1 **ORIGINAL ARTICLE**

2

3 **Delimiting the geographical background in species distribution modelling**

4

5 Pelayo Acevedo^{1,2,3*}, Alberto Jiménez-Valverde¹, Jorge M. Lobo⁴ and Raimundo Real¹

6

7 1 - Biogeography, Diversity, and Conservation Research Team, Faculty of Sciences,
8 University of Málaga, E-29071 Málaga, Spain.

9 2 - CIBIO/UP – Centro de Investigação em Biodiversidade e Recursos Genéticos da
10 Universidade do Porto, Campus Agrario de Vairão, 4485-661 Vairão, Portugal.

11 3 - Instituto de Investigación en Recursos Cinegéticos (CSIC-UCLM-JCCM), E-13071
12 Ciudad Real, Spain.

13 4 – Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias
14 Naturales, E-28006 Madrid, Spain.

15

16 * Correspondence: Pelayo Acevedo, Biogeography, Diversity, and Conservation
17 Research Team, Faculty of Sciences, University of Málaga, E-29071 Málaga, Spain.

18 E-mail: pacevedo@uma.es

19

20 **Running title:** The geographical background in species distribution modelling

21

22

23 **ABSTRACT**

24 **Aim** The extent of the study area (geographical background, GB) can strongly affect the
25 results of species distribution models (SDMs), but as yet we lack objective and
26 practicable criteria for delimiting the appropriate GB. We propose an approach to this
27 problem using trend surface analysis (TSA) and provide an assessment of the effects of
28 varying the GB extent on the performance of SDMs for four species.

29 **Location** Mainland Spain.

30 **Methods** Using data for four well-known wild ungulate species and different GBs
31 delimited with a TSA, we assessed the effects of the GB extent on the predictive
32 performance of the SDMs; specifically, on model calibration (Miller's statistic) and
33 discrimination (AUC, sensitivity and specificity), and on the tendency of the models to
34 predict environmental potential when they are projected beyond their training area.

35 **Results** In the training area, discrimination significantly increased and calibration
36 decreased as the GB was enlarged. In contrast, as GB was enlarged, both discriminatory
37 power and calibration decreased when assessed in the core area of the species
38 distributions. When models trained using small GBs were projected beyond their
39 training area, they showed a tendency to predict higher environmental potential for the
40 species than those models trained using large GBs.

41 **Main conclusions** By restricting the GB extent using a geographical criterion, model
42 performance in the core area of the species distribution can be significantly improved.
43 Large GBs make models demonstrate high discriminatory power but are barely
44 informative. By delimiting GB using a geographical criterion, the effect of historical
45 events on model parameterization may be reduced. Thus, purely environmental models
46 are obtained which, when projected onto a new scenario, depict the potential

47 distribution of the species. We therefore recommend the use of TSA in geographically
48 delimiting the GB for use in SDMs.

49

50 **Keywords** Calibration, discrimination, environmental potential, extent,
51 geographical background, historical factors, Spain, species distribution models,
52 trend surface analysis, ungulate distributions.

53

54 INTRODUCTION

55 Recent studies have shown that the extent of the study area – or geographical
56 background (GB) – in species distribution modelling (SDM) has a strong effect on the
57 parameterization and evaluation of the models (Barve *et al.*, 2011). If the GB is too
58 small to fully represent the ranges of the species, then the importance of coarse-scale
59 factors such as climate may be underestimated when one delimits the species
60 distribution (Jiménez-Valverde *et al.*, 2011a; Sánchez-Fernández *et al.*, 2011). On the
61 other hand, if the GB is very large then the ability of the models to tease out the fine-
62 scale conditions that actually determine species distribution will be limited (Lobo *et al.*,
63 2010). VanDerWal *et al.* (2009) showed that as the GB extent decreases, so does the
64 number of variables included in the models, which in turn affects the predicted spatial
65 patterns.

66
67 The effects of the extent of the GB on the discriminatory power of a model, i.e. the
68 effectiveness of the scoring rule (suitability value in a broad sense) in separating
69 instances of presence from those of absence, are noteworthy (Lobo *et al.*, 2008; Barve *et al.*,
70 2011). Higher and more significant discriminatory values can be obtained simply by
71 increasing the GB extent such that the number of uninhabited and unsuitable localities
72 under consideration increases. In this way, it is easy to obtain models with high
73 discriminatory power but with little informational content (Jiménez-Valverde *et al.*,
74 2008).

75
76 Anderson & Raza (2010), working with sister species, demonstrated the effects of the
77 GB extent on model transferability, and discussed their results within a potential versus
78 realized distribution framework (see Jiménez-Valverde *et al.*, 2008). They argued that if

79 unoccupied but environmentally suitable areas for the species are considered for model
80 training, then the capacity to predict the species' potential distribution will be reduced.
81 On the other hand, if the models are trained using a small area in which the species may
82 have a high probability of being at equilibrium with the environment, then the models
83 will be able to identify other potential occurrence areas when transferred. Barve *et al.*
84 (2011) went a step further and argued that the appropriate GB for model training,
85 validation, and comparison should comprise the set of localities that a species has
86 'sampled' over its history, i.e. 'the parts of the world that have been accessible to the
87 species via dispersal over relevant periods of time' (Barve *et al.*, 2011, p. 1811). This
88 accessible area is called 'M' in the Biotic, Abiotic and Movement (BAM) diagram
89 terminology (*sensu* Soberón & Peterson, 2005; see also Barve *et al.*, 2011) and it is
90 important to realize that it is specific to each species. Both Anderson & Raza (2010) and
91 Barve *et al.* (2011) recognized that delimiting the appropriate GB is generally not
92 feasible because the biological information required for this purpose is rarely available
93 for most species.

