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 Doyen[7] and Paul Leadley[2,3,4]

Supporting Information

1: INRA, UMR 1041 CESAER, F-21079 Dijon (France)
2: AgroParisTech, F-75231 Paris (France)
3: Université Paris Sud, UMR 8079 ESE, F-91405 Orsay (France)
4: CNRS, F-91405 Orsay (France)
5: CIRAD, UMR ECO & SOLS, F-34398 Montpellier (France)
6: INRA, URFM, F-84914 Avignon (France)
7: CNRS, UMR GREThA, F-33600 Bordeaux (France)

Abstract

This supplementary file contains the supporting information for the paper mentioned above. It contains additional insights about the theoretical structure of the Binary Selection Model (section 1), the data used (section 2), and the results (section 3) of our French application.

List of Tables

List of Figures

1 Theory of the Binary Selection Model

1.1 Estimated probabilities

To provide additional theoretical insights from the proposed Binary Selection Model (BSM), we consider the particular Generalized Linear Model case (GLM) of the more general framework presented in the main text. Both tree species presence and land-use choice are assumed to be linear functions of random predictors. From the notations of the main text, we set :

$$f_p(X_i) = \alpha + \beta x_i \quad \text{and} \quad f_\ell(X_i, W_i) = \eta + \gamma x_i + \theta w_i. \tag{1}$$

We also restrict the predictors x_i and w_i to be of dimension 1 to simplify the notations. Next, we assume that the errors of both equations are jointly distributed according to a bivariate Normal distribution of zero means, unit variances, and correlation ρ. This parametrization corresponds to the typical case of a bivariate probit used in most applications of selection models [4, 7]. The normalization of variances is necessary because the coefficients of binary response models are only identifiable up to scale. This reads as:

$$\begin{bmatrix} \varepsilon \\ \xi \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \tag{2}$$

The theoretical probability of potential presence of the tree species of interest on the site i is $\text{Prob}(\varepsilon_i < \alpha + \beta x_i) = \Phi(\alpha + \beta x_i)$, and the probability of having a compatible land use is $\text{Prob}(\xi_i < \eta + \gamma x_i + \theta w_i) = \Phi(\eta + \gamma x_i + \theta w_i)$. $\Phi(\cdot)$ is the cumulative distribution function of a standardized Normal distribution and, for future reference, $\phi(\cdot)$ is the associated density function. The biased probability (P-O) – see equation (4) in the main text – can be noted as:

$$\begin{aligned} \text{Prob}(m_e = 1 \mid x_i, w_i) &= \text{Prob}(\varepsilon_i < \alpha + \beta x_i \ \cap \ \xi_i < \eta + \gamma x_i + \theta w_i) && (3) \\ &= \int_{-\infty}^{\eta+\gamma x_i+\theta w_i} \int_{-\infty}^{\alpha+\beta x_i} \phi\left(\frac{z - \rho v}{\sqrt{1-\rho^2}} \right) \phi(v) \, \mathrm{d}z \, \mathrm{d}v && (4) \\ &= \int_{-\infty}^{\eta+\gamma x_i+\theta w_i} \Phi\left(\frac{\alpha + \beta x_i - \rho v}{\sqrt{1-\rho^2}} \right) \phi(v) \, \mathrm{d}v && (5) \end{aligned}$$

1.2 Land-use selection bias

The integral of the right-hand side of Equation 5 does not have an analytical closed form without using quadrature procedures or linearization. This makes the intuitions about the effect of ρ more tricky to obtain. Note that the biased probability (P-A) of the main text faces the same problem, as $\text{Prob}(m_e = 1 \mid x_i, m_\ell = 1) = \text{Prob}(m_e = 1 \mid x_i, w_i)/\Phi(\eta + \gamma x_i + \theta w_i)$.

Computing the second order Taylor approximation of the probability (P-A) around $\rho = 0$ leads to the following expression:

$$\begin{aligned} \text{Prob}(m_e = 1 \mid x_i, m_\ell = 1) &\approx \Phi(\alpha + \beta x_i) + \rho \frac{\phi(\alpha + \beta x_i)}{\Phi(\eta + \gamma x_i + \theta w_i)} \int_{-\infty}^{\eta+\gamma x_i+\theta w_i} v\phi(v) \, \mathrm{d}v && (6) \\ &\approx \Phi(\alpha + \beta x_i) + \rho\phi(\alpha + \beta x_i)\lambda(\eta + \gamma x_i + \theta w_i). && (7) \end{aligned}$$

We define the function $\lambda(u) \equiv \phi(u)/\Phi(u)$ as the inverse Mill's ratio, a well-known conditional expectation used to correct the selection bias in linear regression models [4]. Through extensive Monte Carlo simulations, [6] shown that this approximation is accurate for a wide range of value of

ρ (i.e., for $|\rho| \lessapprox 0.8$). In particular, the quality of the approximation depends on how far is ρ from zero. Note that these approximations are only used here to provide some intuitions about the sign and the size of the land-use selection bias, the empirical application presented in the main text is based on a Full Penalized Likelihood procedure which does not depend on them.

Because, by definition, density and cumulative distribution functions are both positive, previous Equation 7 shows unambiguously the direction of the bias when using the probability (P-A) instead of the structural probability of interest $\Phi(\alpha + \beta x_i)$. As it is stated in the main text, this bias is:

Positive when the errors terms are positively correlated. The true probability of potential presence is over-estimated by classical presence-absence SDMs

Negative when the errors terms are negatively correlated. The true probability of potential presence is under-estimated by classical presence-absence SDMs

For the other classical SDMs, namely presence-only (PO) SDMs, the same second order Taylor linearization allows to write the biased probability obtained as:

$$\text{Prob}(m_e = 1 \mid x_i, w_i) \approx \Phi(\alpha + \beta x_i) \times \Phi(\eta + \gamma x_i + \theta w_i) + \rho\phi(\alpha + \beta x_i)\phi(\eta + \gamma x_i + \theta w_i). \quad (8)$$

The linear approximation maintains the general results for $\rho = 0$ presented in the main text, e.g., biased probabilities (P-O) have to be divided by the probability of having a compatible land use (i.e., $\text{Prob}(m_\ell = 1 \mid x_i, w_i) = \Phi(\eta + \gamma x_i + \theta w_i)$) to be unbiased. One additional insight from the second order Taylor linearization operated here is that the absolute value of the bias from predicting probability (P-A) as a true potential presence is increasing with the deterministic part of the utility difference $\eta + \gamma x_i + \theta w_i$. Hence, bias from classical P-A SDMs are higher on the sites the most economically suitable for forest (i.e., the less dependent on land-use changes).

1.3 Exclusion restrictions

In the absence of additional economic predictors w_i, the BSM is technically identified but only by the non-linearity of the smooth conditional expectation $\lambda(\cdot)$, i.e., without additional, external information [4, 7]. This situation is unsatisfactory, as it is shown by Equation 7 of this supplementary file, where, in the absence of w_i, the correction for selection enters in the model as a nonlinear function only of the covariate x_i. Then, the only difference between the selection effect and the true effect of the environmental covariate x_i is the nonlinear structure of $\lambda(\cdot)$. This situation is unsatisfactory because, first, we deal with selection bias by imposing functional form assumptions that are exogenous to the data we analyze and, second, in practice there is often so little variation in $\lambda(\cdot)$ compared to the variation in x_i so that the BSM is unidentified.

One powerful way of improving identification of the BSM is to include extra covariate(s) w_i which appear(s) only in the economic equation of land-use choices. [11] shown that using additional, orthogonal covariate(s) in the selection equation can improve drastically the precision of the estimation by decreasing multi-collinearity. [5] found that the presence of exclusion restrictions is more important than the assumption about the distribution of errors. The economic returns from compatible and incompatible land uses are very intuitive candidates that are shown to be relevant in our application. They act as instrumental variables, producing exogenous variations in the probability of having a compatible land use, all other things equal (e.g., the environmental requirements for the tree species). The instrumental variables should affect the probability of making a particular land-use choice but should not have any direct effect on the potential probability of tree species presence. By including these variables, one insures that the selection effect varies independently of the true effect and that the BSM is not exclusively identified from functional form assumptions [7].

1.4 Spatial covariances

In order to illustrate the dependence between the ecological and economical latent variables in the main text, we use what we call spatial covariances. In particular, we decompose the total spatial covariance ρ_T between these two gradients as the sum of an observed and a unobserved terms (respectively noted ρ_O and ρ_U). This decomposition is based on the following expressions, obtained by substituting the formulas for μ_i and $\widetilde{u}_i$ (respectively equations 1 and 2 in the main text) in the definition of the covariance between two continuous variables:

$$\begin{aligned} \text{cov}(\mu_i, \widetilde{u}_i) &= \text{cov}\big[f_p(X_i) - \varepsilon_i, f_\ell(X_i, W_i) - \xi_i\big] && (9)\\ &= \text{cov}\big[f_p(X_i), f_\ell(X_i, W_i)\big] + \rho. && (10) \end{aligned}$$

We use the fact that, by definition, $\text{cov}\big[f_p(X_i), \xi_i\big] = \text{cov}\big[f_\ell(X_i, W_i), \varepsilon_i\big] = 0$. Previous terms are then normalized by the product of the standard deviations of the latent variables μ_i and $\widetilde{u}_i$, namely $\sigma_T = \sqrt{1+\sigma_p^2} \times \sqrt{1+\sigma_\ell^2}$. Hence, we obtain the decompositions reported in Table 3 of the main text, where every term is inside the unit interval:

$$\rho_T = \text{cov}\big[\mu_i, \widetilde{u}_i\big]/\sigma_T \text{ and } \rho_O = \text{cov}\big[f_p(X_i), f_\ell(X_i, W_i)\big]/\sigma_T \text{ and } \rho_U = \rho/\sigma_T. \quad (11)$$

1.5 Predictions from BSMs

In the same fashion that the total spatial covariance contains an observed and unobserved parts, looking at the dependence between economical and ecological gradients require to take into account the correlation between errors. To construct Figure 4 of the main paper and Figures 9 to 15 in this SI file, we use the observed values of land use m_ℓ that, jointly with errors' correlation, allows to represent the full covariance between economical and ecological gradients. By using the notations of Greene (1998), we set $s = 2 \times m_\ell - 1$ to obtain:

$$\begin{aligned} Prob(m_p = 1|m_\ell, X) =& m_\ell \times Prob(m_p = 1|m_\ell = 1, X) + (1 - m_\ell) \times Prob(m_p = 1|m_\ell = 0, X)\\ =& m_\ell \frac{\Phi(f_\ell(X, W), f_p(X); \rho)}{\Phi(f_\ell(X, W))} + (1 - m_\ell)\frac{\Phi(-f_\ell(X, W), f_p(X); -\rho)}{\Phi(-f_\ell(X, W)} && (12)\\ =& \frac{\Phi(s \times f_\ell(X, W), f_p(X_i); s \times \rho)}{\Phi(s \times f_\ell(X, W))} && (13) \end{aligned}$$

The major difference with classical predictions is that m_ℓ is included in the information set used to perform predictions, so more information is used.

