

Economics of land use reveals a selection bias in tree species distribution models

Jean-Sauveur Ay^{*1,2}, Joannès Guillemot^{*2,3,4,5}, Nicolas Martin-StPaul^{*2,3,4,6},
Luc Doyen⁷ and Paul Leadley^{2,3,4}

*Equal contribution

⁸INRA, UMR 1041 CESAER, F-21079, Dijon, France.

⁹AgroParisTech, Paris, F-75231, France.

¹⁰Université Paris-Sud, Laboratoire ESE (UMR 8079), Orsay, F-91405, France.

¹¹CNRS, Orsay, F-91405, France

¹²CIRAD, UMR ECO&SOLS, F-34398 Montpellier, France.

¹³INRA, URFM, F-84914, Avignon, France.

¹⁴CNRS, GREThA, F-33600, Bordeaux, France.

Email addresses

J.S. Ay: jsay@diijon.inra.fr; J. guillemot: Joannes.guillemot@gmail.com; N. Martin-StPaul:

nicolas.martin@avignon.inra.fr; L. Doyen: luc.doyen@u-bordeaux.fr; P. Leadley: paul.leadley@u-psud.fr

Key-words: species distribution; forest ecology; multi-disciplinary; econometrics; land-use; niche modelling; selection bias.

Running title: Selection bias in species distribution models

Abstract: 293 words, **Main Text:** 5596 words, **Number of References:** 50, **Number of Tables:** 2, **Number of Figures:** 3.

Correspondance: Jean-Sauveur Ay, email: jsay@diijon.inra.fr, tel: +33(0)3.80.77.25.80 , fax: +33(0)3.77.25.71.

28Abstract

29 **Aims.** In human dominated ecosystems, the presence of a given species is the result of both
30the ecological suitability of the site and human impacts such as land-use choices. The influence of
31land-use choices on the predictions of species distribution models (SDMs) has however been often
32neglected. Here, we provide a theoretical analysis of the land-use selection bias affecting classical
33SDMs in the case of either presence-only or presence-absence datasets. Land-use selection bias in
34SDM predictions is then quantified for four widespread European tree species, using NFI data.

35 **Location.** Continental France

36 **Methods.** We describe a Bivariate Selection Model (BSM) that estimates simultaneously the
37economics of land-use choices and species responses to bioclimatic variables. The land-use equa-
38tion, based on an econometric model of landowner choices, is joined to an equation of species re-
39sponse to bioclimatic variables.

40 **Results.** We evidenced a significant land-use selection bias in all the species studied. The
41sign and the magnitude of the bias varied among species and were strongly related to the type of
42dataset used in the SDM calibration (presence-only or presence-absence). In addition, the BSM esti-
43mate the spatial covariance between the probability of presence and the presence of compatible
44land-use. We found that, depending on the species, sites with high ecological suitability could
45present a high probability of compatible land-use (positive covariance) or a low probability (nega-
46tive covariance).

47 **Main Conclusion.** We showed that using classical SDMs in human dominated areas can
48lead to strong miss-estimations of actual species' distributions and could therefore prevents sound
49projections of the effect of climate change. Beyond, the proposed BSM represents a crucial step to
50account for economic forces shaping species distribution in anthropized area and pave the way for a
51direct assessment of trade-offs and opportunities that may arise in a context of global change.

521. Introduction

53 Species distribution models (SDMs) have been increasingly used in a wide range of
54ecological studies, with the general aim of understanding and predicting the distributional area of
55species (Guisan & Thuiller 2005; Thuiller *et al.* 2005; Zimmermann *et al.* 2010; Engler *et al.* 2011;
56Franklin *et al.* 2013). The data used to calibrate or evaluate SDMs often consist in presence-only (P-
57O) data, i.e., a collection of point locations where a species is found, as absence data are difficult to
58obtain and require prohibitively expensive sampling designs. This constraint on data resources has
59been circumvented by using information on the range of the available environment conditions in the
60studied region (referred to as background data) instead of recorded absences (Elith *et al.* 2006).
61However, a fundamental limitation of SDM calibration based on P-O data is its sensitivity to the
62sample selection bias (Phillips *et al.* 2009) that can occur if the observations are not randomly
63distributed over a landscape. For instance predictions are likely to be affected by a sample selection
64bias if occurrences are preferably recorded near roads or towns (observer bias, Warton *et al.* 2013),
65or if sites expected to be suitable for a given species are over sampled (Jiménez-Valverde 2012;
66Smith 2013). In these cases, there is a risk that the recorded occurrences only partially describe the
67suitable species area of the studied landscape, which lead to wrong SDM predictions.

68 Because the sample selection bias affects both presence and absence records, its effect is
69usually thought to cancel out in the predictions of SDMs calibrated using presence-absence (P-A)
70data (Elith *et al.* 2011). But SDM predictions based on P-A data remain actually confronted to the
71general issue of selection bias, due to the uncertain nature of absence data (Jiménez-Valverde *et al.*
722008; Lobo *et al.* 2010; Duputié *et al.* 2014). Indeed, the absence of a species from any location
73cannot unambiguously be related to unsuitable environmental conditions as it can also result from
74internal processes related to the species ecology (biotic interactions, limited dispersion, habitat
75fragmentation) or from external processes driven by human activities (e.g., land-use choices,
76hunting intensity, resource management). In these cases, selection bias cannot be reduced by

77improving the sampling scheme or increasing the number of sampled sites of presences and/or
78absences. The biases related to the imperfect detection of potential distribution are therefore very
79common in all the species distribution studies, even in the case of sessile species such as plants
80(Chen *et al.* 2013) and could strongly impact SDM performances (Lahoz-Monfort *et al.* 2014).

81 The effect of land-use in SDM predictions has long been acknowledged. Indeed, SDMs are
82usually evaluated at continental scale and coarse resolution, at which climatic tolerance is a primary
83factor (Pearson & Dawson 2003; Thuiller *et al.* 2004). However, the degradation of habitats due to
84land-use changes remains a major cause of biodiversity loss (Bellard *et al.* 2012) and land-use
85choices could exert a strong control over species distribution evaluated at medium to fine spatial
86scales (1–10 km resolution, Pearson & Dawson 2003). Models integrating climate and land-use
87change components have been developed to produce sound projections of the future of species
88distribution (Brook *et al.* 2008). A few studies have integrated static (Pearson *et al.* 2004; Thuiller
89*et al.* 2004), or dynamic (Stanton *et al.* 2012; Martin *et al.* 2013) land-use variables along with
90climate variables in SDMs. All these studies came to the general conclusion that land-use effects on
91species distribution are largely over-ride by climatic effects. However, despite its potential
92importance, the effect of the selection bias created by land-use choices on SDM predictions has
93never been evaluated.

