table A.1). Landscape
164 metrics were: mean edge contrast index distribution (ECON_MN), area weighted mean fractal
165 dimension index (FRAC_AM), mean perimeter–area ratio (PARA_MN), percentage of landscape
166 (PLAND), mean shape index (SHAPE_MN) and total edge contrast index (TECI). These landscape
167 metrics were previously found to be good predictors of coastal morphology (Chefaoui, 2014) and of
168 the presence of *C. nodosa* (for a complete description see Chefaoui et al. (2016) and Appendix A in
169 the Supplementary material, Table A.2). From this entire set we selected just the variables which
170 were also available for LGM climate scenarios and were not highly correlated ($r \geq |0.80|$, $p < 0.001$),
171 which resulted in a set of 9 variables: minimum SST of winter, maximum SST of summer, salinity,
172 and the six landscape metrics. In addition, we also tested the two SST variables selected to perform
173 the modeling approach using filters (minimum SST of winter and maximum SST of summer). This
174 last set was chosen on the basis of a good performance of SST determining northern and southern
175 range limits for *C. nodosa* in Chefaoui et al. (2016). We used the "biomod2" package (Thuiller et
176 al., 2014) to perform six presence–absence techniques: generalized linear model (GLM),
177 generalized additive model (GAM), generalized boosting model (GBM), flexible discriminant
178 analysis (FDA), multiple adaptive regression splines (MARS), randomForest (RF), and subsequent
179 ensembles. A total of 180 models were run for each group of predictors (10 iterations x 3 pseudo-
180 absence sets x 6 methods), using the same procedure described by Chefaoui et al. (2016). We
181 computed afterwards the "committee averaging" ensemble (the average of binary predictions) as
182 this method obtained better accuracy scores than the ensemble produced estimating the mean
183 probabilities in Chefaoui et al. (2016) for *C. nodosa*. The Wilcoxon signed-rank test was used to

compare the performance of the models produced using a reduced number of variables with those obtained with the complete set.

2.2. Estimation of sampling bias and amelioration by filters

First, we estimated the existence of a climatic bias using the Kolmogorov-Smirnov two-sample test to compare the climatic values of descriptor variables in occurrence cells with a random distribution within the same range of values in the study area. To assess also the existence of a geographical bias related to different survey effort, we tested if frequencies of observed occurrences of *Cymodocea nodosa* in each country were different from random. To correct these possible biases we created three filters, two for presence data and one for background data to be used in MaxEnt (Phillips et al., 2006). An environmental filter (ENV) and a geographic filter (GEO) were produced to discard aggregated presences according to the procedure described by Varela et al. (2014). Minimum SST of winter and maximum SST of summer (the same set of two SST variables as in the previous experiment) were selected to produce the ENV filter and the models. GEO filter for presence data was elaborated using latitude and longitude. For both filters, each pair of variables were the axes of a grid used as stratification to extract subsamples using the "gridSample" function of dismo packge (Hijmans et al., 2014).

To account for the geographical boundaries bias, we created a filter to be applied to the background data. A weighted sampling probability filter (from now on, "WEIGHT") was produced for each country delimited by terrestrial and marine political boundaries. To create the WEIGHT filter we took into account: i) the number of occurrences of *Cymodocea nodosa*; ii) the number of occurrences available in GBIF for all species of marine plants present in the study area of the same order as *C. nodosa* (order Alismatales) occupying similar marine habitats and therefore likely to receive a similar sampling effort: 4877 records representing 5 genera (*Cymodocea*, *Halophila*, *Posidonia*, *Ruppia* and *Zostera*); iii) data publishing activity in the GBIF for each country (www.gbif.org/country/); and iv) the level of development in each country according to the United Nations (www.naturalearthdata.com). To create the WEIGHT filter, these data were classified by

countries and contrasted using a set of condition alternatives by means of decision rules (see Appendix A in the Supplementary material, Figs. A.1 and A.2, Table A.3).