94

95 Barve *et al.* (2011) discussed several methods for delimiting the proper GB for an SDM
96 analysis. They suggested that the most workable approach could be to use biotic
97 regions, i.e. climatic and geographic units with organisms that share broad
98 environmental adaptations and history. Another method would be to use SDMs in a
99 two-step procedure: using the results of a first round of modelling to help define the
100 appropriate GB to be considered in a second round. Finally, they noted that the most
101 interesting but also the most challenging approach would be to use information related
102 to the dispersal capacity of the species, phylogeographic data and palaeoclimatic data.
103 However, Barve *et al.* (2011) recognized the excessive simplicity, the circularity, and

104 the lack of operability, respectively, of their proposals. In this study, we propose and
105 assess a species-specific, practicable procedure to delimit the GB based on the global
106 surface-fitting procedure known as trend surface analysis (TSA; Legendre & Legendre,
107 1998).

108

109 We argue that to develop purely environmental models in SDM (i.e. the so-called
110 ecological niche models *sensu* Soberón, 2010), a GB which maximizes the likelihood
111 that the targeted species is interacting with the environment should first be delimited.
112 This can be done by controlling for the broad-scale geographical structure of the data
113 that may be caused by numerous factors such as dispersal limitation, geographical
114 characteristics or historical events, among others (McGlone, 1996; Soberón & Peterson,
115 2005; Svenning & Skov, 2005). As the present distributional range of a species is
116 determined by its past distribution and population dynamics, the geographical universe
117 delimited with the TSA may be considered a reflection of the history of the ecological
118 interactions of the species (e.g., Real *et al.*, 2003). With the TSA, we can delimit the
119 area that has the highest probability of being accessible to a species given its present
120 distributional pattern and at the same time avoid the inclusion of geographical regions
121 that, due to their spatial remoteness, may be uninformative for an ecological model
122 (Lobo *et al.*, 2010). By accounting for the broad-scale spatially structured variation of
123 species occurrence data, the GB on which SDMs should be trained is defined. Because
124 SDMs are eco-geographical, once the broad-scale geographical structure has been
125 accounted for, the models parameterized within the GB (delimited with the TSA) can be
126 considered to be largely environmental, and these are the models that can be projected
127 onto new territories to identify favourable locations for the species.

128

129 In this context, our main objective is to propose an approach to delimiting the GB by
 130 using TSA to objectively identify the area within which SDMs should be built. Based
 131 on this approach, we also assess the effects of the GB extent on the predictive
 132 performance of the SDMs, and specifically on model calibration and discrimination. To
 133 the best of our knowledge, the effects of the GB extent on model calibration have not
 134 been evaluated. However, this is not surprising because calibration, i.e. how closely the
 135 predicted probabilities match the observed proportions of occurrence (Pearce & Ferrier,
 136 2000), is rarely assessed in SDM. We also evaluate the effects of the GB extent on the
 137 tendency of the models to predict environmental potential when they were projected
 138 onto a new scenario. To this end, we modelled the distribution of four mammal species
 139 with well known and contrasting distribution patterns in mainland Spain.

140

141 MATERIALS AND METHODS

142 The data

143 The study area was mainland Spain, an environmentally heterogeneous territory with a
 144 complex geological history (Font, 2000; Hevia, 2004). For modelling purposes we used
 145 the Universal Transverse Mercator (UTM) 10 km $\times$ 10 km squares as territorial units (n
 146 = 4684 squares in the study area). We modelled the distribution of four well-known
 147 native species (see Fig. 1): red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*),
 148 Iberian wild goat (*Capra pyrenaica*), and Pyrenean chamois (*Rupicapra pyrenaica*).
 149 The red deer is a common species ($n = 1530$ presences), and is widely distributed
 150 throughout the study area. The roe deer is also a common species ($n = 1782$ presences)
 151 in the northern half of Spain. The Iberian wild goat is mainly distributed in the eastern
 152 mountain ranges ($n = 621$ presences). Finally, the Pyrenean chamois has a very limited
 153 distribution ($n = 173$ presences) restricted to the northern mountain ranges. Presence

154 data for these species were extracted from Palomo *et al.* (2007), and information on the
155 Iberian wild goat was updated using data from Acevedo & Cassinello (2009). The rate
156 of false absence data can be considered negligible.