94 In this study, we use an econometric model of land-use choices to correct the land-use
95selection bias that may affect tree species response to bioclimatic variables in classical SDMs. We
96provide a joint ecological and economical framework for modelling tree species distribution, taking
97into account the potential – observed and unobserved – selection bias induced by anthropogenic
98land-use choices. We first provide a theoretical analysis of the land-use selection bias in SDMs,
99generalizing the two classical SDM practices of P-O and P-A calibrations. Our conceptual approach
100also brings new insights into the ecological meaning of results obtained from both P-O and P-A tree
101distribution models in terms of effective and potential distributions (Jiménez-Valverde *et al.* 2008;

102 Lobo *et al.* 2010). Secondly, we estimate Bivariate Selection Models (BSMs) using empirical data
 103 from the French national forest inventory. In addition to the classically used environmental
 104 predictors (climate, soil and topography), BSMs draws on economic predictors (returns from crops,
 105 forests and urban uses). Thirdly, we compared the BSM predictions with those classical SDMs
 106 calibrated with P-O and P-A datasets, for four widespread European tree species: sessile oak
 107 (*Quercus petraea*), pubescent oak (*Quercus pubescens*), common beech (*Fagus sylvatica*) and silver
 108 fir (*Abies alba*) across a range of spatial resolutions (2 km, 4 km and 8 km). BSM simulations are
 109 shown to increase the interpretative content of niche modelling by (i) distinguishing between
 110 potential and effective probability of presence and (ii) explicitly estimating the spatial interactions
 111 between the ecological suitability for a tree species and the economical conditions of having a
 112 forested land-use.

113 2. Methods

114 2.1. General framework

115 We model the potential presence of a particular species through the continuous latent (i.e.,
 116 not directly observed) variable μ . Environmental conditions of a site i are potentially suitable
 117 for this species if the corresponding value μ_i is greater than a given threshold, normalized to 0.
 118 On such sites with $\mu_i > 0$, the species of interest can be potentially observed. This latent variable
 119 is parametrized by an unknown smooth function f_p , a set of abiotic environmental variables
 120 X and an additive random component ε , with $\mu_i = f_p(X_i) - \varepsilon_i$. The subscript p marks
 121 terms relative to the potential distribution. By noting $(m_p = 1 | X_i)$ the event of "potentially
 122 observing the species of interest in the environment X_i ", we have:

123

$$124 \text{ Prob}(m_p = 1 | X_i) = \text{Prob}(\mu_i > 0) = \text{Prob}(\varepsilon_i < f_p(X_i)) \quad (1)$$

125

126 However, in human-dominated ecosystems, the observation of a forest tree species is also
127strictly conditioned by the presence of a compatible land-use, i.e., the presence of forest rather than
128agricultural or urban areas. As it is common in the econometric literature concerning discrete
129individual choices, we modelled land-use choices using the random utility framework (Lubowski *et*
130*al.* 2008; Lewis 2010; Ay *et al.* 2014). Accordingly, we assumed that observed land-use is the result
131of utility maximization. Utility is specified as the sum of monetary returns a landowner can expect
132from land-uses and a random term representing the unobserved part of utility. This unobserved part
133potentially accounts for non-measurable variables or non-monetary values.

134 We assumed that the forested land-use is the only one compatible with the observation of
135forest tree species. Considering that tree species inventories are typically restricted to forests, this
136assumption is implicitly made by many other empirical studies using such data. We can then note
137 \tilde{u}_i the difference between the utility from forest and the maximum utility from any other land-
138uses, which are incompatible with the presence of forest tree species. Hence, the compatible forest
139land-use is chosen by landowners when this utility difference is strictly positive. This utility
140difference is parametrized with a unknown smooth function f_l , the abiotic environmental
141variables X , the monetary returns W expected by landowner, and an additive random
142component ξ , $\tilde{u}_i = f_l(X_i, W_i) - \xi_i$. The subscript l marks terms relative to land-use choices.
143By noting $(m_l = 1 | X_i, W_i)$ the event of "having a compatible land-use in the environment X_i
144with the economic conditions W_i ", we note:

145

146 $\text{Prob}(m_l = 1 | X_i, W_i) = \text{Prob}(\tilde{u}_i > 0) = \text{Prob}(\xi_i < f_l(X_i, W_i)) \quad (2)$

147

1482.2. Land-use selection bias

149 In the following, we develop a formal description of the land-use selection bias by
150modelling jointly tree species distribution and land-use choices. Note that the two previous events
151(i.e., equations (1) and (2)) require observations of very different nature to be estimated.
152Realizations of $(m_p=1|X_i)$ are theoretical (i.e., related to potential presence) and are not
153directly observable from non-experimental data. Conversely, realizations of the land-use event
154 $(m_l=1|X_i, W_i)$ can be directly observed from any land-cover dataset. The land-use selection
155bias occurs because classical tree distributional data only contains the intersection of these two
156events, namely effective presence. As a consequence, what is classically observed and recorded in
157species distribution datasets is the event of “effectively observe the species in the environment
158 X_i with the economic conditions W_i ”, that we note $(m_e=1|X_i, W_i)$. The subscript e
159marks terms relative to the effective distribution that takes into account the land-use constraint.

160 Thus, the observable value m_e depends on the observable value m_l and the
161unobservable value m_p :

162

$$163 \quad m_e = m_p \times m_l \quad (3)$$

164

165Eqn 3 illustrates the uncertain nature of effective species absences (i.e., the event $m_e=0$) which
166is potentially attributable to $m_l=0$ and/or $m_p=0$. Consequently, m_p is not directly
167deductible from m_e . Additional information on land-use compatibility has the potential to clarify
168the link between m_e and m_p events in classical SDM modelling.

169 The land-use selection bias arising from confounding m_e and m_p can be expressed
170differently for SDM calibrated with presence-only (P-O) and presence-absence (P-A) data, both
171being currently used in tree species distribution studies (e.g. Dorazio 2014; Fithian *et al.*, 2014 for

P-O; e.g. Cheaib *et al* 2012; Gritti *et al*, 2013 for P-A). We choose to define P-A dataset as the case where absence records are limited to a compatible land use (i.e., forested sites where $m_l=1$). The absences recorded in the P-A dataset therefore only correspond to sites with unsuitable ecological conditions. This allows providing two contrasted benchmarked case study for modelers with clear mathematical results.

Predictions obtained from our P-O dataset correspond to the case where all sites are used in the SDM calibration (i.e., including sites with all types of land-use). This is equivalent to the typical case of occurrence records combined with naïve background data (e.g., Elith *et al.* 2006). The predictions obtained from our P-A dataset correspond to the case where only sites with a compatible land-use are used in the calibration (i.e., forested sites where $m_l=1$). The absences recorded in the P-A dataset therefore only correspond to sites with unsuitable ecological conditions, i.e. “confirmed” absences.