2.3. Modelling approach and LGM projections using filters

To test the effect of sampling bias, OGCMs, and thresholds on LGM predictions, we used the sets of predictors which had available LGM variables: the set of nine variables, and the set of two SST variables (Fig. 2). Paleoclimate variables were obtained from the LGM experiment pertaining to the Coupled Model Intercomparison Project (CMIP5, http://cmip-pcmdi.llnl.gov/cmip5/). From all available LGM models, CCSM4 (over 100 years) and CNRM-CM5 (over 200 years) were selected on the basis of these criteria: (a) to have a resolution ≤ 1 degree, (b) inclusion of Black Sea data, (c) availability of both SST and salinity variables. We calculated maximum summer SST and minimum winter SST from the LGM long-term monthly SST means of each model. Additionally, the long-term mean sea surface salinity (SSS) for both models was also obtained. To elaborate our LGM land-sea mask and delimit the study area, a mean sea-level change of -116 m was calculated from bathymetric data derived from the General Bathymetric Chart of the Oceans (GEBCO; http://www.gebco.net/data_and_products/gridded_bathymetry_data/). This estimated sea-level change is consistent with the ice-sheet reconstruction used in the Paleoclimate Modelling Intercomparison Project Phase III (PMIP 3, pmip3.lsce.ipsl.fr/). From this coastline, we recalculated for the LGM the six landscape metrics used in the set of nine variables. As it is beyond the scope of this study to examine the variability caused by different modelling techniques, we chose MaxEnt (Phillips et al., 2006), a maximum entropy technique which uses presence and background data. This choice is due to replicate methodologies for filtering (Varela et al., 2014), and also for being widely used for projections. We used MaxEnt with "dismo" package to model the niche using the different filters and sets of variables. MaxEnt models were generated by splitting raw data (n=210) into a calibration set (n=167 ~ 80%) and a validation set (n=43 ~ 20%). Filters ENV and GEO were obtained from the calibration set as well as a random presence set for comparison. Each presence set size was equal to 100, an amount sufficient according to Varela et al. (2014). To calibrate the

236 models, we used WEIGHT filter and a random set (n=4000, each) as background data. The
237 combinations of presence, background data and variables sets originated eight different models
238 (Fig. 2) which were iterated 100 times.
239 To validate the models we measured the area under the receiver operating characteristic
240 (ROC) curve (AUC), sensitivity (presences correctly predicted), and specificity (absences correctly
241 predicted). We calculated three thresholds using "dismo" package in R to transform predicted
242 probability into binary values: i) the prevalence: threshold at which the predicted prevalence
243 (proportion of locations where the species is predicted to be present) is closest to the observed
244 prevalence; "Spec_sens": the value at which the sum of the sensitivity and specificity is highest;
245 and "No omission": the highest threshold at which there are no omission errors (all presence data
246 are classified as presences). We used a Wilcoxon signed-rank test to compare AUC measures
247 among filters using the same threshold. The eight models were projected into the CCSM4 and
248 CNRM-CM5 scenarios as well as an ensemble of both calculated as the mean value. All analyses
249 were run in R (R Core Team, 2014).
250 *2.4. Comparison of regions of persistence with genetic diversity*
251 The LGM projections obtained with our approaches were used to identify the distribution of the
252 regions of long-term persistence (where high probability of presence is found for the species under
253 current and LGM conditions). These were then compared with the long-term persistence zones
254 hypothesized based on population genetic diversity data in a previous study (Alberto et al. 2008).
255 The purpose here was to estimate the congruence between hypotheses derived from independent
256 data: those based on genetic data and our niche models. The allelic richness (Â, number of alleles)
257 and expected heterozygosity (He, gene diversity; Nei, 1978) obtained by Alberto et al. (2008) were
258 used as measures of genetic diversity. That study analysed genetic diversity of 47 populations
259 covering the Mediterranean and Atlantic distribution of *C. nodosa*. We calculated the mean values
260 of Â and He for the four genetic clusters identified by Alberto et al. (2008): low-latitude Atlantic
261 (AL), high-latitude Atlantic (AH), Western Mediterranean basin (WM), and Eastern Mediterranean

basin (EM) (Fig. 1). The frequencies of cells classified as having higher probability of occurrence than the threshold were computed after a binary transformation of current and LGM SDMs using the mean value of the three thresholds explained in section 2.3 for each model. Afterwards, we estimated the intersection among SDMs for current and LGM periods and each genetic region to calculate the habitat area of persistence regions, where the species could have found appropriate long-term conditions. Similarity between raster predictions was estimated using a Pearson correlation test.

The existence of a previous study inferring the network of directions of gene flow among meadows of *C. nodosa* (Masucci et al. 2012), based on computing the pairwise difference of the genetic information shared or exclusive of the gametes from each meadow, allowed further assessment of congruence between hypotheses raised by distinct studies. We thus compared also if this gene flow network among genetic clusters is in agreement with the refugia found in this study.

3. RESULTS

3.1. Sampling bias estimation

The Kolmogorov-Smirnov test found a distribution of the presences significantly different from random both in climatic (minimum SST of winter: D = 0.3, p-value = 1.238e-08; maximum SST of summer: D = 0.3095, p-value = 3.66e-09) as in geographic space (D = 0.4194, p-value = 0.008579), thus evidencing the observed sampling bias. The distribution of marine plants records (Alismatales) seemed to have a representation similar to the overall GBIF records in each country (Pearson's product-moment correlation (r) = 0.97, $p < 0.001$). Although marine plant occurrences were also correlated to those of *C. nodosa* ($r = 0.78$, $p < 0.001$), we used the differences found in the number of records to estimate the sampling effort for *C. nodosa* in our study area in relation to plants of the same order (Alismatales, angiosperms growing also in marine habitats) by means of decision rules (Appendix A in the Supplementary material, Figs. A.2 and A.3) and produce the WEIGHT filter (Fig. 3).