157

158 **Delimiting GB**

159 A third-degree TSA was fitted, as this is recommended for exploring processes that
160 occur at the same or a higher spatial scale than the study area (Legendre & Legendre,
161 1998; p. 742). The saturated functions of TSA, i.e., without the selection of predictors,
162 were used to obtain comparable models for the different species. For each species,
163 seven GBs of different extent were delimited using the TSA predicted values. The first
164 extent was delimited by the lowest TSA value assigned to a presence (GB_{LOW}); the
165 reasoning behind this GB is that it seems logical to select a GB that includes all the
166 presence records currently known for the species. Next, the GB was restricted by
167 selecting as thresholds the TSA values which correspond to excluding 1% (GB₋₁), 5%
168 (GB₋₅) and 10% (GB₋₁₀) of the presences with the lowest TSA values. Similarly, the
169 extent was enlarged including 1% (GB₊₁), 5% (GB₊₅) and 10% (GB₊₁₀) of the absences
170 that had the highest TSA values lower than the values for any presence. Finally, the
171 total study area (mainland Spain) was also included as an additional extent (GB_{MS}). In
172 summary, eight GBs of different extents were considered for each species and these
173 GBs were each used to assess the effects of the GB extent on the performance of the
174 models.

175

176 **Species distribution models**

177 Logistic regressions were performed for each species and criterion ($n = 32$ models) with
178 28 environmental predictors related to topography (2 variables), climate (22 variables),

179 human activity (3 variables), and lithology (1 variable; see Table 1). Variables were
180 chosen on the basis of availability and potential predictive power for wild ungulates in
181 Spain (see Acevedo & Real, 2011). As investigating the environmental factors that
182 modulate species distribution was not the aim of this study, we have not described the
183 variables further; details can be found in Barbosa *et al.* (2003).

184

185 Logistic models were forward–backward stepwise fitted using a 0.05 significance
186 threshold for the inclusion of the variables and 0.10 for their exclusion. Models were
187 trained on each extent (eight different GBs) and projected onto mainland Spain for each
188 species. To compare the results of each model obtained from species with different
189 prevalences, the favourability function was applied to convert logistic probabilities (P)
190 into favourability values (F) that are independent of sample prevalence (for further
191 details about this function see Real *et al.*, 2006).

192

193 **Training and evaluation data sets**

194 For each species and GB, a distribution model was trained using a 70% random sample
195 of the data. The predictive performance of the models was assessed on three evaluation
196 data sets: (1) on independent data and within the GB considered in the training process,
197 i.e., on the remaining 30% of the data (evaluation in the training area); (2) only on the
198 independent data that are included within GB₋₁₀ (evaluation in the core area –
199 independent data); and (3) in order to avoid problems because of a small sample size in
200 the previous evaluation data set, models were also evaluated using all the localities
201 included in GB₋₁₀ for each species (evaluation in the core area – full data). Different
202 evaluation data sets were selected to analyse the effects of GB extent in different
203 contexts of the distribution of a species (core area in relation to the complete training

204 area). The ‘core area – full data’ data set was constant across GBs for each species and
205 therefore provided a way of comparing performance in a quasi-standardized manner
206 between all models for each species.

207

208 **Predictive performance**

209 Sensitivity (Se , the ratio of correctly predicted presences to the total number of
210 presences), specificity (Sp , the ratio of correctly predicted absences to the total number
211 of absences), and the AUC (area under the curve of the receiver operating characteristic
212 plot) were computed to assess the discriminatory power of the models on each
213 evaluation data set. Se and Sp were calculated using a cut-off of $F = 0.5$ according to the
214 favourability concept (Real *et al.*, 2006). Calibration of the P -values was assessed using
215 Miller’s statistic, which is based on the hypothesis that the calibration line – perfect
216 calibration – has an intercept of zero and a slope of one (for details see Miller *et al.*,
217 1991; Pearce & Ferrier, 2000). The R script provided by Wintle *et al.* (2005) was used
218 for calculating Miller’s statistic. Finally, for each species and GB, the number of
219 territorial units predicted as presences ($F > 0.5$) in the whole study area was calculated
220 as a proxy of the environmental potential predicted by models.

221

222 **Assessing the effect of the GB extent**

223 The effects of the GB extent on each of the performance measures and evaluation data
224 sets were assessed using general linear mixed models because performance measures
225 are not independent (Zuur *et al.*, 2009). The species was included as a random factor
226 and the GB extent – measured as number of territorial units – as a covariable. The
227 normality of the residuals of each model was determined using the Kolmogorov–

228 Smirnov test (Zar, 1999). All models were assessed using the statistical package SPSS
229 18.

230

231 **RESULTS**

232 The results of TSA provided evidence of broad-scale spatial trends in the distribution of
233 the four species (Fig. 2). The favourability maps obtained from the 32 models are
234 shown in Appendix S1 in the Supporting Information; the case of the roe deer is
235 presented as an example (Fig. 3). In general, visual assessment of the geographical
236 patterns shows that the predictions of the models for each species are quite similar in the
237 core area (GB₋₁₀), and that the highest variability between the different models is
238 obtained when making predictions outside the training data sets (see Appendix S1).