Depending on whether P-O or P-A data are used for the calibration, classical SDMs respectively estimate one of the following probabilities instead of the probability of interest $\text{Prob}(m_p=1 | X_i)$:

$$(P-O) \quad \text{Prob}(m_e=1 | X_i, W_i) \text{ or } (P-A) \quad \text{Prob}(m_e=1 | X_i, m_l=1) \quad (4)$$

The (P-O) case is the probability of the intersection of the events $m_p=1$ and $m_l=1$ for all sites:

$$\text{Prob}(m_e=1 | X_i, W_i) = \text{Prob}(\varepsilon_i < f_p(X_i) \cap \xi_i < f_l(X_i, W_i)) \quad (5)$$

Note that, as the events $m_l=1$ and $m_l=0$ are a partition of the probability space (i.e., their

probabilities sum to one), the probability of interest is the sum of this (P-O) probability and the probability of potential presence for incompatible land-uses:

199

$$\text{Prob}(m_p=1 | X_i, W_i) = \text{Prob}(\varepsilon_i < f_p(X_i) \cap \xi_i < f_l(X_i, W_i))$$

$$+ \text{Prob}(\varepsilon_i < f_p(X_i) \cap \xi_i \geq f_l(X_i, W_i)) \quad (6)$$

202

Consequently, we can express the bias from (P-O), i.e., the difference between the (P-O) probability and the probability of interest, as:

205

$$\text{(P-O) bias} = \text{Prob}(m_e=1 | X_i, W_i) - \text{Prob}(m_p=1 | X_i) \quad (7) = -\text{Prob}(\varepsilon_i < f_p(X_i) \cap \xi_i \geq f_l(X_i, W_i)) \quad (8)$$

207

Next, according to Bayes' theorem, the probability of interest can be expressed as:

209

$$\text{Prob}(m_p=1 | X_i, W_i) = \text{Prob}(m_p=1 | X_i, m_l=1) \times \text{Prob}(m_l=1 | X_i, W_i). \quad (9)$$

211

Note that, if $m_l=1$, the events $m_p=1$ and $m_e=1$ are equivalent. So the P-A probability of equation (4) is similar to the term $\text{Prob}(m_p=1 | X_i, m_l=1)$ from equation (9), and the associated P-A bias can be written as:

215

$$\text{P-A bias} = \text{Prob}(m_e=1 | X_i, m_l=1) - \text{Prob}(m_p=1 | X_i) \quad (10) = \frac{\text{Prob}(\varepsilon_i < f_p(X_i) \cap \xi_i < f_l(X_i, W_i))}{\text{Prob}(\xi_i < f_l(X_i, W_i))} - \text{Prob}(\varepsilon_i < f_p(X_i)) \quad (11)$$

217

218 The eqn (8) and (11) allow us to determine unambiguously the directions of the bias that
 219 affects classical SDM modelling. In case of P-O calibration (eqn 8) the bias is negative, so classical
 220 P-O SDMs always under-estimate the probability of potential presence. Moreover, it appears from
 221 eqn (8) that the P-O bias depends on the intersection of two inequalities on errors, meaning that it
 222 depends on the correlations between the errors of the ecological and economical equations. More
 223 precisely, the bias decreases as the coefficient value of the correlation between the errors increases
 224 from -1 to 1 (see the formal proof in section 1.2 of Supporting Information, SI).

225 To interpret the bias that affects P-A SDMs (equation 10-11) we should recall that, for
 226 independently distributed variables, the probability of the intersection is the product of the marginal
 227 probabilities. So, one can show by simplifying the numerator that the P-A bias is zero if the errors
 228 are independent. For positively correlated variables the probability of the intersection is greater than
 229 the product of marginal probabilities. In this case, the bias is positive and classical P-A SDMs over-
 230 estimate the probability of potential presence (see Section 1.2 of SI). In the case of negatively
 231 correlated errors, this P-A bias is negative.

232

233 **2.3. Empirical Implementation**

234 We applied this theoretical framework on tree species distribution modelling by estimating
 235 Bivariate Selection Models (BSMs, Marra et al. 2014) that included both ecological and economical
 236 drivers. BSMs were estimated with the package “SemiParBIVProbit” available on CRAN (Marra *et*
 237 *al.* 2014). BSMs require exclusion restrictions through the variables W_i for technical reasons
 238 presented in Section 1.3 of SI. BSMs allowed us estimating the correlations between the errors of
 239 the land-use choices and ecological equations and therefore inferring the potential interactions
 240 between land-use choices and responses of tree species (Section 1.4 of SI). More details regarding
 241 the technical aspects of BSM predictions are provided in Section 1.5 of SI.

242 BSMs are compatible with the generalized linear model (GLM) and with the generalized

additive model (GAM) frameworks, as they both allow estimating $f_l(X_i, W_i)$ jointly with $f_p(X_i)$ and taking into account the potential dependence between equation errors. BSMs were therefore estimated in both GLM and GAM frameworks, using probit-linked functions. The GAMs analyses were performed to ensure the robustness of the GLM results, as GAMs reduce errors due to model misspecification and therefore reduce the risk of misleading interpretations of the error correlation between the ecological and economical equations (Ahn & Powell 1993). The BSMs estimated using GLMs and those estimated using semiparametric GAMs yielded quantitatively similar results (Section 3 of SI). For the sake of conciseness only the GLM results are presented in the main text. Note that selection biases are not restricted to the GLM or GAM cases but apply to any model assuming errors independently distributed from land-use choice.

2.4. Data

2.4.1 Land-use and species distribution data

Land-use and species distribution data were derived from the French national forest inventory which provides a systematic record of tree species presence/absence on a regular 1 km grid over the mainland territory (Figure 1 of SI) for sessile oak (*Quercus petraea*), pubescent oak (*Quercus pubescens*), common beech (*Fagus sylvatica*) and silver fir (*Abies alba*). This dataset therefore allowed separating land-uses in two categories: forest and non-forest. To test for the effect of spatial resolution on BSMs calibration and predictions, the dataset was upscaled on squared grid at three different resolutions 2 km, 4 km and 8 km sides, in accordance with the environmental data. The upscaling procedure was straightforward: a pixel indicated the presence (of a species or of the forested land-use) at the coarser resolution if it included at least one pixel where the species or the land-use was indicated present at the finer resolution.

2.4.2 Environmental data

All the environmental variables were scaled at the different resolutions: 4 km², 16 km² and 8

267km². Climate variables were derived from the SAFRAN re-analysis which includes temperature,
268rainfall, and radiation on a 3 hourly basis at 8 km² resolutions (Vidal et al 2010). These variables
269were averaged at a monthly time step and downscaled at 2 km² and 4 km² resolutions using a thin
270plate spline interpolation procedure with 3 predictors (elevation, latitude and longitude),
271implemented in the packages “fields” and “raster” in R (R Development Core Team 2012). This
272methodology has been validated with surface observations of temperature and rainfall over a region
273of southern France by Ruffault et al (2014). From the downscaled climatic variables, we derived 6
274variables considered critical to plant physiological function and survival, as in Cheaib et al (2012),
275which are summarized in Table 2 of SI. The slope and exposure data were computed by applying
276the “terrain” function of the “raster” package of R to digital elevation models at each resolution.
277The 1 km French soil data base developed by the INRA (Infosol Unit, INRA, Orleans, Jamagne et
278al. 1995) was averaged at the different resolutions and used to calculate soil water holding capacity
279(Cheaib et al 2012). To include realistic environmental conditions and reduce multicollinearity, we
280selected the first two axes of a principal component analysis (PCA) based on monthly climate
281variables and the first axis of a PCA made on soil and topographic variables (see Figure 2 of SI).

2822.4.3 Economic data

283 Environmental predictors are also used in the econometric equation of land-use choices, in
284addition to some proxies of economic returns from the work of Ay *et al.* (2014). These authors
285approximated the monetary returns from crops by the land prices from the French ministry of
286agriculture in 2005, available at a regional scale named *Petites Régions Agricoles*. Monetary returns
287from forests were approximated by multiplying raw productions and unitary wood prices, divided
288by forest acreages. Monetary returns from urban area were approximated by population densities. A
289full description of the sample and the variables is reported in Table 1 and 2 of SI. Accessing
290monetary returns from different land uses is nevertheless not a necessary condition to estimate a
291BSM. Some other proxy variables could provide sufficient exclusion restrictions in numerous

situations (such as FAO suitability index or NASA human footprint map).