3.2. Effect of the number of variables

There were significant differences among the three sets of predictors in terms of accuracy and predicted probability of presence for the present. The Wilcoxon signed-rank test revealed significantly lower AUC values in the models obtained using nine and two variables in comparison to the complete set (Table 1). These differences in AUC were consistent across the techniques used. In general, specificity and sensitivity also decreased, though specificity was more influenced by the reduction of variables. Just one method (GLM) showed a sensitivity that was significantly higher using nine or two variables in comparison to the complete set.

In order to establish a comparison among the probability of presence predicted by each set, we selected the ensemble that achieved the best scores in the complete set (the "committee averaging" ensemble computed with the average of binary predictions of the GAM models) as there was no agreement for a best method for all sets. Comparing the "committee averaging" of GAMs, we found that sets with fewer variables tended to overpredict and obtained a higher density of cells of elevated probability of presence (Fig. 4).

3.3. Differences among filters

Mean AUC values obtained by each filter differed according to the threshold used (Fig. 5). Thus, there was not a filter that consistently improved model results. "No omission" and "prevalence" were the thresholds which optimized the sensitivity of the models with values close to 1, while "Spec_sens" increased specificity (Appendix A in the Supplementary material, Fig. A.4). Wilcoxon test used to compare the scores of AUC among filters revealed more significant differences among models using "prevalence" threshold (89.28% of 28 cases were significant: 22 at p-value< 0.001 and 3 at p-value< 0.05) than "No omission" (57.14% significant: 12 at p-value< 0.001 and 4 at p-value< 0.05), or "Spec_sens" (42.85% significant at the 0.001 level). Using "prevalence" threshold, non filtered models achieved significantly better AUC scores than filtered ones in the set of two variables, and there were no significant differences between filtered or not using the nine variables set (Fig. 5). Similarly, models obtained with nine variables (filtered and non-filtered) showed better or worse results than the two variables set depending on the threshold used. Models produced with

315 the WEIGHT filter were completely correlated (r=1) with their homologous in presence data sets
316 using two variables. Just LGM projections using CNRM model showed remarkably little variation
317 (Table 2).

318 *3.4. Congruence between genetic diversity and LGM SDMs*

319 There was a high similarity in terms of probability of presence among the raster predictions of each
320 filter. All filters produced with the same set of variables showed Pearson's correlation coefficients
321 higher than 0.9, and maps obtained using both sets of variables (two or nine) showed also
322 correlations > 0.7(Table 2). EM, the region which obtained a higher genetic diversity both in terms
323 of allelic richness and expected heterozygosity (Fig. 1), was also the one showing a major area with
324 higher probability of presence in most of the models under current and past conditions (Fig. 6).
325 However, there were differences regarding the persistence regions between CCSM and CNRM
326 models and the threshold used. Although all LGM scenarios (CCSM, CNRM and the ensemble of
327 both) found a possible persistence of the species in AL and EM regions, just the threshold
328 "Spec_sens" applied on CCSM showed also the genetic discontinuity identified by Alberto et al.
329 (2008) (Fig. 7).

330

331 **4. DISCUSSION**

332 Applying different filters to presence and background data did not enhance model performance,
333 though our findings suggest that known occurrences of *Cymodocea nodosa* show a biased
334 distribution both in the geographical and environmental space. Despite finding significant
335 differences in predictive power among filters, those were dependent on the threshold used to
336 validate the models. Therefore, a better performance of geographic or environmental filters cannot
337 be concluded, as even non-filtered models performed better according to some thresholds. Contrary
338 to expectations, our specific WEIGHT filter to ameliorate bias due to differences in survey effort
339 from each country produced a negligible effect on performance in comparison with the models
340 obtained with the same set of presences, filtered or not. At the moment, we can only compare our

results on reduction of sampling bias with studies of terrestrial species. Manipulation of the background data by (Kramer-Schadt et al., 2013) also caused weak improvements. Our results are also consistent with other studies which did not find major improvements correcting geographical bias (Syfert et al., 2013; Varela et al., 2014), though differ from those which obtained an increase in accuracy (e.g. Kramer-Schadt et al., 2013).