References

[1] Jean-Sauveur Ay, Raja Chakir, Luc Doyen, Frederic Jiguet, and Paul Leadley. Integrated models, scenarios and dynamics of climate, land use and common birds. *Climatic Change*, 126(1-2):13–30, 2014.

[2] Alissar Cheaib, Vincent Badeau, Julien Boe, Isabelle Chuine, Christine Delire, Eric Dufrêne, Christophe François, Emmanuel S Gritti, Myriam Legay, Christian Pagé, et al. Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty. *Ecology letters*, 15(6):533–544, 2012.

[3] Marra G. and Radice R. *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*, 2014. R package version 3.2-10.

[4] James J Heckman. Sample selection bias as a specification error. *Econometrica*, pages 153–161, 1979.

[5] Arthur Lewbel. Endogenous selection or treatment model estimation. *Journal of Econometrics*, 141(2):777–806, 2007.

[6] Cheti Nicoletti and Franco Perracchi. Two-step estimation of binary response models with sample selection. *Working Paper*, Faculty of Economics(Tor Vergata University):Rome, 2001.

[7] Patrick Puhani. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68, 2000.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[9] Julien Ruffault, Nicolas K Martin-StPaul, Carole Duffet, Fabien Goge, and Florent Mouillot. Projecting future drought in Mediterranean forests: Bias correction of climate models matters! *Theoretical and applied climatology*, 117(1-2):113–122, 2014.

[10] Wood S.N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

[11] Francis Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.

[12] Jean-Philippe Vidal, Eric Martin, Laurent Franchistéguy, Martine Baillon, and Jean-Michel Soubeyroux. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *International Journal of Climatology*, 30(11):1627–1644, 2010.

2 Data description

estimated using probit-linked functions for both Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs). While other models exist for generating SDMs in the literature (e.g., through maximum entropy, simulated annealing, neural networks), we argue that they are also sensitive to the land-use selection bias studied here as they all assume errors independently distributed from land-use choices. Hence, the bias presented is more general than the GLM and GAM cases. Taking into account selection bias in more complex models is an interesting issue that we put outside the scope of this paper.

We assume that functions $f_p(X_i)$ and $f_\ell(X_i, W_i)$ are additive polynomials of order 2 in GLMs and penalized additive splines in GAMs. As mentioned above, P-O SDMs consist in estimating $f_p(X_i)$ using the whole data sample while P-A SDMs restrict the sample to forested sites (see Table 1 of SI). Both classical SDMs are estimated with R, using the widely used `mgcv` package [10]. BSMs are compatible both with the GLM and GAM frameworks, while they estimate $f_\ell(X_i, W_i)$ jointly with $f_p(X_i)$ and take into account the potential absence of independence between errors. BSMs are estimated with the package `SemiParBIVProbit` available on CRAN [3]. BSMs require exclusion restrictions through the variables W_i for technical reasons presented in Section 1.3 of this SI. BSMs allowed us to estimate the correlations between errors of the land-use choices and the ecological equation and therefore to infer the potential interactions between the land-use choices and the responses of tree species. Such interactions are studied through what we call spatial covariances, a formal definition of them is available at Section 1.4 of SI. Finally, the details for performing predictions from BSMs are presented in Section 1.5 of SI.

The detailed results obtained from the BSM estimated using GLMs and bivariate Normal errors or with semiparametric GAMs were quantitatively similar and are provided in SI. We performed the GAMs analysis to ensure the robustness of the GLM results since GAMs reduce the errors due to model misspecification and therefore reduce the risk of misleading interpretations of the error correlation between the ecological and economical equations. We also estimated models with errors based on copula functions without founding significant differences.

2.1 Data

The BSMs models were applied over France under historical environmental conditions to four widespread tree species with contrasted distributions (Figure 1): sessile oak (*Quercus petraea*), pubescent oak (*Quercus pubscens*), common beech (*Fagus sylvatica*) and silver fir (*Abies alba*).

2.1.1 Land use and species distribution data

Land use and species presence / absence data were derived from the French national forest inventory which provides a systematic record of tree species presence/absence on a regular 1km grid over the mainland territory (Figure 1). This dataset therefore allowed to separate land uses in two categories: forest and non-forest (Figure 1). To test for the effect of spatial resolution on BSMs calibration and predictions, the dataset was upscaled at three different resolutions 2km^2, 4km^2 and 8 km^2 in accordance with the environmental data. The upscaling procedure was straightforward: For a pixel to indicate the presence (of a species or of the land use forest) at the upper resolution, it must include at least one pixel where the species or the land use is indicated present at the higher resolution (see Figure 1 below).

2.1.2 Environmental data

The environmental predictors were computed with the same climatic (temperature, precipitations, etc.) and pedo-topographic (water holding capacity, slope, etc.) databases used in [2]. In the present study, to test the effect of the spatial resolution, all the environmental variables were scaled at $2km^2$, $4km^2$ and $8km^2$ resolutions. Climate variables were derived from the SAFRAN re-analysis which includes temperature, rainfall, and radiation on a 3 hourly basis at 8 km^2 resolutions [12]. These variables were averaged at a monthly time step and downscaled at 2 km^2 and 4 km^2 resolutions using a thin plate spline interpolation procedure with 3 predictors (elevation, latitude and longitude), implemented in the packages `fields` and `raster` in R [8]. This methodology has been validated with surface observations of temperature and rainfall over a region of southern France by [9]. From the downscaled climatic variables, we derived 6 variables considered critical to plant physiological function and survival as in [2], which are summarized in Table 2. The slope and exposure data were computed by applying the `terrain` function of the `raster` package of R to digital elevation models at each resolution. The 1km French soil data base developed by the INRA (Infosol Unit, INRA, Orleans) and described in [2] was was averaged at the different resolutions.

Table 1: **Tree species presences and prevalences according to land use.**
For the three spatial resolutions of interest (2, 4, and 8 km), our data contain respectively 134,328 and 33,626 and 8,427 grids on the whole contiental France. For each resolution, the first row of the Table below reports the number of grids where a considered species is observed (presence), the second rows report the % of forest that these grids represent and the third rows the % of all grids (i.e., prevalence).

Resolution	Statistics	*Q.petraea*	*Q.pubescens*	*F.sylvatica*	*A.alba*
2 km	Presence (# of grids)	8693	4660	8641	3404
	Prevalence (% of forest)	20.61	11.05	20.49	8.07
	Prevalence (% of all)	6.47	3.47	6.43	2.53
4 km	Presence (# of grids)	4717	2502	4598	1810
	Prevalence (% of forest)	21.51	11.41	20.97	8.25
	Prevalence (% of all)	14.03	7.44	13.67	5.38
8 km	Presence (# of grids)	1965	1139	1720	655
	Prevalence (% of forest)	25.73	14.91	22.52	8.58
	Prevalence (% of all)	23.32	13.52	20.41	7.77

To include realistic environmental conditions and reduce multicollinearity, we selected the first two axes of a principal component analysis (PCA) based on monthly climate variables and the first axis of a PCA made on pedo-topographic variables (see Figure 2).