3. Results

3.1. Estimated parameters and response curves

Explanatory variables included in the BSM all explained significantly the land-use choices and the tree species presences (Table 3 to 8 of SI). The estimation of the econometric equation (2) yielded adjusted R^2 ranging from 0.09 to 0.16 with economically intuitive effects: the probability of having a compatible land-use (i.e., forest) increased with the return from forests and decreased with the returns from cropland and urban areas. The estimation of the ecological equation (1) yielded adjusted R^2 ranging from 0.05 to 0.34 with significant bioclimatic predictors. We found a 95% significant selection bias for the four tree species at 2 and 4 km resolutions while at 8 km only *F. sylvatica* presents a significant selection bias. At 2 km resolution, the two Oak species present positive correlations of the errors (respectively 0.95 and 0.96 for *Q. petraea* and *Q. pubescens*, Table 3 of SI) while the two other species present negative covariance (respectively -0.75 and -0.82 for *F. sylvatica* and *A. alba*, Table 3 of SI).

We simulated the effects of environmental predictors on the probabilities of species presence (response curves). Environmental factors were first projected in PCAs, and the first PCA axes were used as covariates (Figure 2 of SI). The predicted responses to a given covariate were computed by predicting the probability of presence while holding the other covariates at their sample means. The response curves estimated at 2 km resolution strongly differed between the BSM and the classical P-O and P-A SDMs for most of the species (Figure 1). The direction and magnitude of the deviation from the BSM response curves in the four species varied according to i) the sign of the correlation between the errors obtained from the environmental and economic equations and ii) the type of dataset used for the calibration of the classical SDMs (P-O or P-A). On the one hand, the two oak

species, with positive correlation presented a similar pattern: the classical P-A SDMs overestimated the probabilities compared to the BSM whereas classical P-O SDMs yielded response curves very close to those of the BSM (Figure 1, panel A and B). On the other hand, for the two other species with negative correlation the response curves obtained with either the P-O or the P-A SDMs strongly underestimated the probabilities of presence compared to the BSM predictions (Figure 1, panel C and D). For these latter species, the magnitude of the deviation from the BSM was always greater for P-O SDMS.

Expected location of Figure 1]

3.2. Prediction errors and spatial distribution

We compared the probabilities predicted by the BSM and by the classical P-O and P-A SDMs, at the 2 km, 4 km and 8 km resolutions. We used the root mean of squared error (RMSE) to quantify the differences in predictions (eqn 12), taking the BSM predictions at 2 km as reference:

$$RMSE_{m,r} = \sqrt{\frac{\sum_{i=1}^n (Prob_{i,m,r} - Prob_{i,BSM,2km})^2}{n}} \quad (12)$$

Where Prob is the estimated probability of presence, m={P-O SDM, P-A SDM, BSM}, r={2 km, 4 km, 8 km resolution}, n is the number of sites. For a given resolution, the estimates of P-O and P-A SDMs strongly diverged from the estimates of BSM for most species, yielding high RMSEs (Table 1). Decreasing the spatial resolution had very strong effects on the prediction of all SDMs, and on the resulting RMSE. The effect of the spatial resolution on SDM predictions is

indeed of the same order of magnitude than the effect related to the choice of the modelling approach (BSM or classical SDMs).

[Expected location of Table 1]

As an external source of data about tree species potential presence, we use the EuroVegMap dataset (www.floraweb.de/vegetation/dnld_eurovegmap.html) to provide an additional evaluation. Such data could be considered closer to the potential distribution relatively to IFN data because they were built at larger scale, upon an ensemble of criteria (e.g. presence of accompanying species, soil type and climate type) that may discriminate if an area of land is suitable for a given species independently from current land use. Figure 18 in SI 3.9 shows that, for species with negative correlations (*F.sylvatica*, *A.alba*), the relative performance of BSM increases highly, and could even slightly exceeded the levels of classical SDM for *A.alba*. This shows that BSM perform better at explaining the differences between the two data sources than classical SDMs.

The spatial patterns of predicted potential distribution differed markedly among the different models for each species at 2 km (Figure 2). This resulted in a bias (see Figure 15b in SI 3.8) affecting the predicted species distributions over the North Eastern-part and the Southern part of the territory in *Q. petraea* and *Q. pubescens*, respectively. Classical SDMs appeared to strongly underestimate the potential presence of *F. sylvatica* and *A. alba* with the exception of mountain areas where much smaller distribution areas were predicted.

[Expected location of Figure 2]

3633.3 *Interaction between ecological and economical gradients*

364
365 The additional information contained in BSM compared to classical SDMs can be assessed
366by computing the covariance between the ecological gradient μ_i and the economic gradient \tilde{u}_i .
367The total correlation ρ_T between the variables describing potential tree species presence and
368compatible land-use is decomposed between an observed part (ρ_O , due to the included
369predictors) and an unobserved part (ρ_U , due to errors' correlation) as formally shown in section
3701.4 of SI:

$$371 \quad \rho_T = \frac{\text{cov}(\mu_i, \tilde{u}_i)}{\sigma_T} = \frac{\text{cov}(f_p(X_i), f_l(X_i, W_i))}{\sigma_T} + \frac{\text{cov}(\varepsilon_i, \xi_i)}{\sigma_T} = \rho_O + \rho_U \quad (13)$$

372 A positive ρ_O indicates that the predictors have the same effects on ecological suitability
373and economic incentive of having a forested land-use, and this was the case for all species and for
374all the resolutions (Table 2). By contrast, the sign of ρ_U varied between tree species. An
375important result is that ρ_U had a greater influence on the total correlation than ρ_O : for every
376species, ρ_T and ρ_U had the same sign. This implies that neglecting the correlation between
377errors (as it is typically done by classical SDMs) induces a strong misestimation of the interactions
378between the ecological and economical gradients.

379

380Expected location of Table 2]

381

382 In the case of species with positive total correlations (*i.e.* *Quercus*), sites with high
383probability of presence are also sites with high probability of compatible land-use (Figure 3). This
384indicates a synergy between the two dimensions: the best suitable sites for the species are actually
385forested. The opposite is obtained for species with negative correlation, for which the hidden part of

the niche is composed of sites with above-than-average probabilities of potential presences (Figure 3). This indicates an antagonism between the ecological and the economical dimensions. Similar results were found at 4 km and 8 km resolution, highlighting the importance of ρ_T to characterise the interaction between ecological and economical gradients and the importance of BSMs to estimate it.

[Expected location of Figure 3]

4. Discussion

This study examines the influence of human decisions about land-use on the calibration and prediction of tree species distribution models (SDMs). We did so by developing a bivariate selection model (BSM) that is based on the coupling of an ecological equation of species distribution and an econometric equation of land-use choices. An empirical application on four tree species using standard statistical methods (GLMs and GAMs) illustrates and confirms the theoretical developments. It shows that classical SDMs may, depending on the species and the spatial resolution of the calibration dataset, either over or underestimate the potential probabilities of presence. We show that the bias is important at fine spatial resolution and decreases at coarse resolution. Hereafter, we discuss the causes of between-species variations and spatial resolution on the direction and magnitude of the bias. Finally, we discuss the implications of coupling ecological models of species distributions with econometric models of land-use choices.