The performance of the environmental filter was an improvement in some studies (Varela et al., 2014; de Oliveira et al., 2014), but our findings differ, even from those obtained by Varela et al. (2014) who - using just one threshold value and a virtual species - used a similar approach. This discrepancy could be attributed to intrinsic characteristics of AUC, which is influenced by the threshold, the extent of the study area, and it is only accurate when using true absences (Lobo et al., 2008; Jiménez-Valverde, 2012). In addition, these seemingly counterintuitive results may be due to the idiosyncrasy of coastal areas, defined by its linear shape and a lower number of cells than terrestrial studies. The particularities of linear study areas as is the case of coastal areas, affect the proportion of occupied area by the species in relation to the total area, known as the relative occurrence area (ROA; Lobo, 2008; Lobo et al., 2008), which has been identified as an important factor influencing performance of the models, because species with lower ROA are predicted more accurately (Chefaoui et al., 2011; Tessarolo et al., 2014). A species with low ROA shows a similar relation between presences/absences in its habitat as a specialist species, which implies a higher discrimination power of its predictive models (Lobo et al., 2008; Jiménez-Valverde et al., 2008). The linearity of coastal studies indefectibly involves a higher ROA than most terrestrial studies - regardless of the studied species - thus, the effect of ROA might reduce predictability, increasing the variability between models and, therefore, their uncertainty. The smaller extent of coastal areas may also contribute to reduce the heterogeneity of the study area and mitigate the effects of both geographical and environmental filtering in comparison to the larger areas usually found in terrestrial studies.

Concerning other sources of uncertainty, our results are in agreement with those of Nenzén

and Araújo (2011), who demonstrated that the threshold selected to produce a binary transformation affected greatly the projections of bioclimatic envelope models. We also found great variability caused by the election of this value, evidencing the need for using a range of thresholds in comparisons among filtered models for confirmation of consistent results. Besides, this study has shown that the number of variables also caused differences in accuracy and projections. Though a more complete set of variables increased the discriminative power and reduced the potential area predicted by the models, the SST set (minimum SST of winter and maximum SST of summer) were also able to produce accurate models (Table 1, Fig. 4). Thus, just two SST variables seem enough to summarize changes in the range of distribution of *C. nodosa* across time. These two variables were also sufficient to provide accurate models for *Posidonia oceanica* in a previous study (Chefaoui et al 2017). This may be explained because Mediterranean east-west LGM temperature gradients were wider in comparison to present day gradients (Hayes et al., 2005). However, as the lack of variables is not a problem for terrestrial species as much as for marine ones, this source of uncertainty is not very well explored. Though we have not found in the particular case of *C. nodosa* a relevant reduction of accuracy by only using two SST variables, that might not be the case for other marine species whose distribution might be more influenced by other variables different from SST and salinity. Thus, in waiting for the availability of a major variety of variables in OGCMs, further studies would be needed to estimate if there is a generality among marine species.

The predictions of long term climatic refugia from previously studied distribution of population genetic diversity and differentiation along the species range were in agreement with the long term persistence zones inferred by our data. Despite the large differences between hindcasts obtained from the CCSM and CNRM scenarios, they all had a remarkable agreement in reference to the long-term persistence in the AL and EM regions, which correspond to the locations with the highest genetic diversity, a pattern that indicates long-term persistence of large populations that accumulate diversity over time in the absence of major bottlenecks. Both CCSM and CRNM hindcasts were also congruent in finding the major potential habitat in EM. However, only the

393 model produced with CCSM supported the vicariance hypothesis proposed by Alberto et al. (2008).
394 These results are not surprising as a multimodel comparison among LGM SST simulations (Wang
395 et al. 2013) found large discrepancies in the mid-latitude ocean. We have also detected that the
396 anomalies between CCSM and CNRM differ in distribution between the minimum and the
397 maximum SST (see Appendix A in the Supplementary material, Fig. A.5). Thus, not just the
398 OGCM used, but also the relative relevance of each measure of SST in the model, for a particular
399 marine species, can be a potential source of uncertainty in the hindcasts. Ensemble approaches have
400 been proposed to average uncertainties (Araújo & New, 2007; Thuiller et al., 2009), but it is the
401 marine researcher who has to find a compromise among the number of OGCMs used, the available
402 resolution of those OGCMS, and their completeness. In our case, we produced an ensemble using
403 just the two OGCMs which fitted our resolution and completeness criteria (e.g. including the Black
404 Sea). This ensemble described a high probability of presence for the AL region and most of the
405 southern Mediterranean Sea (Fig. 7). Other studies could test if the use of more OGCMs could help
406 reducing uncertainty despite sacrificing data resolution.
407 The model that was most congruent with the independent evidence derived from genetic
408 data, was the LGM CCSM (Fig. 7). This model supported the hypothesis that suitable habitat for the
409 species could have existed during the LGM in two very distant regions: the low-latitude Atlantic
410 range edge (AL, the western coast of Saharian Africa) and the Eastern Mediterranean basin (EM),
411 which coincide nowadays with very genetically differentiated groups. Those regions showing
412 higher SST in the LGM and currently, could have acted as glacial refugia, in contrast with habitats
413 with lower probability of presence in the past, which were located in the AH and WM regions, thus
414 supporting the vicariance hypothesis proposed by Alberto et al. (2008). A recent study by Masucci
415 et al. (2012) inferred the pathways of gene flow of the species (Fig. 7), confirming a main flow
416 from Western Africa and Morocco (AL region) towards the Canary Islands, and another from Sicily
417 (limit between EM and WM region) to the Western Mediterranean basin (WM). This vicariance
418 hypothesis cannot be definitely supported by our results because it is only congruent with our