239

240 The results of the statistical models used to assess the effect of the GB extent on the
241 measures of model performance are summarized in Table 2 (statistical parameters are
242 given in Appendix S2). The residuals of each model were normally distributed ($P >$
243 0.05 in all cases). In most cases, the GB extent was significantly associated with the
244 four performance measures. There was a negative association between the
245 discriminatory power – AUC and Sp – and the GB extent when models were assessed in
246 the core areas and a positive association when the evaluation was performed on the
247 training area. The relationship with Se was positive in all cases, although it was not
248 always significant. Miller’s statistic, in which high values indicate poorly calibrated
249 models, was positively associated with the GB extent when the models were assessed
250 on both the training and the core area data sets. Finally, there was a negative association
251 between the GB extent and the area predicted as suitable for mainland Spain ($F_{1,27} =$
252 6.023, $P = 0.021$; species was included as a random factor: $F_{1,27} = 62.022$, $P < 0.001$).

253 In summary, as the GB extent increases, the discriminatory power within the considered
254 GB improves but on closer inspection, when only the performance in the core area is
255 assessed, the discriminatory capacity worsens due to overprediction. Calibration is
256 always negatively affected by increasing the GB extent. Furthermore, when the models
257 are projected beyond their training area, the smaller the GB extent, the higher the
258 capacity to predict environmental potential becomes (Fig. 3; see also Appendix S1).

259

260 **DISCUSSION**

261 Our results demonstrate that the GB selected has visible effects on the parameters used
262 to measure the predictive performance of SDMs, namely model discrimination (Lobo *et al.*
263 *et al.*, 2008; VanDerWal *et al.*, 2009; Barve *et al.*, 2011), calibration, and the model's
264 capacity to predict environmental potential (Anderson & Raza, 2010). Unfortunately,
265 the GB has usually been defined using geopolitical criteria with no real biological
266 justification (Meyer & Thuiller, 2006). Sufficient evidence has now been accumulated
267 showing the effects of GB on SDMs, and steps to delimit it are beginning to be
268 contemplated; the approach that is proposed in this study is a practical and objective
269 way to do so.

270

271 The inclusion of absences from beyond the geographical domain of the species, i.e.,
272 increasing GB, is an easy way to obtain models with a high capacity to discriminate
273 between instances of presence and instances of absence (Lobo *et al.*, 2008; Barve *et al.*,
274 2011). This is corroborated by the positive association between the GB extent and the
275 discrimination measures obtained when the models were evaluated on the training area
276 data sets. However, discriminatory power decreased as GB increased when assessed on
277 the core area of the species distribution (i.e. GB₋₁₀, the minimum extent). On the other

278 hand, calibration improved when models were built using smaller GBs. In other words,
279 if absences from beyond the geographical domain of the species are included, then the
280 models will not effectively reflect the probability of presence. In summary, increasing
281 GB produces apparently better (in terms of discrimination) but barely informative
282 models (see also Lobo *et al.*, 2010).

283
284 Our results indicate that larger favourable areas in mainland Spain were predicted when
285 using smaller GBs (see Fig. 3 and Appendix S1). Although there is no objective way to
286 assess the accuracy of the estimations of environmental potential, these were consistent
287 with expert opinions for the studied species. Anderson & Raza (2010) found a similar
288 pattern and explained that using a large GB in SDM could make the models prone to
289 overfit the environmental conditions present in the region occupied by the species. This
290 may happen because the algorithm recognizes spurious environmental differences
291 between the inhabited localities and localities that could be inhabited but are not, which
292 may be due, for example, to barriers preventing species dispersion or other historical
293 events restricting the species current distribution. By delimiting the GB using
294 geographical criteria, we may be excluding – or at least minimizing – the effect of
295 historical events on model parameterization. Thus, we may be able to obtain
296 environmental models which, when projected onto a new scenario, may help to depict
297 the potential distribution of the species more reliably. The extrapolation of models is
298 risky and requires caution and careful consideration (Jiménez-Valverde *et al.*, 2011b).
299 For instance, it is necessary to highlight the areas that have environmental values that
300 are beyond those shown in the training region, because the predictions there are more
301 uncertain (Elith *et al.*, 2010; Jiménez-Valverde *et al.*, 2011c). It is also advisable to
302 check for maintenance of the correlation structure among the independent variables in

303 the new geographical area with respect to the training region (see Jiménez-Valverde *et*
304 *al.*, 2011c). It is also interesting to highlight that the patterns obtained in this study
305 using presence–absence data and logistic regression follow the same trend as those
306 obtained by other authors using presence–background data and MAXENT (Anderson &
307 Raza, 2010; Barve *et al.*, 2011), which suggests that they do not depend on the
308 modelling technique.