Figure 1: **Species and forest distributions according to the scales.**

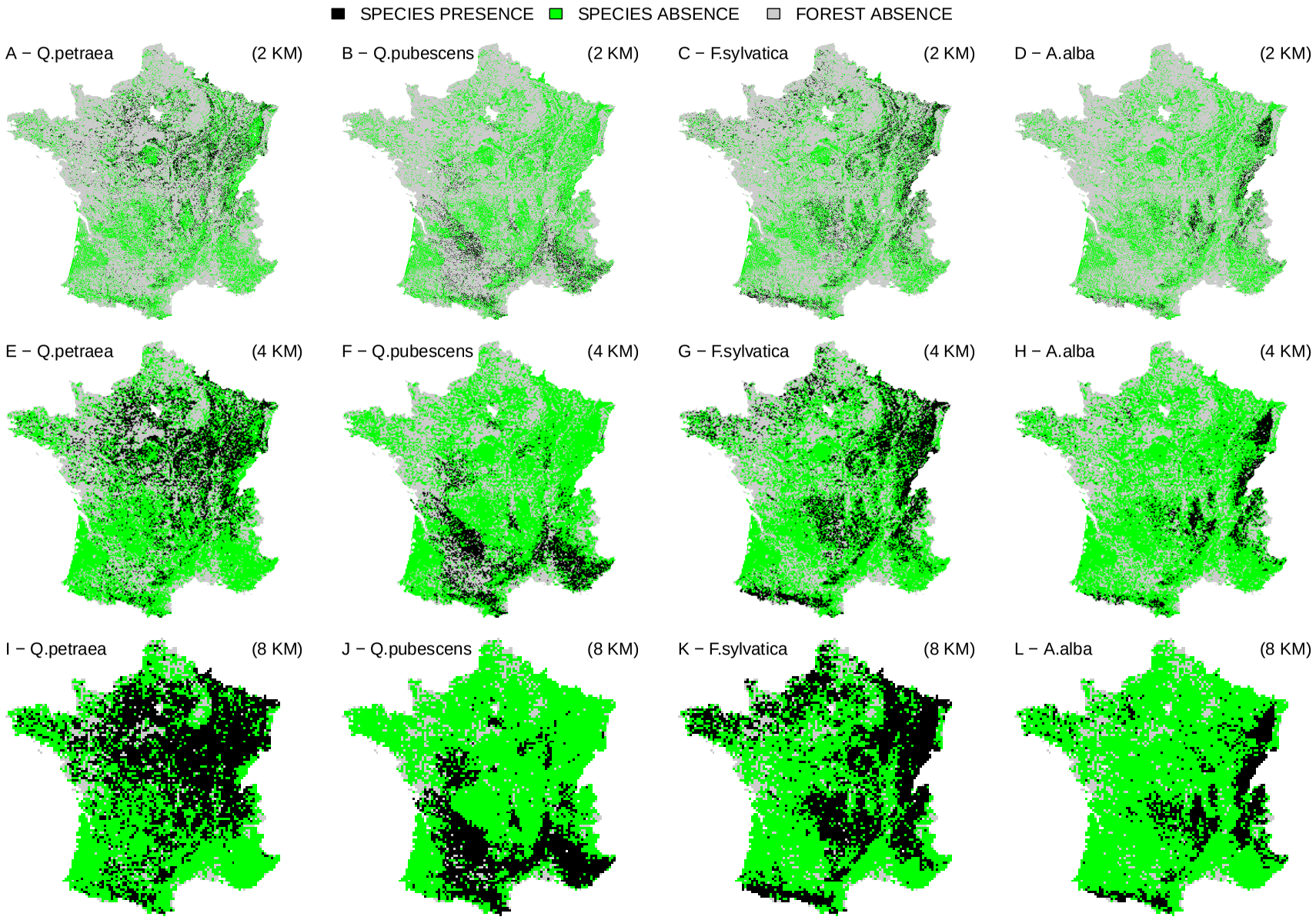
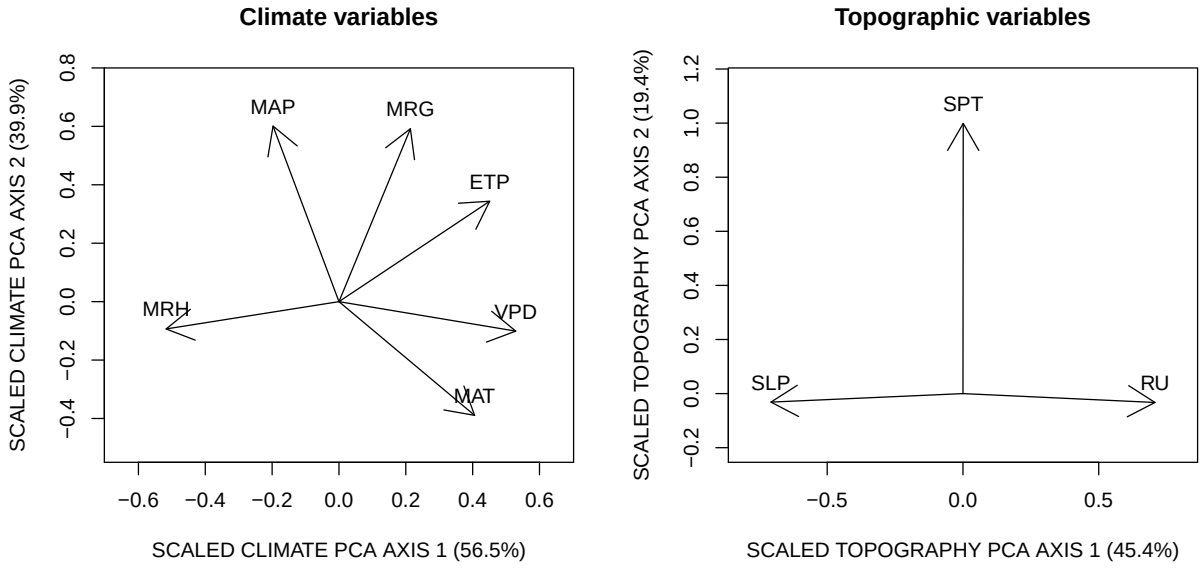


Figure 2: **Principal component analysis of raw climatic and pedo-topographic variables.** The labels of the initial raw variables are described in Table 2 of this SI. In our regression analysis, we keep only the 2 best principal axis of climatic variables and the first principal axis of pedo-topographic variables. The 2 climatic principal axis account respectively for 56.5 and 39.9% of the total variance of climatic variables, and the pedo-topographic axis accounts for 45.4%.



2.1.3 Economic variables

Previous variables are also used in the econometric equation of land-use choices, in addition to some proxies of economic returns from the work of [1]. They approximated the monetary returns from crops by the land prices from the French ministry of agriculture in 2005, available at a regional scale named Petites Régions Agricoles. Monetary returns from forests were approximated by multiplying raw productions and unitary wood prices, divided by forest acreages. Monetary returns from urban area were approximated by population densities. A full description of the sample and the variables is reported in Table 2.

Table 2: **Summary statistics for predictors used in models (4km).**
See section 2.3 about empirical implementation in the main text for the details of computations and downscaling.

Statistic	Mean	St. Dev.	Min	Max
MAT : Mean of Annual Temperatures (°C)	10.908	1.997	−4.293	14.927
MAP : Cumulative Annual Precipitations (mm)	934.951	144.681	761.930	1,996.730
ETP : Potential Evapotranspiration (mm)	701.569	73.899	552.751	886.242
VPD : Vapor Pressure Deficit (KPa)	0.543	0.151	0.067	1.164
MRH : Mean of Relative humidity (%)	83.837	3.018	64.202	93.464
MRG : Mean of annual Solar Radiations (MJ)	5,159.772	430.648	4,318.510	6,586.610
WHC : Water Holding Capacity (mm)	130.816	51.038	12.000	487.006
SLP : Mean of the Slope (degree)	0.087	0.103	0.000	0.641
SPT : Geographical Aspect (radian)	3.178	0.452	1.024	5.369
PCC1 : Climate PCA axis 1 (scaled)	0.000	1.834	−5.708	6.296
PCC2 : Climate PCA axis 2 (scaled)	0.000	1.547	−2.288	9.434
PCT1 : Topography PCA axis 1 (scaled)	0.000	1.209	−5.114	5.529
RT.FOR : Returns from forest (00 € /ha.yrs)	1.088	0.130	1.000	1.792
RT.AGR : Returns from croplands (00 € /ha.yrs)	1.184	0.102	1.000	2.630
RT.POP : Population Density (000 Pop./km^2)	0.183	0.503	0.005	15.168

3 Additional results

3.1 Regression tables

Table 3: **Raw GLM results at 2 km resolution.**
For the 4 tree species of interest (in columns) we report the estimated coefficients and standard errors (in parenthesis) for the two equations: the land-use choice and the SDM equations. The values in brackets for ρ represent the confidence intervals at 95%.

	Q.petraea		*Q.pubescens*		*F.sylvatica*		*A.alba*	
Variables	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.
$(Intercept)$	−1.155	−1.036	−1.347	−1.972	−1.117	0.171	−1.112	−0.377
	(0.052)	(0.009)	(0.049)	(0.012)	(0.053)	(0.053)	(0.053)	(0.079)
$PCC1$	−0.064	−0.188	−0.074	0.545	−0.067	−0.339	−0.064	−0.339
	(0.003)	(0.006)	(0.003)	(0.01)	(0.004)	(0.018)	(0.004)	(0.023)
$I(PCC1^2)$	−0.017	−0.126	−0.016	−0.112	−0.016	−0.034	−0.017	−0.009
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.005)
$PCC2$	0.186	0.079	0.197	0.174	0.193	0.043	0.189	0.199
	(0.006)	(0.009)	(0.006)	(0.009)	(0.006)	(0.014)	(0.006)	(0.024)
$I(PCC2^2)$	−0.079	−0.159	−0.083	−0.303	−0.081	−0.034	−0.079	−0.05
	(0.002)	(0.005)	(0.002)	(0.012)	(0.002)	(0.006)	(0.002)	(0.008)
$PCT1$	−0.238	−0.114	−0.237	−0.055	−0.238	0.018	−0.237	−0.035
	(0.005)	(0.008)	(0.005)	(0.005)	(0.005)	(0.017)	(0.005)	(0.025)
$I(PCT1^2)$	0	−0.005	0		−0.002	0.01	−0.001	−0.005
	(0.002)	(0.004)	(0.002)		(0.002)	(0.004)	(0.002)	(0.006)
$RT.FOR$	0.94		1.051		0.891		0.93	
	(0.027)		(0.026)		(0.028)		(0.027)	
$RT.AGR$	−0.126		−0.066		−0.105		−0.15	
	(0.037)		(0.033)		(0.038)		(0.037)	
$RT.POP$	−0.024		−0.01		−0.04		−0.031	
	(0.009)		(0.007)		(0.009)		(0.009)	
ρ	0.95	$[0.89, 0.98]$	0.96	$[0.89, 0.99]$	−0.75	$[-0.81, -0.66]$	−0.82	$[-0.87, -0.76]$
N	134.328	42.173	134.328	42.173	134.328	42.173	134.328	42.173
R^2	0.0911	0.1266	0.0911	0.0532	0.0911	0.1499	0.0911	0.2272

Table 4: **Raw GAM results at 2 km resolution.**
For the variables that enter non-parametrically, we report the χ^2 values corresponding to the joint significance of the associated terms. For the variables that enter parametrically, we report the coefficients and standard errors, as above. The values in brackets for ρ represent the ionfidence intervals at 95%.