4.1 Land-use selection bias affects the predictions of potential presence

We found a significant spatial correlation between the economic requirements of the forested land-use and the ecological requirements of every studied tree species. It is long recognized that in

anthropogenized areas, forests are preferentially located on poor soils with low water holding capacity, where no profitable agriculture is possible (Ricardo 1817). Therefore, tree species tolerant to such constraining environmental conditions will be more frequently found at forested sites than species with ecological requirements that match conditions favouring high economic returns for a different land-use (e.g. agriculture). This is well illustrated in our results: for *Quercus* species -- which are characterised by a relatively high tolerance to dry sites (Arend *et al.* 2011; Michelot *et al.* 2012; Lévesque *et al.* 2014, see also Figure 2) -- we found a positive correlations between the errors of the ecological and economical equations (Table 3). By contrast, for *Abies* and *Fagus* we found a negative correlation between errors of the two equations (Table 2). This is consistent with the fact that these latter species are relatively more sensitive to dry climate and shallow soil (Aranda *et al.* 2005), and grow preferentially on fertile sites of temperate areas which are therefore predominantly used for agriculture.

In agreement with the theory developed in this study, the correlation between ecological and economical gradients found in every species resulted in a significant land-use selection bias that affected the predictions of classical SDMs. We found that P-O SDMs always underestimate the probability of presence, whatever the species considered (Figure 3). All sites with land-uses incompatible with forests are indeed considered as absences in P-O SDM calibration, which consequently keeps some of the environmental conditions that are suitable for a given species out of the predicted potential niche. Contrastingly, the sign of the bias affecting P-A SDM predictions varies among species: positive bias in the two *Quercus* species and negative bias in *Abies* and *Fagus* (Table 2). Unlike P-O SDMs, P-A SDMs characterize the environmental conditions of unsuitable sites in order to predict species distributions and niches (Lobo *et al.*, 2010). However, if the ecological requirements of a tree species are preferably found at locations where land-use is incompatible (i.e., there is a negative correlation between ecological and economical gradients), the proportion of observed absence data is higher than in the case of a randomly distributed species. In

such case, P-A SDMs underestimate the predicted probability of presence (Figure 3, *Abies* and *Fagus*). Conversely, if the ecological requirements of a tree species are preferably found at locations where land-use is compatible, P-A SDMs overestimate the predicted probability of presence (Figure 3, *Quercus* sp.).

Overall, our results demonstrate that the land-use selection bias can lead to strong spatial miss-estimations of the predicted probabilities of presence of different tree species (Figure 2). Not accounting for this bias may have great implications for the many studies that deal with the probability of presence derived from classical P-O-SDM or P-A-SDM dataset in anthropized area. Furthermore, it is worth mentioning that such a bias will affect the predicted distribution of any species that is more specifically found within a certain land-use. Consequently, the projections of future distribution areas, that are a major tool for conservation strategy (Wilson *et al.* 2011), are likely to be biased as well. It is crucial to consider land-use selection bias in SDMs if we are to provide sound prediction of the future of species distributions in Western Europe and other anthropized areas.

4.2. How to evaluate SDMs in anthropized areas?

Our results question the relevance of the validation of statistical models based on presence-absence data from forest inventories. Indeed, the maps of probabilities computed with the BSM significantly differ from the presence-absence (P-A) records used to calibrate them, and such differences are far more important than for both classical SDMs (Figure 2, section 3.9 of SI). We argue that by accounting for the land-use selection bias, our BSM yields probabilities more closely related to the unobserved potential niche of the species than a classical SDM. This interpretation is consistent with (i) studies proposing that part of the differences between the realized and the potential niche is shaped by land cover modification related to human activities (Svenning *et al.*

4592010); and ii) with studies that found a significant role of land-use in the distribution of plant
460species independently of climate soil or topographic variables (e.g. Duckworth et al 2000).

461 It would be of great value to find new ways of validation/assessment for distribution models
462based on independent data. Such an evaluation is nevertheless quite a hard task given the high
463discrepancies reported when multiple sources of tree distribution data are compared (Duputie *et al.*,
4642013). Evaluation over pre-historical periods, when vegetation cover was not affected by human
465induces modifications of the land covers, by using distributions databases built from paleo materials
466may be a solution (Svenning *et al.*, 2011). However paleo-data are subjected to uncertainty far more
467important than classical inventory data (Svenning *et al.*, 2011). Another interesting way forward is
468the use of process-based models, including dynamic global vegetation models, that capture the
469physiological mechanisms determining plant survival and reproduction (Dormann *et al.* 2012).
470Because these models are based on the biophysical rules driving plant functioning, we expect their
471predictions to be unaffected by land-use selection bias, when they are calibrated on experimental
472data (Benito Garzón *et al.* 2011; Anderegg *et al.* 2015). In such case, they provide valuable estimate
473of the potential niche of the species (Morin *et al.* 2007; Hickler *et al.* 2012). A recent study
474conducted over France (Cheaib et al. 2012), showed that species distribution estimates from
475process-based models are close to the distribution observed in temperate oaks, but predict a larger
476distributional area than what is observed in common Beech. These additional presence areas
477correspond to regions that are mostly covered by agricultural land-use (e.g. Beauce region in
478Southern Paris, Normandie in North West France). This qualitative assessment echoes the results
479obtained in our theoretical development about the sign of the correlation between errors. Indeed,
480*Quercus* exhibits a positive correlation sign and is thus weakly affected by land use selection bias,
481whereas *Fagus* shows highly negative correlation leading to an important selection bias.

482 However, we should recall that process-based model predictions remain contingent on the
483hypothetical biological processes considered, and that they may be more uncertain than correlative

approaches estimates (Buckley *et al.* 2010). Promising attempts have been made to circumvent the limitations of individual modelling approaches, either in building hybrid models with both statistical and mechanistical features (Dormann *et al.* 2012), or by combining the predictive power of several independent models (Romero *et al.* 2015). These advances will likely help reducing the bias currently affecting SDM predictions. Given the current state of the art in species distribution modelling, comparing the simulations of a variety of approaches (Gritti *et al.* 2013) against multiple sources of data -- including expert knowledge -- remains necessary to ensure adequate evaluations of model outputs (Cheaib *et al.*, 2012; Duputie *et al.*, 2013).