CCSM hindcast, therefore this remains an open question.

This study has found that the threshold, the OGCM and, to a lesser extent, the number of variables used, represent larger sources of uncertainty than the environmental or geographical bias of existing distributional records of *C. nodosa*. Though the use of environmental filters seems an intuitive approach improving predictions of terrestrial species, the linear shape of coastal studies may originate a higher ROA and/or different environmental heterogeneity, which might result in that the filtering process may not cause equivalent improvements in coastal studies. More research on coastal areas would help us to establish if differing levels of reported georeferenced data for species presence, as found in GBIF or other databases, can influence our understanding of the distribution of a species, and more robust conclusions on the effect of the ROA. As improvements occur in the OGCMs for LGM, projections to the past will gain reliability to allow further testing of the hypothesis of existence of just two climatic refugia for *Cymodocea nodosa* during the LGM, located at its warmer range limits in Africa and the eastern Mediterranean (a hypothesis supported by one model and by the present genetic data), versus the alternative hypothesis of existence of a wider glacial refugia.

ACKNOWLEDGEMENTS

We thank Matthew Fitzpatrick and anonymous referees for their useful comments on a previous version of the manuscript. We acknowledge funding by the Fundação para a Ciência e a Tecnologia (FCT, Portugal) as postdoctoral fellowship SFRH/BPD/85040/2012 to RMC and funding program UID/Multi/04326/2013, and the Pew Foundation (USA).

REFERENCES

Aiello-Lammens M.E., Boria R.A., Radosavljevic A., Vilela B., & Anderson R.P. (2015) spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, **38**, 541–545.

Alberto F., Massa S., Manent P., Diaz-Almela E., Arnaud-Haond S., Duarte C.M., & Serrão E. a. (2008) Genetic differentiation and secondary contact zone in the seagrass *Cymodocea nodosa*

across the Mediterranean-Atlantic transition region. *Journal of Biogeography*, **35**, 1279–1294.

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–7.

Araújo M.B. & Peterson A.T. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Assis J., Serrão E.A., Claro B., Perrin C., & Pearson G.A. (2014) Climate-driven range shifts explain the distribution of extant gene pools and predict future loss of unique lineages in a marine brown alga. *Molecular ecology*, **23**, 2797–810.

Assis J., Coelho N.C., Lamy T., Valero M., Alberto F., & Serrão E.A. (2016) Deep reefs are climatic refugia for genetic diversity of marine forests. *Journal of Biogeography*, **43**, 833–844.

Beale C. & Lennon J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 247–258.

Beck J., Böller M., Erhardt A., & Schwanghart W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, **19**, 10–15.

Braconnot P., Harrison S.P., Kageyama M., Bartlein P.J., Masson-Delmotte V., Abe-Ouchi A., Otto-Bliesner B., & Zhao Y. (2012) Evaluation of climate models using palaeoclimatic data. *Nature Climate Change*, **2**, 417–424.

Chefaoui R., Lobo J., & Hortal J. (2011) Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, **34.2**, 229–247.

Chefaoui R.M. (2014) Landscape metrics as indicators of coastal morphology: A multi-scale approach. *Ecological Indicators*, **45**, 139–147.

Chefaoui R.M., Assis J., Duarte C.M., & Serrão E.A. (2016) Large-Scale Prediction of Seagrass Distribution Integrating Landscape Metrics and Environmental Factors: The Case of *Cymodocea nodosa* (Mediterranean–Atlantic). *Estuaries and Coasts*, **39**, 123–137.

Chefaoui R.M., Duarte C.M., & Serrão E.A. (2017) Palaeoclimatic conditions in the Mediterranean explain genetic diversity of *Posidonia oceanica* seagrass meadows. *Scientific Reports*, in press, DOI: 10.1038/s41598-017-03006-2.