	Q.Petraea		*Q.pubescens*		*F.sylvatica*		*A.alba*	
Variables	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.
$s(PCC1)$	1053.172	2285.084	1915.397	1326.27	1034.575	567.221	1006.165	317.979
$s(PCC2)$	1101.953	1517.165	2300.703		1308.538	149.44	1283.092	171.417
$s(PCT1)$	2726.503	304.577	2167.857		3082.75	164.49	3058.386	58.76
$RT.FOR$	1.016		1.104		0.91		0.929	
	(0.0286)		(0.0274)		(0.0307)		(0.0306)	
$RT.AGR$	−0.207		−0.21		−0.061		−0.085	
	(0.0392)		(0.0373)		(0.0411)		(0.0406)	
$RT.POP$	−0.041		−0.007		−0.047		−0.044	
	(0.00892)		(0.00784)		(0.00928)		(0.0092)	
ρ	0.89	$[0.81, 0.93]$	0.76	$[0.71, 0.8]$	−0.44	$[-0.61, -0.25]$	−0.54	$[-0.68, -0.35]$
N	134.328	42.173	134.328	42.173	134.328	42.173	134.328	42.173
R^2	0.0981	0.1318	0.0908	0.0449	0.0997	0.1553	0.0997	0.2396

Table 5: **Raw GLM results at 4 km resolution.**
For the 4 tree species of interest (in columns) we report the estimated coefficients and standard errors (in parenthesis) for the two equations: the land-use choice and the SDM equations. The values in brackets for ρ represent the confidence intervals at 95%.

	Q.petraea		*Q.pubescens*		*F.sylvatica*		*A.alba*	
Variables	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.
$(Intercept)$	−0.588	−0.282	−0.806	−1.504	−0.599	−0.096	−0.468	−1.038
	(0.115)	(0.017)	(0.111)	(0.019)	(0.119)	(0.029)	(0.118)	(0.064)
$PCC1$	−0.07	−0.202	−0.081	0.747	−0.076	−0.492	−0.069	−0.504
	(0.007)	(0.009)	(0.007)	(0.017)	(0.007)	(0.015)	(0.007)	(0.023)
$I(PCC1^2)$	−0.023	−0.156	−0.022	−0.145	−0.021	−0.033	−0.023	−0.019
	(0.002)	(0.005)	(0.002)	(0.004)	(0.002)	(0.005)	(0.002)	(0.007)
$PCC2$	0.25	0.083	0.264	0.081	0.267	0.226	0.259	0.408
	(0.012)	(0.017)	(0.012)	(0.02)	(0.012)	(0.018)	(0.012)	(0.03)
$I(PCC2^2)$	−0.099	−0.167	−0.101	0.071	−0.103	−0.123	−0.1	−0.13
	(0.004)	(0.008)	(0.004)	(0.006)	(0.004)	(0.008)	(0.004)	(0.01)
$PCT1$	−0.35	−0.118	−0.344	−0.384	−0.345	−0.062	−0.344	−0.174
	(0.012)	(0.017)	(0.012)	(0.019)	(0.012)	(0.023)	(0.012)	(0.041)
$I(PCT1^2)$	0.008	0.007	0.005	−0.056	0.004	0.028	0.008	0.005
	(0.005)	(0.008)	(0.005)	(0.009)	(0.006)	(0.007)	(0.006)	(0.011)
$RT.FOR$	1.273		1.511		1.317		1.298	
	(0.064)		(0.062)		(0.066)		(0.065)	
$RT.AGR$	−0.03		−0.066		−0.049		−0.149	
	(0.078)		(0.074)		(0.08)		(0.079)	
$RT.POP$	−0.051		−0.025		−0.063		−0.059	
	(0.015)		(0.011)		(0.015)		(0.015)	
ρ	0.8	$[0.59, 0.92]$	0.94	$[0.79, 0.98]$	−0.6	$[-0.7, -0.49]$	−0.6	$[-0.71, -0.46]$
N	33.626	21.927	33.626	21.927	33.626	21.927	33.626	21.927
R^2	0.156	0.1593	0.156	0.2855	0.156	0.2206	0.156	0.2869

Table 6: **Raw GAM results at 4 km resolution.**
For the variables that enter non-parametrically, we report the χ^2 values corresponding to the joint significance of the associated terms. For the variables that enter parametrically, we report the coefficients and standard errors, as above. The values in brackets for ρ represent the confidence intervals at 95%.

	Q.Petraea		Q.pubescens		F.sylvatica		A.alba	
Variables	Sel. Eq.	S.D.M.	Sel. Eq.	S.D.M.	Sel. Eq.	S.D.M.	Sel. Eq.	S.D.M.
$s(PCC1)$	370.667	2159.439	632.72	1282.201	342.545	821.763	323.918	1049.308
$s(PCC2)$	488.29	968.408	617.647		505.011	288.411	476.171	327.08
$s(PCT1)$	1033.99	205.993	1068.834		1012.308	70.453	1044.651	141.251
$RT.FOR$	1.408		1.68		1.385		1.358	
	(0.066)		(0.0686)		(0.0711)		(0.0716)	
$RT.AGR$	−0.287		−0.18		−0.149		−0.216	
	(0.0797)		(0.0818)		(0.0836)		(0.0855)	
$RT.POP$	−0.065		−0.039		−0.077		−0.072	
	(0.0145)		(0.0128)		(0.0152)		(0.0153)	
ρ	0.98	$[-0.98, 1]$	0.78	$[0.71, 0.83]$	−0.66	$[-0.76, -0.54]$	−0.22	$[-0.42, 0.02]$
N	33.626	21.927	33.626	21.927	33.626	21.927	33.626	21.927
R^2	0.1618	0.1753	0.1618	0.0779	0.162	0.2202	0.162	0.3012

Table 7: **Raw GLM results at 8 km resolution.**
For the 4 tree species of interest (in columns) we report the estimated coefficients and standard errors (in parenthesis) for the two equations: the land-use choice and the SDM equations. The values in brackets for ρ represent the confidence intervals at 95%.

	Q.petraea		*Q.pubescens*		*F.sylvatica*		*A.alba*	
Variables	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.
$(Intercept)$	0.666	0.242	0.093	−0.664	0.766	−0.109	0.466	−1.159
	(0.417)	(0.022)	(0.34)	(0.024)	(0.364)	(0.022)	(0.386)	(0.03)
$PCC1$	0.02	−0.153	−0.006	0.58	0.001	−0.413	0.001	−0.351
	(0.017)	(0.013)	(0.012)	(0.016)	(0.013)	(0.012)	(0.013)	(0.015)
$I(PCC1^2)$	−0.015	−0.079	−0.01	−0.073	−0.013	−0.031	−0.013	−0.021
	(0.003)	(0.007)	(0.002)	(0.001)	(0.003)	(0.005)	(0.003)	(0.006)
$PCC2$	0.185	−0.163	0.169	0.214	0.183	0.214	0.179	0.234
	(0.023)	(0.018)	(0.022)	(0.021)	(0.023)	(0.017)	(0.023)	(0.025)
$I(PCC2^2)$	−0.034	0.011	−0.033	−0.048	−0.038	−0.033	−0.037	−0.029
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)
$PCT1$	−0.264	−0.175	−0.275	−0.436	−0.301	−0.29	−0.28	−0.588
	(0.034)	(0.039)	(0.032)	(0.03)	(0.034)	(0.027)	(0.034)	(0.043)
$I(PCT1^2)$	−0.08	−0.067	−0.079	−0.039	−0.077	−0.121	−0.082	−0.126
	(0.015)	(0.017)	(0.014)	(0.016)	(0.014)	(0.012)	(0.014)	(0.016)
$RT.FOR$	1.644		1.897		1.352		1.594	
	(0.313)		(0.278)		(0.255)		(0.277)	
$RT.AGR$	−0.547		−0.334		−0.372		−0.341	
	(0.215)		(0.132)		(0.216)		(0.213)	
$RT.POP$	−0.121		−0.151		−0.116		−0.117	
	(0.03)		(0.004)		(0.03)		(0.03)	
ρ	−0.54	$[-0.87, 0.1]$	1	$[-1, 1]$	0.62	$[0.36, 0.8]$	0.08	$[-0.2, 0.34]$
N	8.426	7.638	8.426	7.638	8.426	7.638	8.426	7.638
R^2	0.0914	0.1363	0.0914	0.2594	0.0914	0.0701	0.0914	0.3362

Table 8: **Raw GAM results at 8 km resolution.**
For the variables that enter non-parametrically, we report the χ^2 values corresponding to the joint significance of the associated terms. For the variables that enter parametrically, we report the coefficients and standard errors, as above. The values in brackets for ρ represent the confidence intervals at 95%.

	Q.Petraea		*Q.pubescens*		*F.sylvatica*		*A.alba*	
Variables	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.	Land use	S.D.M.
$s(PCC1)$	66.101	524.234	940.136	200.721	69.622	1014.766	59.135	579.716
$s(PCC2)$	71.055	59.191	70.208		70.011	96.334	65.786	55.146
$s(PCT1)$	133.107	202.64	67.595		119.272	127.688	111.746	176.876
$RT.FOR$	1.608		0.296		1.989		1.837	
	(0.24)		(0.231)		(0.32)		(0.287)	
$RT.AGR$	−0.188		−0.438		−0.468		−0.683	
	(0.22)		(0.172)		(0.22)		(0.227)	
$RT.POP$	−0.106		−0.118		−0.11		−0.134	
	(0.0305)		(0.0202)		(0.0332)		(0.0322)	
ρ	0.97	$[-0.99, 1]$	−1	$[-1, 1]$	−0.75	$[-0.87, -0.54]$	−0.47	$[-0.7, -0.15]$
N	8.426	7.638	8.426	7.638	8.426	7.638	8.426	7.638
R^2	0.095	0.1453	0.095	0.1266	0.1	0.3358	0.1	0.3695

3.2 Spatial Covariances

3.2.1 For GLM

Table 9: **Decomposition of spatial covariance between ecological and economical latent variables for GLM.**
The formula used to decompose the total covariance between a observed term and an unobserved term is described in Section 1.4 of this SI file. The results below show the importance of the unobserved covariance to estimate the sign of the total spatial covariance.