4.3 The importance of land-use selection bias increases at fine resolution

An increasing number of studies have pointed out the importance of working at fine spatial resolution to properly assess the outcome of climate change on species persistence (Trivedi *et al.* 2008; Randin *et al.* 2009). These scientific results are followed by an increasing demand from applied researchers for fine resolution data. This leads to a growing importance of the land-use selection bias presented in this study, as we showed an increasing importance of the land-use selection bias with the spatial resolution (Table 2). Our results are in agreement with other studies showing that land-use can help to improve the explanatory power of species distribution models based on climate variable, mostly at fine resolution (Pearson *et al.* 2004; Luoto *et al.* 2007). More generally our results are in line with the premise that different factors shape the distribution of species according to a hierarchy of spatial scale; in which climate is more important at coarse scale and land-use at finer scale (Pearson *et al.* 2003). However and more importantly, the changes in the shape of the response curves to environmental variables between fine and coarse resolution have likely great implications for climate projections (Figure 2). Indeed we can reasonably expect that the projections under climate change scenario based on BSM calibrated at fine resolution will deeply differ from the projections based either on classical SDMs at fine resolution (without land-use) or on BSM at coarse resolution. Such pronounced effect of land-use contrasts with the different

509 studies showing little changes in the projected probabilities when land-use is accounted for (Barbet-
510 Massin et al 2012b; Martin et al 2013). Hence, our work highlights the need to jointly estimate
511 land-use choices and species responses to bioclimatic variables in a bivariate context to capture
512 their interacting effects on species distribution (Clavero *et al.* 2011; Mantyka-pringle *et al.* 2012)
513 and to deal with the land-use selection bias affecting the SDM calibration in human-dominated
514 ecosystems. Consider for instance *F. sylvatica*, a major European species whose distribution is
515 predicted to strongly shrink due to climate changes (Cheaib *et al.* 2012). It is in fact probable that
516 some suitable areas will last on the territory for decades, but we highlight in this study that the best
517 sites for this species are often not under a compatible land-use (i.e. they are not forested areas). So
518 maintaining *F. sylvatica* in France will have a cost that it is necessary to anticipate in order to refine
519 the predictions of the outcomes of global changes on species distributions (Fisher *et al.* 2011).

520

521 **4.4 Going further the selection bias: toward an integrated projecting**

522 The BSM that we present here constitutes an important step to account for the economic forces
523 shaping the observed species distributions. Selection bias is proven to be important for tree
524 distribution in France (Table 1) but other types of bias could persist in modelling species
525 distributions. One typical example is the high proportion of absence for species with a highly
526 negative correlation. There are some alternative models based on zero-inflated probability
527 distributions that could be used and even mixed with the BSM to correct for this. We believe that
528 this issue would deserve proper researches to improve modeling practices.

529 Our empirical estimates of the interactions between economic and ecological gradients (Figure
530 3; Table 2) revealed how the studied species are differently influenced by environmental and
531 economic variables. The use of explicit economic data to predict land-use paves the way for more
532 direct quantification of synergies and trade-off between environmental protection and economical
533 welfare. Using BSM for future projections imply a qualitative change of the role of economic
534 variables. From their technical role in removing the selection bias, economic variables become key

535 predictors for future land use choices. Hence, instead of using exogenous global/regional scenario
536 of future land-use (as generally done is SDM, e.g., Martin et al., 2013), the BSM would be based on
537 the trajectory of monetary returns from different land-use and may help to quantify the cost for
538 conservation of a given species. In our opinion, this alternative approach is promising to help
539 decision making by policies.

540

541

542 **Supporting Information:**

543 **1. Theory of the Bivariate Selection Model**

544 **2. Data description**

545 **3. Additional results**

546

547 **Biosketch** : The authors belong to a multidisciplinary project founded by the French government
548 (LabEx BASC) that aims to provide new knowledge regarding species distribution drivers and to
549 design innovative guidelines toward sustainable land management. The present team included both
550 ecologists interested in the functioning of plants in relation to the environment (Paul Leadley,
551 Nicolas K. Martin-StPaul and Joannès Guillemot) and economists specialized in land-use
552 economics and its influence on biodiversity (Luc Doyen, Jean-Sauveur Ay). The authors share an
553 interest in the understanding of species distribution in anthropized areas, and in the use of SDM to
554 gain knowledge and inform environmental decisions.

555

556 **5. References**

557 Ahn, H. & Powell, J.L. (1993). Semiparametric estimation of censored selection models with a
558 nonparametric selection mechanism. *Journal of Econometrics*, **58**, 3–29.

559 Ahn, H. & Powell, J.L. (1993). Semiparametric estimation of censored selection models with a
560 nonparametric selection mechanism. *Journal of Econometrics*, **58**, 3–29.

561 Anderegg, W.R.L., Flint, A., Huang, C., Flint, L., Berry, J.A., Davis, F.W., Sperry, J.S. & Field, C.B.
562 (2015). Tree mortality predicted from drought-induced vascular damage. *Nature Geosci*, **8**,

- 564Ay, J.-S., Chakir, R., Doyen, L., Jiguet, F. & Leadley, P. (2014). Integrated models, scenarios and
565 dynamics of climate, land use and common birds. *Climatic Change*, **126**, 13–30.
- 566Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W. & Courchamp, F. (2012). Impacts of climate
567 change on the future of biodiversity. *Ecology letters*, **15**, 365–377.
- 568Benito Garzón, M., Alía, R., Robson, T.M. & Zavala, M.A. (2011). Intra-specific variability and
569 plasticity influence potential tree species distributions under climate change. *Global Ecology
570 and Biogeography*, **20**, 766–778.
- 571Brook, B.W., Sodhi, N.S. & Bradshaw, C.J.A. (2008). Synergies among extinction drivers under
572 global change. *Trends in ecology & evolution*, **23**, 453–460.
- 573Buckley, L.B., Urban, M.C., Angilletta, M.J., Crozier, L.G., Rissler, L.J. & Sears, M.W. (2010). Can
574 mechanism inform species' distribution models? *Ecology letters*, **13**, 1041–1054.
- 575Cheaib, A., Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrêne, E., François, C., Gritti, E.S., Legay,
576 M., Pagé, C., Thuiller, W., Viovy, N. & Leadley, P. (2012). Climate change impacts on tree
577 ranges: model intercomparison facilitates understanding and quantification of uncertainty.
578 *Ecology letters*, **15**, 533–44.
- 579Chen, G., Kéry, M., Plattner, M., Ma, K. & Gardner, B. (2013). Imperfect detection is the rule rather
580 than the exception in plant distribution studies. *Journal of Ecology*, **101**, 183–191.
- 581Clavero, M., Villero, D. & Brotons, L. (2011). Climate change or land use dynamics: do we know
582 what climate change indicators indicate? *PLoS One*, **6**, e18581.
- 583Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin,
584 X., Römermann, C. & Schröder, B. (2012). Correlation and process in species distribution
585 models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.
- 586Duputié, A., Zimmermann, N.E. & Chuine, I. (2014). Where are the wild things? Why we need
587 better data on species distribution. *Global Ecology and Biogeography*, **23**, 457–467.
- 588Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R.,
589 Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion,
590 G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J.
591 Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S.
592 Wisz, M. & E. Zimmermann, N. (2006). Novel methods improve prediction of species'
593 distributions from occurrence data. *Ecography*, **29**, 129–151.
- 594Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011). A statistical
595 explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.
- 596Engler, R., Randin, C.F., Thuiller, W., Dullinger, S., Zimmermann, N.E., Araújo, M.B., Pearman,
597 P.B., Le Lay, G., Piedallu, C. & Albert, C.H. (2011). 21st century climate change threatens
598 mountain flora unequally across Europe. *Global Change Biology*, **17**, 2330–2341.
- 599Fisher, B., Edwards, D.P., Larsen, T.H., Ansell, F.A., Hsu, W.W., Roberts, C.S. & Wilcove, D.S.
600 (2011). Cost-effective conservation: calculating biodiversity and logging trade-offs in
601 Southeast Asia. *Conservation Letters*, **4**, 443–450.
- 602Franklin, J., Davis, F.W., Ikegami, M., Syphard, A.D., Flint, L.E., Flint, A.L. & Hannah, L. (2013).
603 Modeling plant species distributions under future climates: how fine scale do climate
604 projections need to be? *Global Change Biology*, **19**, 473–483.
- 605Gritti, E.S., Gaucherel, C., Crespo-Perez, M.-V. & Chuine, I. (2013). How can model comparison
606 help improving species distribution models? *PloS one*, **8**, e68823.
- 607Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat

models. *Ecology letters*, **8**, 993–1009.