309

310 To our knowledge, only Barve *et al.* (2011) have previously presented a framework for
311 thinking about and estimating the GB in the context of SDM; they suggested several
312 potential approaches to objectively delimit ‘M’ (see Introduction). The most promising
313 approach would be to use information related to the dispersal capacity and history of the
314 species, but the data required are rarely available. More feasible procedures such as the
315 use of biogeographical regions overlook the species-specificity of the GB and may not
316 be entirely satisfactory. In this study, we propose a simple but practical and species-
317 specific way to delimit GB using purely geographical criteria. TSA is a simple method
318 that accounts for broad-scale spatial structures and shows the main geographical trends
319 in the data (see Legendre & Legendre, 1998). Thus, we argue that TSA is a useful
320 method for use in delimiting the area in order to maximize the likelihood that the target
321 species is currently interacting with the environment. At the same time, it minimizes the
322 probability of including regions that are suitable for the species but that are
323 uninformative for an ecological model due to their spatial remoteness from the current
324 geographical range (see Lobo *et al.*, 2010). The TSA should be considered a working
325 procedure intended to minimize the role played by the factors that operate beyond the
326 area inhabited by the species. Strictly speaking, the spatial pattern generated with the
327 TSA cannot be considered a geographical representation of ‘M’ because the concepts

328 that underlie each approximation do not necessarily converge on the same geographical
329 space. Most likely, the longer the species has been present in the accessible area, the
330 lower the similarity between the spatial patterns yielded by the two approximations will
331 be. Nevertheless, under such extreme circumstances (long time periods), estimating ‘M’
332 is very difficult if not impossible. Whether ‘sampled’ unoccupied localities that are far
333 away from the present distribution range should be considered in the modelling process,
334 or should be excluded because they are not informative about the interaction of the
335 species with the environment (see Lobo *et al.*, 2010), is debatable. We show that TSA is
336 a practical approach that can be used as a reference for future studies aimed at
337 developing new ideas in delimiting GB. We also anticipate that other spatial pattern
338 analytical procedures may merit future investigation, and that the delimitation of the GB
339 is a promising line of research and debate.

340

341 **ACKNOWLEDGEMENTS**

342 This study was presented at the International Biogeography Society Crete meeting. We
343 are grateful to A. Townsend Peterson and one anonymous referee for their useful
344 comments and suggestions on previous versions of the manuscript. Lucía Maltez kindly
345 reviewed the English. P.A. and A.J.-V. were supported by the Juan de la Cierva
346 research program awarded by the Ministerio de Ciencia e Innovación–Fondo Social
347 Europeo, and partly by the project CGL2009-11316/BOS—Fondos FEDER. P.A. is in
348 Portugal thanks to a José Castillejo fellowship (2010–2011) granted by the Ministerio
349 de Ciencia e Innovación.

350

351

352 **REFERENCES**

- 353 Acevedo, P. & Cassinello, J. (2009) Biology, ecology and status of Iberian ibex *Capra*
354 *pyrenaica*: a critical review and research prospectus. *Mammal Review*, **39**,17–
355 32.
- 356 Acevedo, P. & Real, R. (2011) Biogeographical differences between the two *Capra*
357 *pyrenaica* subspecies, *C. p. victoriae* and *C. p. hispanica*, inhabiting the
358 Iberian Peninsula: implications for conservation. *Ecological Modelling*, **222**,
359 814–823.
- 360 Anderson, R.P. & Raza, A. (2010) The effect of the extent of the study region on GIS
361 models of species geographic distributions and estimates of niche evolution:
362 preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela.
363 *Journal of Biogeography*, **37**, 1378–1393.
- 364 Barbosa, A.M., Real, R., Olivero, J. & Vargas, J.M. (2003) Otter (*Lutra lutra*)
365 distribution modeling at two resolution scales suited to conservation planning
366 in the Iberian Peninsula. *Biological Conservation*, **114**, 377–387.
- 367 Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson,
368 A.T., Soberón, J. & Villalobos, F. (2011) The crucial role of the accessible area
369 in ecological niche modeling and species distribution modeling. *Ecological*
370 *Modelling*, **222**, 1810–1819.
- 371 Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species.
372 *Methods in Ecology and Evolution*, **1**, 330–342.
- 373 Font, I. (1983) *Atlas climático de España*. Instituto Nacional de Meteorología, Madrid.
- 374 Font, I. (2000) *Climatología de España y Portugal*. Ediciones Universidad de
375 Salamanca, Salamanca.

- 376 Hevia, I.M. (2004) *Geología de España. Una historia de seiscientos millones de años.*
377 Rueda S. L., Madrid.
- 378 IGME (1979) *Mapa hidrogeológico Nacional. Explicación de los mapas de lluvia útil,*
379 *de reconocimiento hidrogeológico y de síntesis de los sistemas acuíferos*, 2nd
380 ed. Instituto Geológico y Minero de España (IGME), Madrid.
- 381 IGN (1999) *Mapa de carreteras. Península Ibérica, Baleares y Canarias.* Instituto
382 Geográfico Nacional (IGN)/Ministerio de Fomento, Madrid.
- 383 Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the
384 importance of concepts in species distribution modelling. *Diversity and*
385 *Distributions*, **14**, 885–890.
- 386 Jiménez-Valverde, A., Lira-Noriega, A., Barve, N., Maher, S.P., Nakazawa, Y., Papes,
387 M., Soberón, J., Sukumaran, J. & Peterson, A.T. (2011a) Dominant climate
388 influences on North American birds. *Global Ecology and Biogeography*, **20**,
389 114–118.
- 390 Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J., Aragón, P. & Lobo, J.
391 M. (2011b) Use of niche models in invasive species risk assessments.
392 *Biological Invasions*, **13**, 2785–2797.
- 393 Jiménez-Valverde, A., Decae, A.E. & Arnedo, M.A. (2011c) Environmental suitability
394 of new reported localities of the funnelweb spider *Macrothele calpeiana*: an
395 assessment using potential distribution modelling with presence-only
396 techniques. *Journal of Biogeography*, **38**, 1213–1223.
- 397 Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier
398 Science, Amsterdam.