		TOTAL		OBSERVED		UNOBSERVED	
Resol.	Species	ρ_T	CI 95%	ρ_O	CI 95%	ρ_U	CI 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	$[-0.12, 0.21]$

3.2.2 For GAM

Table 10: **Decomposition of spatial covariance between ecological and economical latent variables for GAM.**
The formula used to decompose the total covariance between a observed term and an unobserved term is described in Section 1.4 of this SI file. The results below show the importance of the unobserved covariance to estimate the sign of the total spatial covariance.

		TOTAL		OBSERVED		UNOBSERVED	
Resol.	Species	ρ_T	CI 95%	ρ_O	CI 95%	ρ_U	CI 95%
2 KM	Q.petraea	0.774	$[0.62, 0.9]$	0.110	$[0.02, 0.2]$	0.664	$[0.6, 0.7]$
	Q.pubescens	0.735	$[0.57, 0.89]$	0.113	$[-0.01, 0.23]$	0.622	$[0.58, 0.66]$
	F.sylvatica	-0.282	$[-0.4, -0.13]$	0.040	$[0.04, 0.04]$	-0.322	$[-0.44, -0.18]$
	A.alba	-0.249	$[-0.36, -0.1]$	0.104	$[0.09, 0.12]$	-0.353	$[-0.45, -0.22]$
4 KM	Q.petraea	0.777	$[-0.61, 0.87]$	0.119	$[0.03, 0.2]$	0.658	$[-0.65, 0.67]$
	Q.pubescens	0.762	$[0.55, 0.96]$	0.172	$[0.01, 0.33]$	0.590	$[0.54, 0.63]$
	F.sylvatica	-0.409	$[-0.51, -0.28]$	0.045	$[0.01, 0.09]$	-0.454	$[-0.52, -0.37]$
	A.alba	0.094	$[-0.09, 0.3]$	0.216	$[0.15, 0.28]$	-0.122	$[-0.24, 0.02]$
8 KM	Q.petraea	0.725	$[-0.63, 0.76]$	0.057	$[0.05, 0.07]$	0.668	$[-0.67, 0.69]$
	Q.pubescens	-0.742	$[-0.86, 0.98]$	0.060	$[-0.06, 0.18]$	-0.803	$[-0.8, 0.8]$
	F.sylvatica	-0.295	$[-0.44, -0.09]$	0.127	$[0.05, 0.2]$	-0.422	$[-0.49, -0.29]$
	A.alba	-0.074	$[-0.31, 0.23]$	0.199	$[0.09, 0.3]$	-0.272	$[-0.41, -0.07]$

3.3 Consistency of predictions

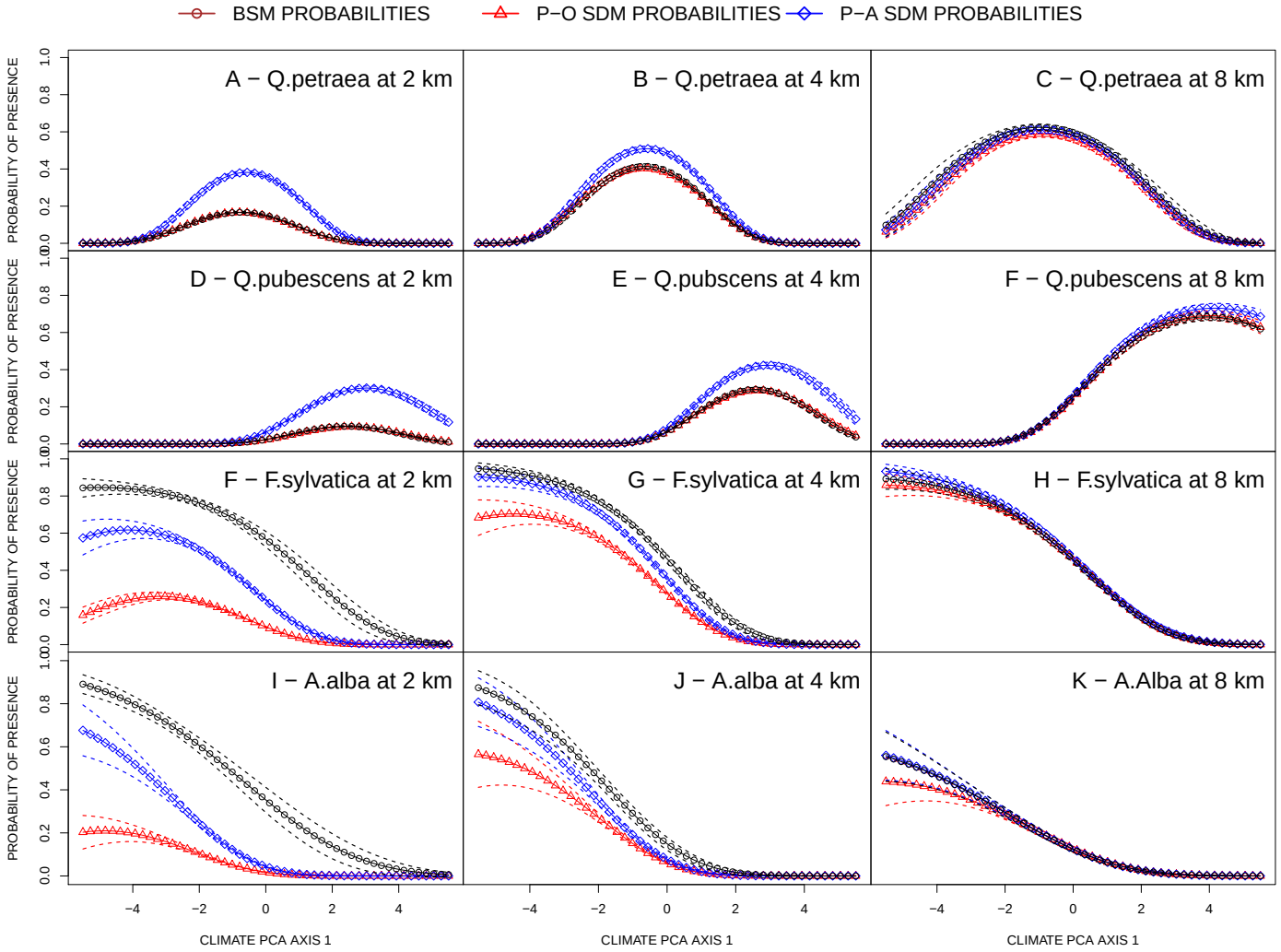
Table 11: **Consistency between predictions of potential presence.**
For each species and and scale, we report the root mean of squared errors (RMSE) of the predicted probabilities relatively to the the BSM predictions at the finest 2 km scale. This Table allows to evaluate the importance of taking into account the selection bias in fine scale models. For *F.sylvatica* and *A.alba*, the BSM taking into account the land-use selection bias at 4 km produces better prediction that classical SDM at 2km.

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

3.4 GLM response curves

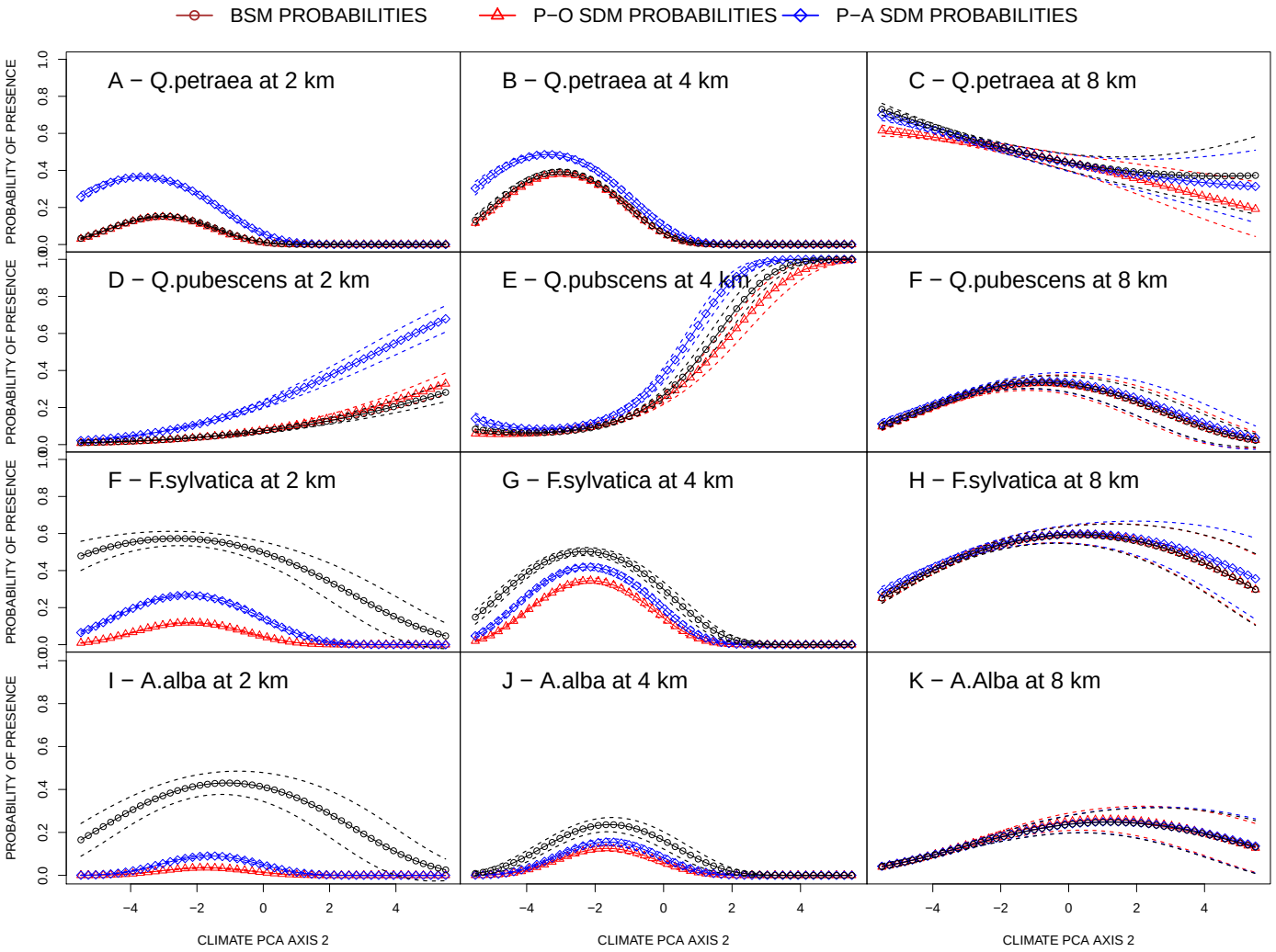
3.4.1 For the Climate PCA 1

Figure 3: **GLM response curves for Climate PCA Axis 1**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Climate PCA 1 axis here (we interpret it as an index of water availability, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