Hickler, T., Vohland, K., Feehan, J., Miller, P.A., Smith, B., Costa, L., Giesecke, T., Fronzek, S., Carter, T.R. & Cramer, W. (2012). Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global Ecology and Biogeography*, **21**, 50–63.

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.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Wintle, B.A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515.

Lewis, D.J. (2010). An economic framework for forecasting land-use and ecosystem change. *Resource and Energy Economics*, **32**, 98–116.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Lubowski, R.N., Plantinga, A.J. & Stavins, R.N. (2008). What drives land-use change in the United States? A national analysis of landowner decisions. *Land Economics*, **84**, 529–550.

Mantyka-pringle, C.S., Martin, T.G. & Rhodes, J.R. (2012). Interactions between climate and habitat loss effects on biodiversity: a systematic review and meta-analysis. *Global Change Biology*, **18**, 1239–1252.

Marra, G., Radice, R. & Marra, M.G. (2014). Package ‘SemiParBIVProbit’.

Martin, Y., Van Dyck, H., Dendoncker, N. & Titeux, N. (2013). Testing instead of assuming the importance of land use change scenarios to model species distributions under climate change. *Global Ecology and Biogeography*, **22**, 1204–1216.

Morin, X., Augspurger, C. & Chuine, I. (2007). Process-based modeling of species’ distributions: what limits temperate tree species’ range boundaries? *Ecology*, **88**, 2280–2291.

Pearson, R.G. & Dawson, T.P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global ecology and biogeography*, **12**, 361–371.

Pearson, R.G., Dawson, T.P. & Liu, C. (2004). Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

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.

Romero, D., Olivero, J., Brito, J.C. & Real, R. (2015). Comparison of approaches to combine species distribution models based on different sets of predictors. *Ecography*.

Smith, A.B. (2013). On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, **19**, 867–872.

Stanton, J.C., Pearson, R.G., Horning, N., Ersts, P. & Reşit Akçakaya, H. (2012). Combining static and dynamic variables in species distribution models under climate change. *Methods in Ecology and Evolution*, **3**, 349–357.

652Thuiller, W., Araujo, M.B. & Lavorel, S. (2004). Do we need land-cover data to model species
653 distributions in Europe? *Journal of Biogeography*, **31**, 353–361.

654Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005). Climate change threats
655 to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United*
656 *States of America*, **102**, 8245–8250.

657Warton, D.I., Renner, I.W. & Ramp, D. (2013). Model-Based Control of Observer Bias for the
658 Analysis of Presence-Only Data in Ecology. *PloS one*, **8**, e79168.

659Wilson, C.D., Roberts, D. & Reid, N. (2011). Applying species distribution modelling to identify
660 areas of high conservation value for endangered species: A case study using *Margaritifera*
661 *margaritifera* (L.). *Biological Conservation*, **144**, 821–829.

662Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B. & Svenning, J. (2010). New
663 trends in species distribution modelling. *Ecography*, **33**, 985–989.

664

665

Table 1: Comparison of the probabilities predicted by the three modelling approaches: BSM, (P-O) and (P-A) SDMs. Figures are RMSEs calculated using BSM predictions at 2 km as reference.

Species	Scale	BSM	PO	PA
<i>Q.petraea</i>	2km	0.00	0.00	0.19
	4km	0.17	0.17	0.28
	8km	0.46	0.39	0.43
<i>Q.pubescens</i>	2km	0.00	0.10	0.23
	4km	0.14	0.14	0.19
	8km	0.29	0.29	0.31
<i>F.sylvatica</i>	2km	0.00	0.48	0.35
	4km	0.24	0.41	0.34
	8km	0.33	0.33	0.31
<i>A.alba</i>	2km	0.00	0.33	0.29
	4km	0.28	0.32	0.32
	8km	0.31	0.30	0.31

Table 2: Decomposition of the spatial covariance between ecological and economical gradients. ρ_T is the total correlation between ecological and economical components. It is decomposed into an observed part ρ_O (due to the included predictors) and a unobserved part ρ_U (due to the errors' correlation). The terms in brackets represent the confidence intervals at 95%.