- 399 Lobo, J.M, Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the
400 performance of predictive distribution models. *Global Ecology and*
401 *Biogeography*, **17**, 145–151.
- 402 Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences
403 and their importance in species distribution modelling. *Ecography*, **33**, 103–
404 114.
- 405 McGlone, M. (1996) When history matters: scale, time, climate and tree diversity.
406 *Global Ecology and Biogeography*, **5**, 309–314.
- 407 Meyer, C.B. & Thuiller, W. (2006) Accuracy of resource selection functions across
408 spatial scales. *Diversity and Distributions*, **12**, 288–297.
- 409 Miller, M.E., Hui, S.L. & Tierney, W. (1991) Validation techniques for logistic
410 regression models. *Statistics in Medicine*, **10**, 1213–1226.
- 411 Montero de Burgos, J.L. & González-Rebollar, J.L. (1974) *Diagramas bioclimáticos*.
412 ICONA, Madrid.
- 413 Palomo, L.J., Gisbert, J. & Blanco, J.C. (2007) *Atlas y Libro Rojo de los mamíferos*
414 *terrestres de España*. Ministerio de Medio Ambiente, Madrid.
- 415 Pearce, J. & Ferrier, S. (2000) An evaluation of alternative algorithms for fitting species
416 distribution models using logistic regression. *Ecological Modelling*, **128**, 127–
417 147.
- 418 Real, R., Barbosa, A.M., Porras, D., Kin, M.S., Márquez, A.L., Guerrero, J.C., Palomo,
419 L.J., Justo, E.R. & Vargas, J.M. (2003) Relative importance of environment,
420 human activity and spatial situation in determining the distribution of terrestrial
421 mammal diversity in Argentina. *Journal of Biogeography*, **30**, 939–947.

- 422 Real, R., Barbosa, A.M. & Vargas, J.M. (2006) Obtaining environmental favourability
423 functions from logistic regression. *Environmental and Ecological Statistics*, **13**,
424 237–245.
- 425 Sánchez-Fernández, D., Lobo, J.M. & Hernández-Manrique, O.L. (2011) Species
426 distribution models that do not incorporate global data misrepresent potential
427 distributions: a case study using Iberian diving beetles. *Diversity and*
428 *Distributions*, **17**, 163–171.
- 429 Soberón, J. (2010) Niche and area of distribution modeling: a population ecology
430 perspective. *Ecography*, **33**, 159–167.
- 431 Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological
432 niches and species' distribution areas. *Biodiversity Informatics*, **2**, 1–10.
- 433 Svenning, J.C. & Skov, F. (2005) The relative roles of environment and history as
434 controls of tree species composition and richness in Europe. *Journal of*
435 *Biogeography*, **32**, 1019–1033.
- 436 VanDerWal, J., Shoo, L.P., Graham, C. & Williams, S.E. (2009) Selecting pseudo-
437 absence data for presence-only distribution modeling: how far should you stray
438 from what you know? *Ecological Modelling*, **220**, 589–594.
- 439 Wintle, B.A., Elith, J. & Potts, J.M. (2005) Fauna habitat modelling and mapping in an
440 urbanising environment: a case study in the Lower Hunter Central Coast region
441 of NSW. *Austral Ecology*, **30**, 719–738.
- 442 Zar, J.H. (1999) *Biostatistical analysis*, 5th edn. Pearson, New York.
- 443 Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. (2009) *Mixed effects*
444 *models and extensions in ecology with R*. Springer, New York.
- 445

446 SUPPORTING INFORMATION

447 Additional Supporting Information may be found in the online version of this article:

448

449 **Appendix S1** Favourability maps for each species and geographical background (GB)
450 extent.

451 **Appendix S2** Statistical parameters of the mixed models used to assess the effect of the
452 GB extent on the calibration and discrimination parameters.

453

454 As a service to our authors and readers, this journal provides supporting information
455 supplied by the authors. Such materials are peer-reviewed and may be re-organized for
456 online delivery, but are not copy-edited or typeset. Technical support issues arising
457 from supporting information (other than missing files) should be addressed to the
458 authors.

459

460 BIOSKETCH

461 **Pelayo Acevedo** is a researcher at the University of Málaga. His interests include the
462 study of factors affecting the distribution and abundance of pathogens and their hosts
463 and vectors, through fragmented habitats.

464

465 Author contributions: P.A. and R.R. conceived the ideas; P.A. and A.J.-V. analysed the
466 data; P.A., A.J.-V, J.M.L. and R.R participated in the discussion of the results and wrote
467 the manuscript.

468

469 Editor: Miles Silman

470

471 **Table 1** Variables used to model the distribution of wild ungulates.