Chefaoui R.M. & Lobo J.M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.

Gould S.F., Beeton N.J., Harris R.M.B., Hutchinson M.F., Lechner A.M., Porfirio L.L., & Mackey B.G. (2014) A tool for simulating and communicating uncertainty when modelling species distributions under future climates. *Ecology and evolution*, **4**, 4798–811.

Green E.P. & Short F.T. (2003) *World Atlas of Seagrasses.* Univ of California Press, Berkeley, USA.

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

Hayes A., Kucera M., Kallel N., Sbaffi L., & Rohling E.J. (2005) Glacial Mediterranean sea surface temperatures based on planktonic foraminiferal assemblages. *Quaternary Science Reviews*, **24**, 999–1016.

Hewitt G.M. (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **359**, 183–195.

Hijmans, Robert J., Phillips, Steven, Leathwick, John, Elith J. (2014) dismo: Species distribution modeling. R package version 1.0-5.

Hortal J., Jiménez-Valverde A., Gómez J., Lobo J., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Jiménez-Valverde A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

Jiménez-Valverde A., Lobo J.M., & Hortal J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, **14**, 885–890.

Kadmon R., Farber O., & Danin A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kramer-Schadt S., Niedballa J., Pilgrim J.D., Schröder B., Lindenborn J., Reinfelder V., Stillfried M., Heckmann I., Scharf A.K., Augeri D.M., Cheyne S.M., Hearn A.J., Ross J., Macdonald D.W., Mathai J., Eaton J., Marshall A.J., Semiadi G., Rustam R., Bernard H., Alfred R., Samejima H., Duckworth J.W., Breitenmoser-Wuersten C., Belant J.L., Hofer H., & Wilting A. (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, **19**, 1366–1379.

Lobo J.M. (2008) More complex distribution models or more representative data? *Biodiversity Informatics*, **5**, 14–19.

Lobo J.M., Jiménez-Valverde A., & Real R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Masucci A.P., Arnaud-Haond S., Eguíluz V.M., Hernández-García E., & Serrão E.A. (2012) Genetic flow directionality and geographical segregation in a *Cymodocea nodosa* genetic diversity network. *EPJ Data Science*, **1**, 11.

Merow C., Smith M.J., & Silander J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Nei M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–90.

Neiva J., Assis J., Fernandes F., Pearson G.A., & Serrão E.A. (2014) Species distribution models and mitochondrial DNA phylogeography suggest an extensive biogeographical shift in the high-intertidal seaweed *Pelvetia canaliculata*. *Journal of Biogeography*, **41**, 1137–1148.

Nenzén H.K. & Araújo M.B. (2011) Choice of threshold alters projections of species range shifts

under climate change. *Ecological Modelling*, **222**, 3346–3354.

de Oliveira G., Rangel T.F., Lima-Ribeiro M.S., Terribile L.C., & Diniz-Filho J.A.F. (2014) Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. *Ecography*, **37**, 637–647.

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

Phillips S.J., Dudík M., Elith J., Graham C.H., Lehmann A., Leathwick J., & Ferrier S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications : a publication of the Ecological Society of America*, **19**, 181–97.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reddy S. & Dávalos L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Svenning J.-C., Normand S., & Kageyama M. (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology*, **96**, 1117–1127.

Syfert M.M., Smith M.J., & Coomes D.A. (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PloS one*, **8**, e55158.

Tessarolo G., Rangel T.F., Araújo M.B., & Hortal J. (2014) Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, **20**, 1258–1269.

Thuiller W., Georges D., & Engler R. (2014) biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64.

Thuiller W., Lafourcade B., Engler R., & Araújo M.B. (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Varela S., Anderson R.P., García-Valdés R., & Fernández-González F. (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, **37**, 1084–1091.

Varela S., Lima-Ribeiro M.S., & Terribile L.C. (2015) A Short Guide to the Climatic Variables of the Last Glacial Maximum for Biogeographers. *PloS one*, **10**, e0129037.

Veloz S.D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, **36**, 2290–2299.

Wang T., Liu Y., & Huang W. (2013) Last glacial maximum sea surface temperatures: A model-

data comparison. *Atmospheric and Oceanic Science Letters*, **6**, 233–239.

Zaniewski A.E., Lehmann A., & McC Overton J. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. **157**, 261–280.

Appendix A {Tables A.1, A.2 and A.3, and Figures A.1, A.2, A.3, A.4 and A.5}

570 **Table 1** Comparison among mean AUC, sensitivity (sens.) and specificity (spec.) values of the models obtained using the complete set of 18 predictors (environmental variables +
571 landscape metrics), the subset of 9 variables available for current and LGM conditions, and just the set of 2 variables (minimum SST of winter and maximum SST of summer) to
572 predict the distribution of *Cymodocea nodosa*. The best scores are highlighted in bold. The last six columns show the change in validation scores (AUC, sensitivity and specificity)
573 when using just nine or two variables in relation to the complete set. The significance of this comparison is expressed as p-values of Wilcoxon test (** p< 0.001; * p<0.05; n.s. = no
574 significant). (s.d.= standard deviation).