673

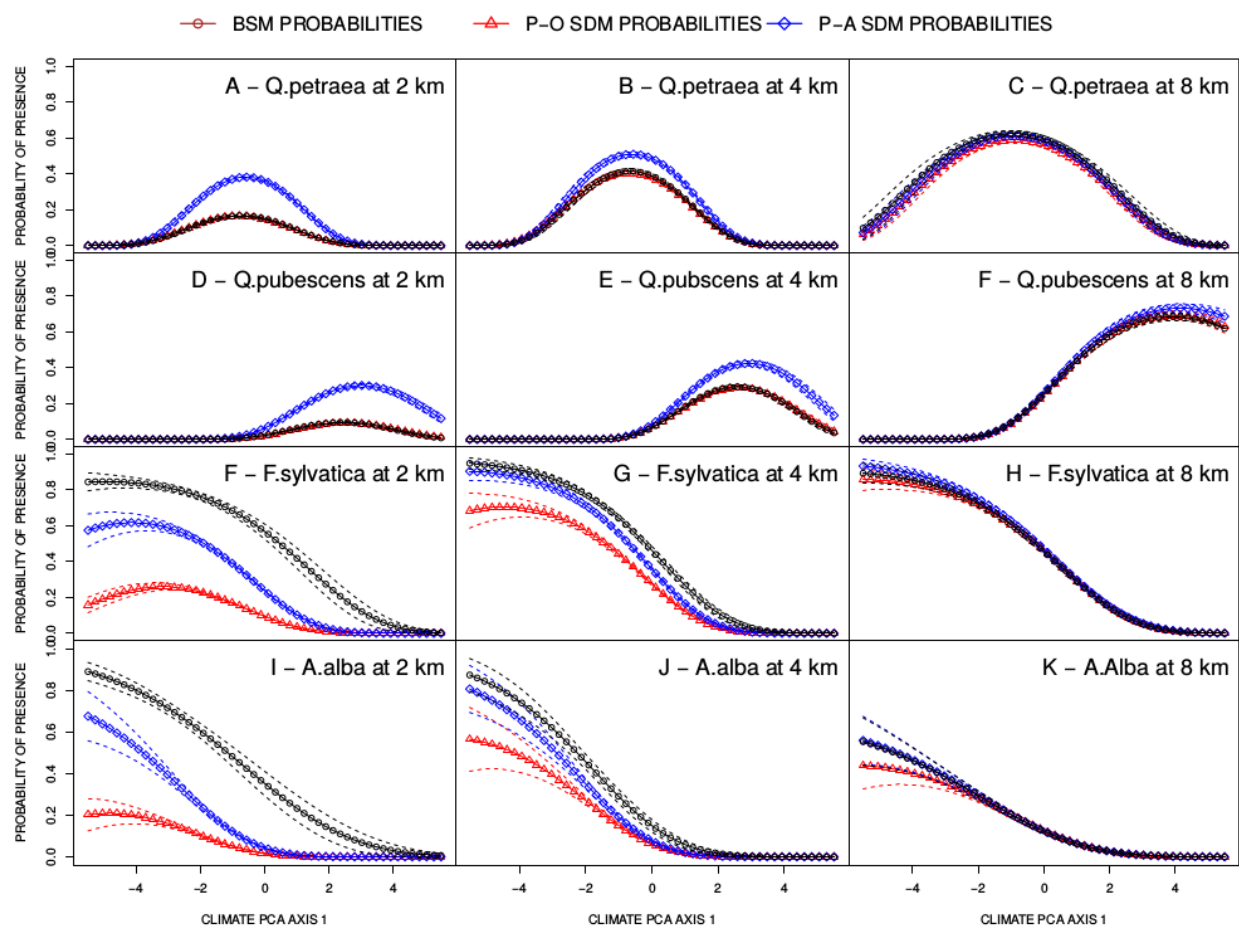
Resol.	Species	TOTAL		OBSERVED		UNOBSERVED	
		ρ_T	IC 95%	ρ_O	IC 95%	ρ_U	IC 95%
2 KM	Q.petraea	0.707	[0.58, 0.81]	0.143	[0.06, 0.23]	0.565	[0.53, 0.58]
	Q.pubescens	0.781	[0.65, 0.88]	0.180	[0.1, 0.27]	0.600	[0.55, 0.62]
	F.sylvatica	-0.568	[-0.66, -0.46]	0.007	[-0.04, 0.05]	-0.575	[-0.62, -0.51]
	A.alba	-0.620	[-0.79, -0.44]	0.041	[-0.09, 0.17]	-0.660	[-0.7, -0.61]
4 KM	Q.petraea	0.567	[0.35, 0.72]	0.136	[0.04, 0.23]	0.431	[0.31, 0.49]
	Q.pubescens	0.727	[0.52, 0.87]	0.226	[0.1, 0.35]	0.502	[0.42, 0.53]
	F.sylvatica	-0.325	[-0.43, -0.21]	0.078	[0.03, 0.12]	-0.403	[-0.46, -0.33]
	A.alba	-0.246	[-0.58, 0.11]	0.153	[-0.1, 0.41]	-0.399	[-0.47, -0.3]
8 KM	Q.petraea	-0.307	[-0.56, 0.14]	0.020	[-0.03, 0.07]	-0.327	[-0.53, 0.07]
	Q.pubescens	0.712	[-0.43, 0.75]	0.163	[0.12, 0.2]	0.549	[-0.55, 0.55]
	F.sylvatica	0.520	[0.21, 0.75]	0.153	[0.02, 0.29]	0.366	[0.19, 0.47]
	A.alba	0.258	[-0.11, 0.63]	0.213	[0.01, 0.42]	0.044	[-0.12, 0.21]

674

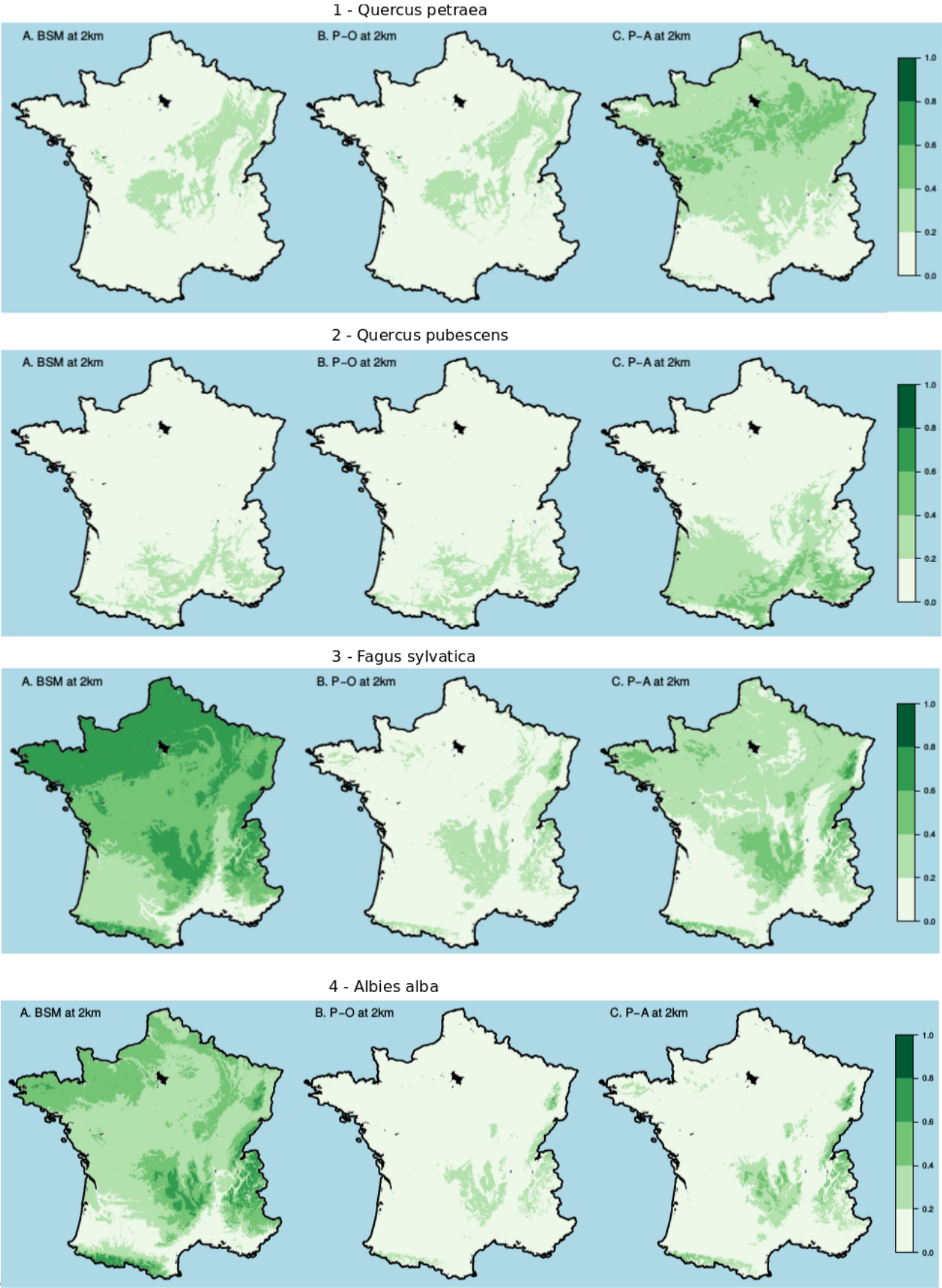
675

676

677**Figure 1.** Predicted response curves of the presence probabilities to the environmental gradient
 678(first axis of the PCA on environmental variables, see Figure 2 of SI).
 679



682**Figure 2.** Spatial distribution of the predicted potential presence at 2 km resolution.



683

684**Figure 3:** Probabilities of potential presence of tree species according to the presence of compatible
685land use (i.e., forested land use).
686