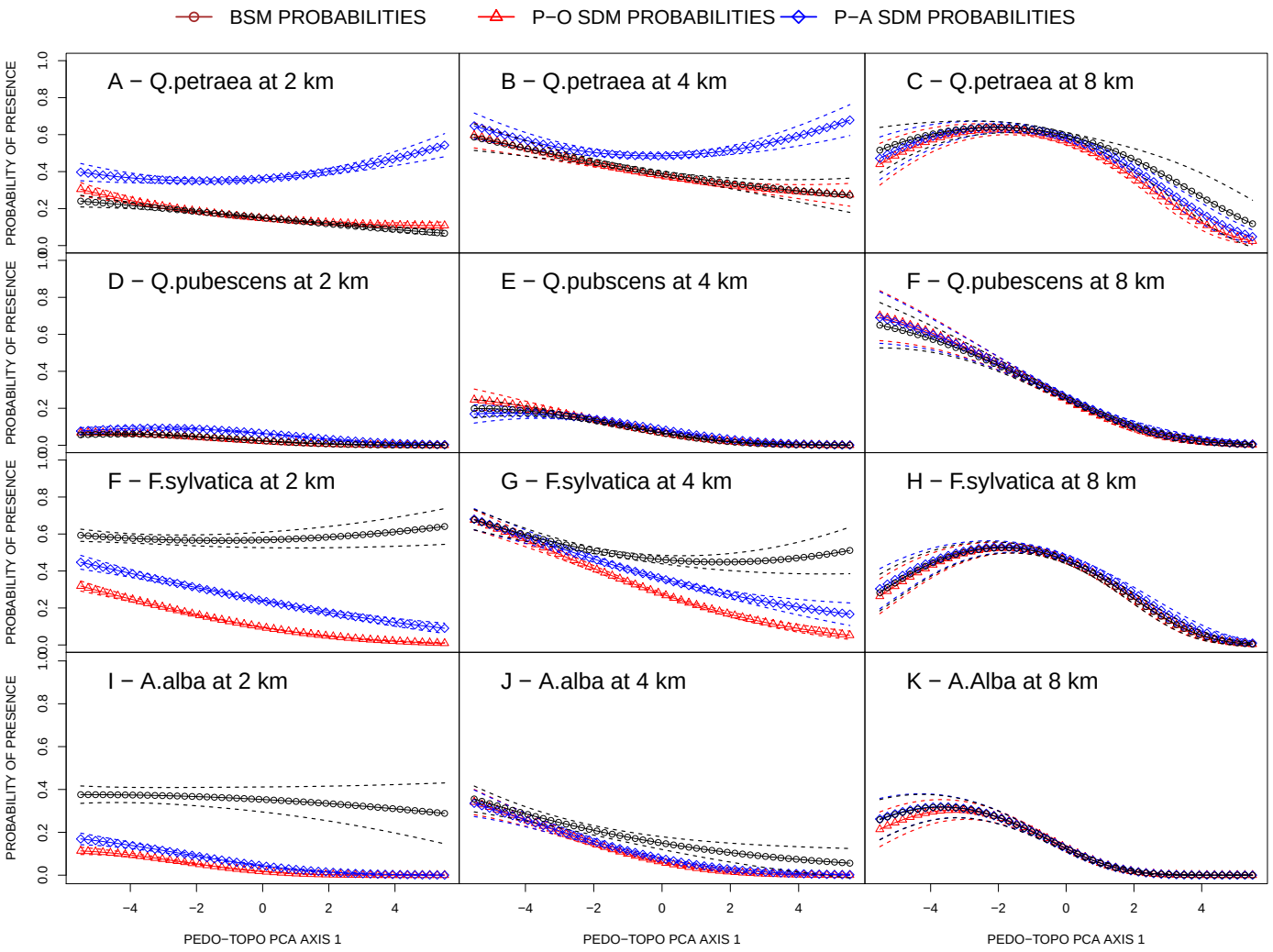
3.4.2 For the Climate PCA 2

Figure 4: **GLM response curves for Climate PCA Axis 2**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Climate PCA 2 axis here (we interpret it as an index of climate aridity, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



3.4.3 For the Topo PCA 1

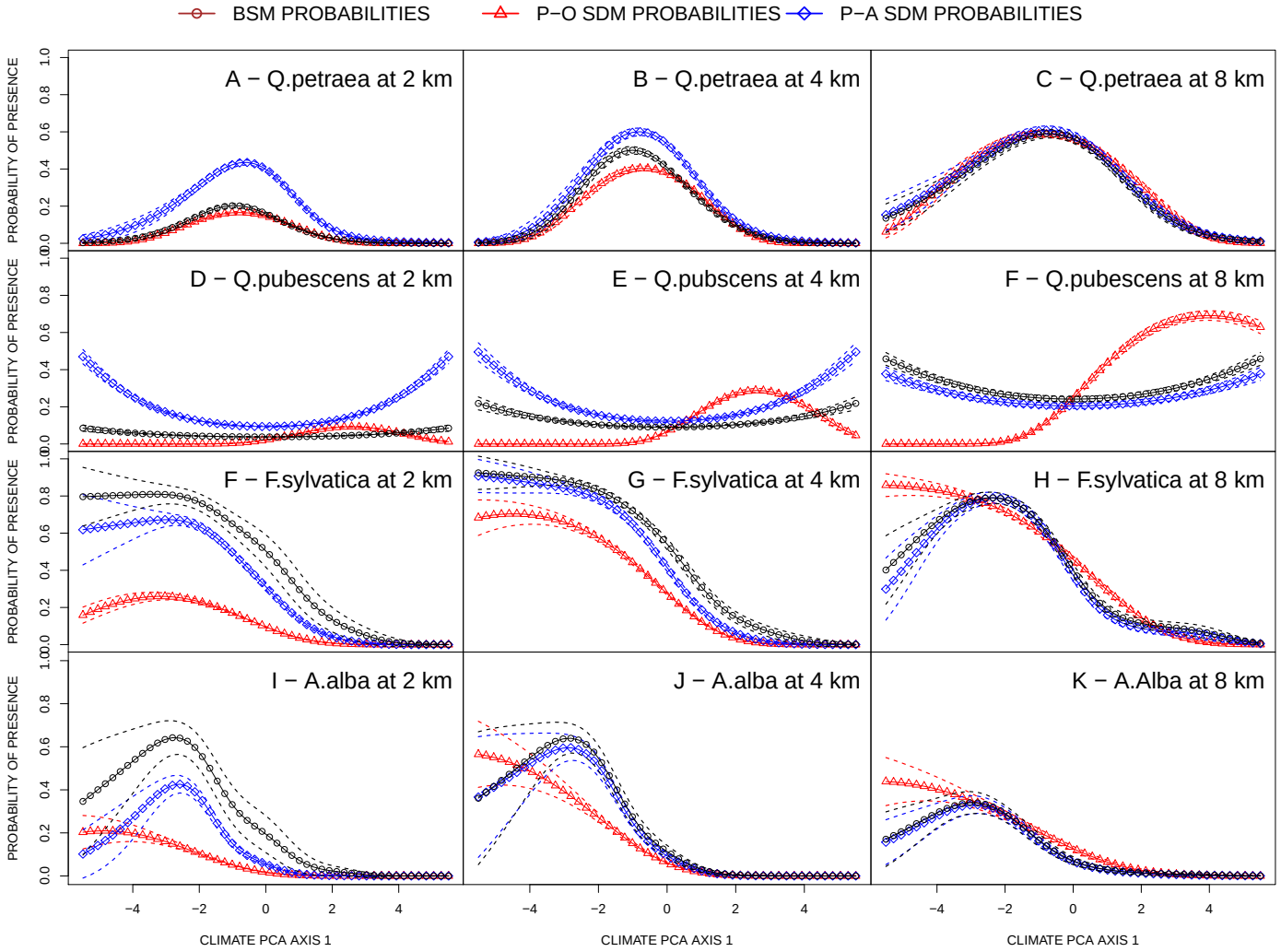
Figure 5: **GLM response curves for Topography PCA Axis 1**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Topography PCA 1 axis here (we interpret it as an index of flatness, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