575

	Complete set (18 variables)			Subset (9 variables)			SST set (2 variables)			Subset minus complete set			SST set minus complete set		
Model	AUC (± s.d.)	Sens. (± s.d.)	Spec. (± s.d.)	AUC (± s.d.)	Sens. (± s.d.)	Spec. (± s.d.)	AUC (± s.d.)	Sens. (± s.d.)	Spec. (± s.d.)	AUC	Sens.	Spec.	AUC	Sens.	Spec.
GLM	0.839 (±0.01)	76.939 (±4.37)	78.630 (±4.16)	0.788 (±0.03)	**81.639** (±3.77)	66.397 (±5.74)	0.691 (±0.06)	**86.011** (±3.34)	52.546 (±6.81)	-0.051 (**)	4.699 (*)	-12.233 (**)	-0.147 (**)	9.071 (**)	-26.083 (**)
GBM	0.874 (±0.01)	79.398 (±4.41)	80.501 (±3.46)	0.837 (±0.02)	71.912 (±4.93)	**81.358** (±4.86)	**0.838** (±0.02)	77.978 (±4.09)	77.123 (±3.55)	-0.036 (**)	-7.486 (*)	0.857 (n.s.)	-0.035 (**)	-1.420 (n.s.)	-3.377 (n.s.)
GAM	**0.957** (±0.01)	**91.858** (±1.67)	**89.450** (±1.24)	0.814 (±0.01)	79.945 (±4.51)	72.326 (±4.83)	0.757 (±0.01)	81.366 (±4.20)	62.777 (±2.44)	-0.142 (**)	-11.912 (**)	-17.123 (**)	-0.199 (**)	-10.491 (**)	-26.672 (**)
FDA	0.836 (±0.02)	76.885 (±5.72)	76.650 (±5.39)	0.810 (±0.02)	78.142 (±7.91)	71.699 (±7.90)	0.804 (±0.02)	77.814 (±4.70)	70.797 (±6.86)	-0.025 (**)	1.256 (n.s.)	-4.950 (n.s.)	-0.031 (**)	0.929 (n.s.)	-5.852 (*)
MARS	0.843 (±0.02)	78.525 (±3.82)	79.532 (±4.59)	0.805 (±0.02)	78.852 (±4.96)	72.211 (±5.90)	0.789 (±0.02)	79.726 (±3.24)	69.928 (±4.58)	-0.037 (**)	0.327 (n.s.)	-7.321 (**)	-0.054 (**)	1.202 (n.s.)	-9.604 (**)
RF	0.891 (±0.01)	80.819 (±2.64)	82.596 (±2.39)	**0.851** (±0.01)	76.284 (±2.99)	80.786 (±3.61)	0.823 (±0.01)	72.076 (±4.63)	**81.743** (±3.41)	-0.040 (**)	-4.535 (*)	-1.809 (n.s.)	-0.067 (**)	-8.743 (**)	-0.852 (n.s.)

576
577

Table 2 Pearson's correlation coefficients among the SDMs predictions (probabilities of presence) obtained for present and Last Glacial Maximum (LGM) conditions according to the different Ocean General Circulation Models and filters used to reduce sampling bias.

		No filter	GEO	GEO _WEIGHT	ENV	ENV _WEIGHT	WEIGHT	No filter _9var	WEIGHT _9var
Present	No filter	1							
	GEO	0.98	1						
	GEO_WEIGHT	0.98	1	1					
	ENV	0.97	0.99	0.99	1				
	ENV_WEIGHT	0.97	0.99	0.99	1	1			
	WEIGHT	1	0.98	0.98	0.97	0.97	1		
	No filter_9var	0.82	0.79	0.79	0.79	0.79	0.82	1	
	WEIGHT_9var	0.81	0.79	0.79	0.78	0.78	0.81	1	1
LGM CCSM	No filter	1							
	GEO	0.95	1						
	GEO_WEIGHT	0.95	1	1					
	ENV	0.96	0.92	0.92	1				
	ENV_WEIGHT	0.96	0.92	0.92	1	1			
	WEIGHT	1	0.95	0.95	0.97	0.97	1		
	No filter_9var	0.81	0.78	0.78	0.85	0.85	0.82	1	
	WEIGHT_9var	0.89	0.83	0.83	0.89	0.89	0.89	0.93	1
LGM CNRM	No filter	1							
	GEO	0.96	1						
	GEO_WEIGHT	0.96	1	1					
	ENV	0.98	0.97	0.98	1				
	ENV_WEIGHT	0.97	0.97	0.98	1	1			
	WEIGHT	1	0.96	0.97	0.98	0.98	1		
	No filter_9var	0.73	0.77	0.78	0.76	0.77	0.74	1	
	WEIGHT_9var	0.7	0.74	0.74	0.72	0.71	0.71	0.92	1
LGM Ensemble	No filter	1							
	GEO	0.95	1						
	GEO_WEIGHT	0.96	1	1					
	ENV	0.97	0.93	0.94	1				
	ENV_WEIGHT	0.97	0.92	0.93	1	1			
	WEIGHT	1	0.95	0.95	0.98	0.97	1		
	No filter_9var	0.87	0.87	0.88	0.85	0.85	0.87	1	
	WEIGHT_9var	0.8	0.82	0.81	0.77	0.76	0.79	0.93	1