472

Factor	Variable description
Orography	Mean elevation (m) *
	Mean slope (degrees; calculated from mean altitude)
Climatology	Mean annual precipitation (mm) [<i>P</i>] [†]
	Maximum precipitation in 24 h (mm) [<i>MP24</i>] [†]
	Relative maximum precipitation (= <i>MP24/P</i>)
	Mean annual number of days with precipitation ≥ 0.1 mm [†]
	Mean annual number of hail days [†]
	Mean annual number of foggy days [†]
	Inter-annual pluviometric irregularity [‡]
	Mean annual potential evapotranspiration (mm) [<i>PET</i>] [†]
	Mean annual actual evapotranspiration (mm) (= min [<i>P</i> , <i>PET</i>])
	Mean relative air humidity in July at 07:00 h (%) [<i>HJL</i>] [†]
	Mean relative air humidity in January at 07:00 h (%) [<i>HJN</i>] [†]
	Annual humidity range (%) (= <i>HJL</i> – <i>HJN</i>)
	Mean annual solar radiation (kW h m ^{–2} day ^{–1}) [†]
	Mean temperature in July (°C) [<i>TJL</i>] [†]
	Mean temperature in January (°C) [<i>TJN</i>] [†]
	Annual temperature range (°C) (= <i>TJL</i> – <i>TJN</i>)
	Mean annual temperature (°C) [†]
	Mean annual number of frost days (minimum temperature ≤ 0 °C) [†]
	Continentality index [¶]
	Humidity index [¶]
Human activity	Mean annual insolation (hours year ^{–1}) [†]
	Mean annual runoff (mm) [‡]
Lithology	Distance to the nearest town with more than 100,000 inhabitants (km) [§]
	Distance to the nearest town with more than 500,000 inhabitants (km) [§]
	Distance to the nearest highway (km) [§]
Lithology	Soil permeability ^{**}
Third-degree polynomial of the trend surface analysis	Mean latitude (°N) [<i>LA</i>] [§]
	Mean longitude (°E) [<i>LO</i>] [§]
	<i>LALO</i> = <i>LA</i> × <i>LO</i>
	<i>LOLA</i> ² = <i>LO</i> × <i>LA</i> ²
	<i>LO</i> ² <i>LA</i> = <i>LO</i> ² × <i>LA</i>
	<i>LA</i> ² = <i>LA</i> × <i>LA</i>
	<i>LO</i> ² = <i>LO</i> × <i>LO</i>

	$LA^3 = LA^2 \times LA$
	$LO^3 = LO^2 \times LO$

473 Sources: * <http://www.etsimo.uniovi.es/~feli/data/datos.html>. [†] Font (1983). [‡] Montero de Burgos &

474 González-Rebollar (1974). [¶] Font (2000). [§] IGN (1999); data on the number of inhabitants of urban

475 centres taken from the Instituto Nacional de Estadística (<http://www.ine.es>). ^{**} IGME (1979).

476

477 **Table 2** Summary of the results of the general linear mixed models used to assess the
 478 effect of the extent of the geographical background (GB) on the performance
 479 (calibration and discrimination) of species distribution models for four species in
 480 mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian
 481 wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Species
 482 was included as a random factor. The predictive performance of the models was
 483 evaluated on different data sets (see text for details). Statistical parameters are reported
 484 in the Supporting Information (Appendix S2).
 485

486

Evaluation data set	Parameter (dependent variable)	GB extent (covariable)	487
Training area	Miller’s statistic	(+)**	488
	Sensitivity	(+) ns	
	Specificity	(+)**	489
	AUC	(+)*	490
Core area – independent data	Miller’s statistic	(+)*	
	Sensitivity	(+) ns	491
	Specificity	(–)**	492
	AUC	(–) #	
Core area – full data	Miller’s statistic	(+)**	493
	Sensitivity	(+)**	
	Specificity	(–)**	494
	AUC	(–)*	495

496 ns, non-significant; #, $P < 0.08$; *, $P < 0.05$; **, $P < 0.01$.
 497

498 **FIGURE LEGENDS**

499 **Figure 1** Current distribution of the focus species in mainland Spain: red deer (*Cervus*
500 *elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and
501 Pyrenean chamois (*Rupicapra pyrenaica*). Presence data were referred to UTM 10 ×
502 10 km grid cells. These were taken from Palomo *et al.* (2007) and Acevedo &
503 Cassinello (2009).

504

505 **Figure 2** Results of the trend surface analysis (TSA) using a third-degree polynomial of
506 the spatial coordinates applied to the occurrence localities of the four species in
507 mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian
508 wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*).

509

510 **Figure 3** Species distribution models (favourability values) obtained using different
511 criteria to delimit the geographical background (GB) extent using roe deer (*Capreolus*
512 *capreolus*) as an example (see also Appendix S1). 'GB_{MS}' indicates the model that
513 included the complete study area (mainland Spain) as a training data set. 'GB_{LOW}'
514 indicates the model in which the training area was delimited by the lowest trend surface
515 analysis (TSA) value assigned to a presence (see text for details). 'GB₊₁₀' indicated the
516 model that included 10% of the absences that, having TSA values lower than any
517 presence, had the highest TSA values. Finally, 'GB₋₁₀' is similar to 'GB_{LOW}' but excludes
518 10% of the presences. The dashed line marks the area delimited with 'GB₋₁₀'.