3.5 GAM response curves

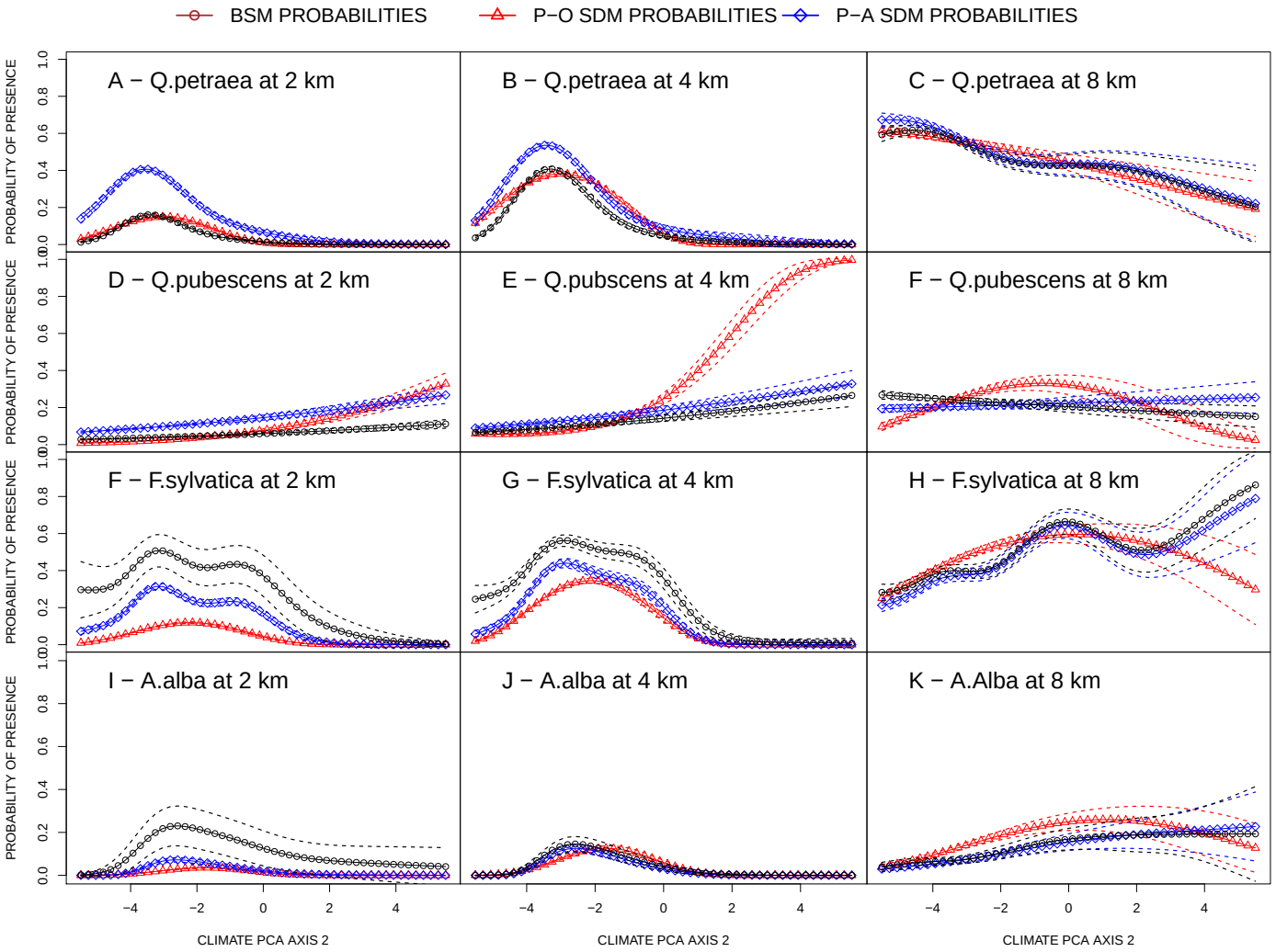
3.5.1 For the Climate PCA 1

Figure 6: **GAM response curves for Climate PCA Axis 1**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Climate PCA 1 axis here (we interpret it as an index of water availability, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



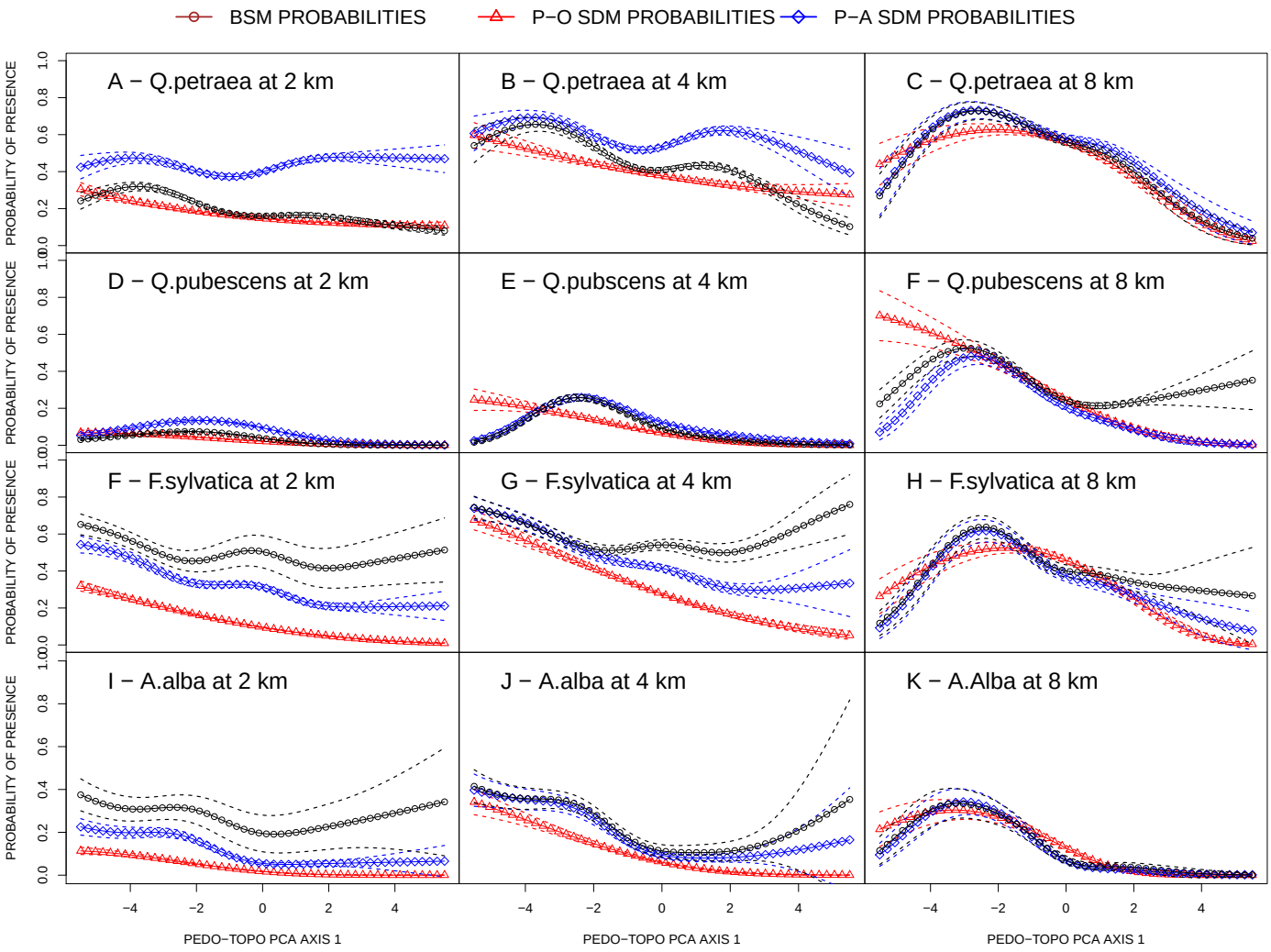
3.5.2 For the Climate PCA 2

Figure 7: **GAM response curves for Climate PCA Axis 2**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Climate PCA 2 axis here (we interpret it as an index of climate aridity, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



3.5.3 For the Topo PCA 1

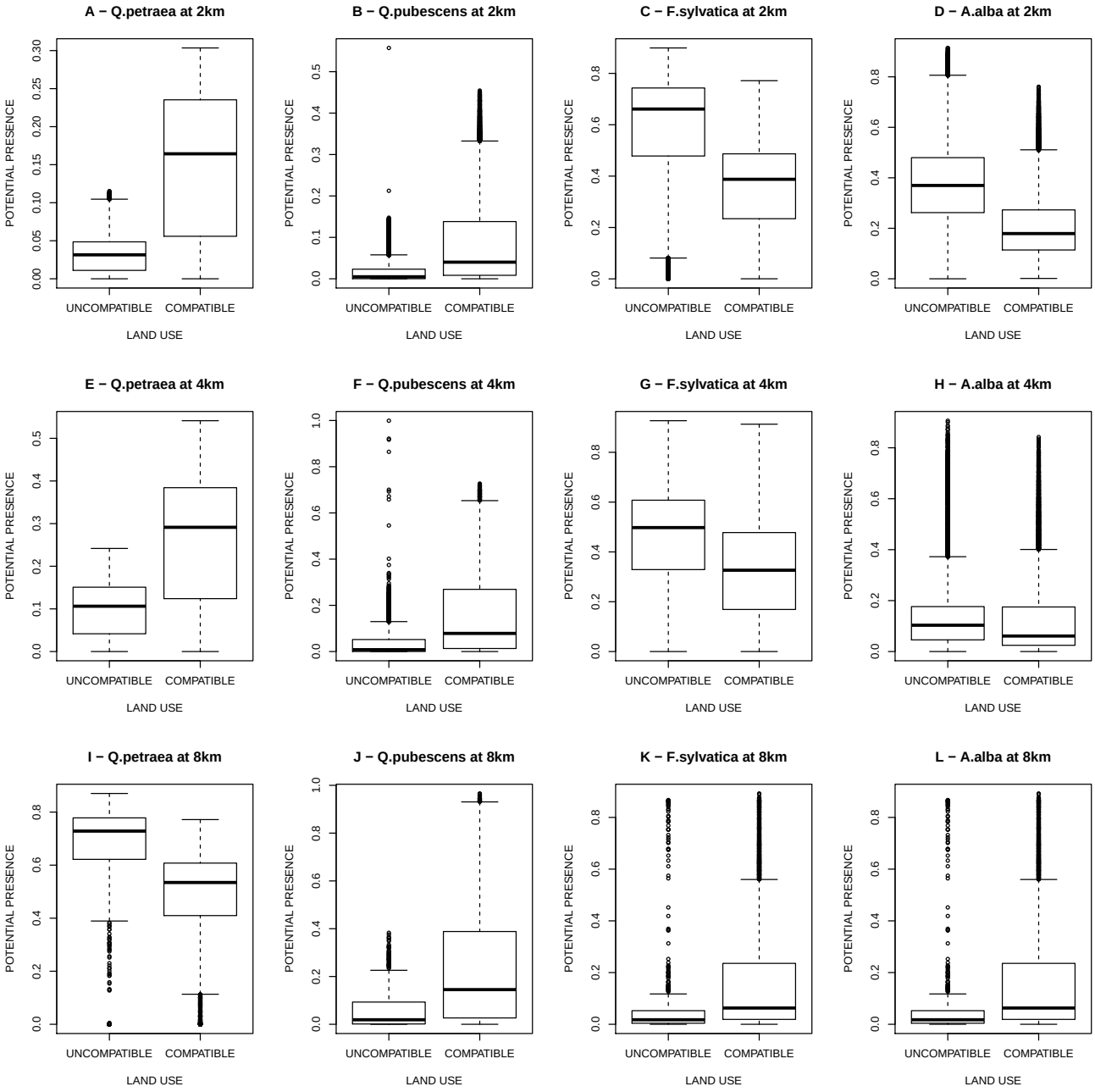
Figure 8: **GAM response curves for Topography PCA Axis 1**
Response curves are predicted probabilities of tree species presence with all the covariates fixed at their sample means except the covariate of interest, the Topography PCA 1 axis here (we interpret it as an index of flatness, see Figure 2 of the SI). Response curves below are differentiated according to the model used (the three curves), the tree species (from the top to the bottom) and the scale of the data (from the left to the right).