583 **FIGURE LEGENDS**

584

585 **Fig. 1** Occurrence records of *Cymodocea nodosa* in its entire range of distribution (blue circles) and
586 location of populations analyzed genetically and genetic regions identified by Alberto et al. (2008)
587 (AH: high-latitude Atlantic; AL: low-latitude Atlantic; EM: Eastern Mediterranean; WM: Western
588 Mediterranean**)**. Mean values of allelic richness (A) and expected heterozygosity (He) for each
589 region, and the assignment of individuals to genetic clusters are also shown. Adapted from Alberto
590 et al. (2008).

591

592 **Fig. 2** Scheme showing the procedure used to test the uncertainties affecting the predictions of a
593 coastal species distribution. We assessed the effect of the number of variables and sampling bias on
594 the niche modelling of *Cymodocea nodosa*, a subtidal seagrass, by using the combinations of
595 variables and techniques shown. We tested the filters for occurrences (GEO and ENV) and the
596 background data filter (WEIGHT) against random sets and compared their performance. We used
597 sets of variables from the complete set: "Two SSTs" (minimum SST of winter and maximum SST
598 of summer), and "9var" (a subset of nine variables: minimum SST of winter, maximum SST of
599 summer, salinity, ECON_MN, FRAC_AM, PARA_MN, PLAND, SHAPE_MN and TECI).

600

601 **Fig. 3** Weighted sampling probability filter (WEIGHT) used to give a statistical weight to the
602 background data used in MAXENT models according to the probability of having been sampled for
603 each country. These probabilities were estimated using a set of rules including variables about
604 GBIF publishing activity (for Alismatales order, and in general) and country development. A
605 weight of 1 represents "very high" reliability of absences, while a weight of 0 represents "very
606 low".

607

608 **Fig. 4** Violin plot of the distribution of probabilities of presence in the three sets of variables using

609 the ensemble committee averaging of Generalized Additive models (GAMs). All: Complete set of
610 18 variables; Subset: set of 9 predictors; and SSTs: minimum sea surface temperature (SST) of
611 winter and maximum SST of summer. The dark blue bar represents the interquartile range and the
612 white dot represents the median.
613
614 **Fig. 5** Mean AUC measures obtained by filtered and non filtered environmental niche models of
615 *Cymodocea nodosa* using different thresholds for validation. No agreement among thresholds on
616 the best filtering option was found. (Env: environmental filter; geo: geographic filter; weight:
617 weighted background data filter according to political boundaries; 9var: subset of nine variables).
618
619 **Fig. 6** Stacked chart illustrating comparisons among habitat area predicted as having higher
620 probability of occurrence after a binary transformation of the models produced using different
621 filters and thresholds ("No omission", "Prevalence" and "Spec_sens") for each genetic cluster
622 (regions; Alberto et al. (2008). AH: high-latitude Atlantic; AL: low-latitude Atlantic; EM: Eastern
623 Mediterranean; WM: Western Mediterranean).
624
625 **Fig. 7** Likelihood of presence of *Cymodocea nodosa* according to the projection to the Last Glacial
626 Maximum climate modelled using the Ocean General Circulation Models CNRM-CM5, CCSM4
627 and an ensemble. In all models the glacial refugia found are coincident with the populations with
628 higher genetic diversity (Alberto et al. 2008) delimiting the low latitude Atlantic region (AL) and
629 the Eastern Mediterranean region (EM), those with higher sea surface temperature during
630 glaciations. But just CCSM model would be congruent with a vicariant hypothesis. Directional
631 genetic flows (grey arrows) identified by Masucci et al. (2012) are also congruent explaining how a
632 post-glacial recolonization could have happened.
633
634