SUPPORTING INFORMATION -- *Journal of Biogeography*

Delimiting the geographical background in species distribution modelling, by P. Acevedo, A. Jiménez-Valverde, J.M. Lobo and R. Real

Appendix S1 Species distribution models (favourability values) for each of the focus species: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). For each species, seven geographical backgrounds (GBs) of different extent were delimited. The first extent (GB_{LOW}) was delimited by the lowest trend surface analysis (TSA) value assigned to a presence. Next, the extent was constrained by selecting as thresholds the TSA values which correspond to excluding 1% (GB₋₁), 5% (GB₋₅) and 10% (GB₋₁₀) of the presences with the lowest TSA values. In a similar way, the extent was enlarged including 1% (GB₊₁), 5% (GB₊₅) and 10% (GB₊₁₀) of the absences that had the highest TSA values lower than that of any presence. Finally, we also included the complete study area (mainland Spain) as an additional GB (GB_{MS}).

Appendix S2 Statistical parameters of the mixed models carried out to assess the effect of the extent of the geographical background (GB) on the performance (calibration and discrimination) of the species distribution models for four species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Species was included as a random factor. The predictive performance of the models was evaluated on different data sets (see text for details).

Data set	Parameter (dependent variable)	Intercept	Species (random factor)	GB extent (covariable)
Training area	Miller's statistic	$F_{1,15.16} = 4.42, P = 0.043$	$F_{3,27} = 8.22, P < 0.001$	(+) $F_{1,27} = 15.11, P = 0.001$
	Sensitivity	$F_{1,5.19} = 342.70, P < 0.001$	$F_{3,27} = 38.77, P < 0.001$	(+) $F_{1,27} = 1.52, P = 0.230$
	Specificity	$F_{1,3.56} = 392.16, P < 0.001$	$F_{3,27} = 139.25, P < 0.001$	(+) $F_{1,27} = 11.81, P = 0.002$
	AUC	$F_{1,3.66} = 483.38, P < 0.001$	$F_{3,27} = 119.07, P < 0.001$	(+) $F_{1,27} = 7.12, P = 0.013$
Core area – independent data	Miller's statistic	$F_{1,20.72} = 9.24, P = 0.089$	$F_{3,27} = 5.38, P = 0.005$	(+) $F_{1,27} = 2.58, P = 0.012$
	Sensitivity	$F_{1,6.82} = 222.17, P < 0.001$	$F_{3,27} = 23.52, P < 0.001$	(+) $F_{1,27} = 1.63, P = 0.213$
	Specificity	$F_{1,14.04} = 536.90, P < 0.001$	$F_{3,27} = 9.04, P < 0.001$	(–) $F_{1,27} = 9.02, P = 0.006$
	AUC	$F_{1,3.91} = 706.30, P < 0.001$	$F_{3,27} = 87.20, P < 0.001$	(–) $F_{1,27} = 2.35, P = 0.077$
Core area – full data	Miller's statistic	$F_{1,7.90} = 3.77, P = 0.092$	$F_{3,27} = 18.89, P < 0.001$	(+) $F_{1,27} = 26.92, P < 0.001$
	Sensitivity	$F_{1,3.87} = 270.75, P < 0.001$	$F_{3,27} = 90.77, P < 0.001$	(+) $F_{1,27} = 14.40, P = 0.001$
	Specificity	$F_{1,17.51} = 1491.59, P < 0.001$	$F_{3,27} = 6.83, P = 0.001$	(–) $F_{1,27} = 45.75, P < 0.001$
	AUC	$F_{1,3.19} = 898.98, P < 0.001$	$F_{3,27} = 403.79, P < 0.001$	(–) $F_{1,27} = 5.56, P = 0.026$

Current distribution of the focus species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Presence data were referred to UTM 10 × 10 km grid cells. These were taken from Palomo et al. (2007) and Acevedo & Cassinello (2009).
165x136mm (300 × 300 DPI)

Results of the trend surface analysis (TSA) using a third-degree polynomial of the spatial coordinates applied to the occurrence localities of the four species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*).
196x192mm (300 x 300 DPI)

Species distribution models (favourability values) obtained using different criteria to delimit the geographical background (GB) extent using roe deer (*Capreolus capreolus*) as an example (see also Appendix S1). 'GB_{MS}' indicates the model that included the complete study area (mainland Spain) as a training data set. 'GB_{LOW}' indicates the model in which the training area was delimited by the lowest trend surface analysis (TSA) value assigned to a presence (see text for details). 'GB₊₁₀' indicated the model that included 10% of the absences that, having TSA values lower than any presence, had the highest TSA values. Finally, 'GB₋₁₀' is similar to 'GB_{LOW}' but excludes 10% of the presences. The dashed line marks the area delimited with 'GB₋₁₀'.
152x116mm (300 x 300 DPI)