3.6 Potential / Effective

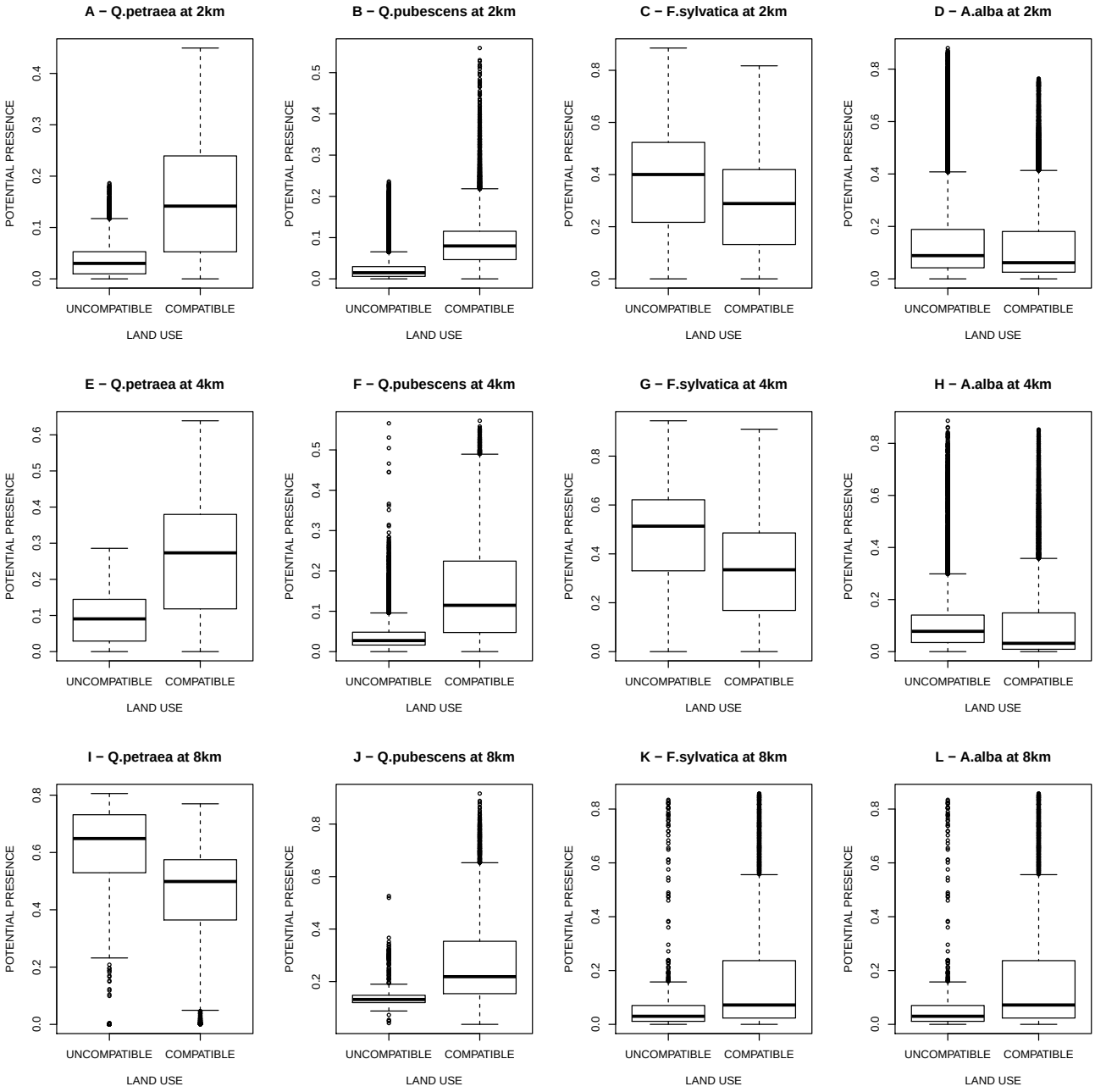
3.6.1 For GLM

Figure 9: **Probabilities of potential presence according to land-use compatibility for GLM**



3.6.2 For GAM

Figure 10: **Probabilities of potential presence according to land-use compatibility for GAM**



3.7 Maps of predicted GLM probabilities

Figure 11: **Maps of predicted GLM probabilities for *Q.petraea***

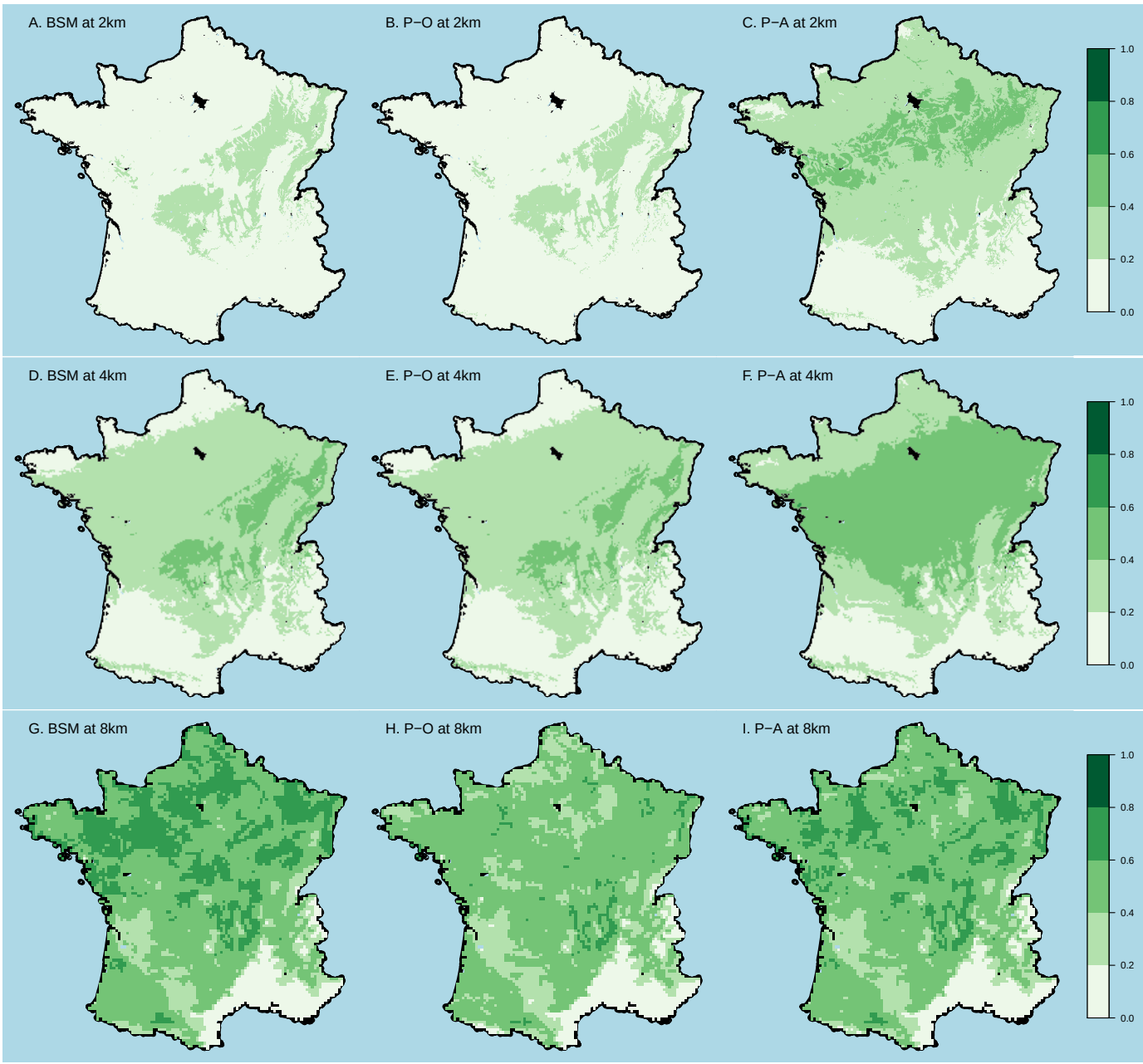


Figure 12: **Maps of predicted GLM probabilities for *Q.pubescens***

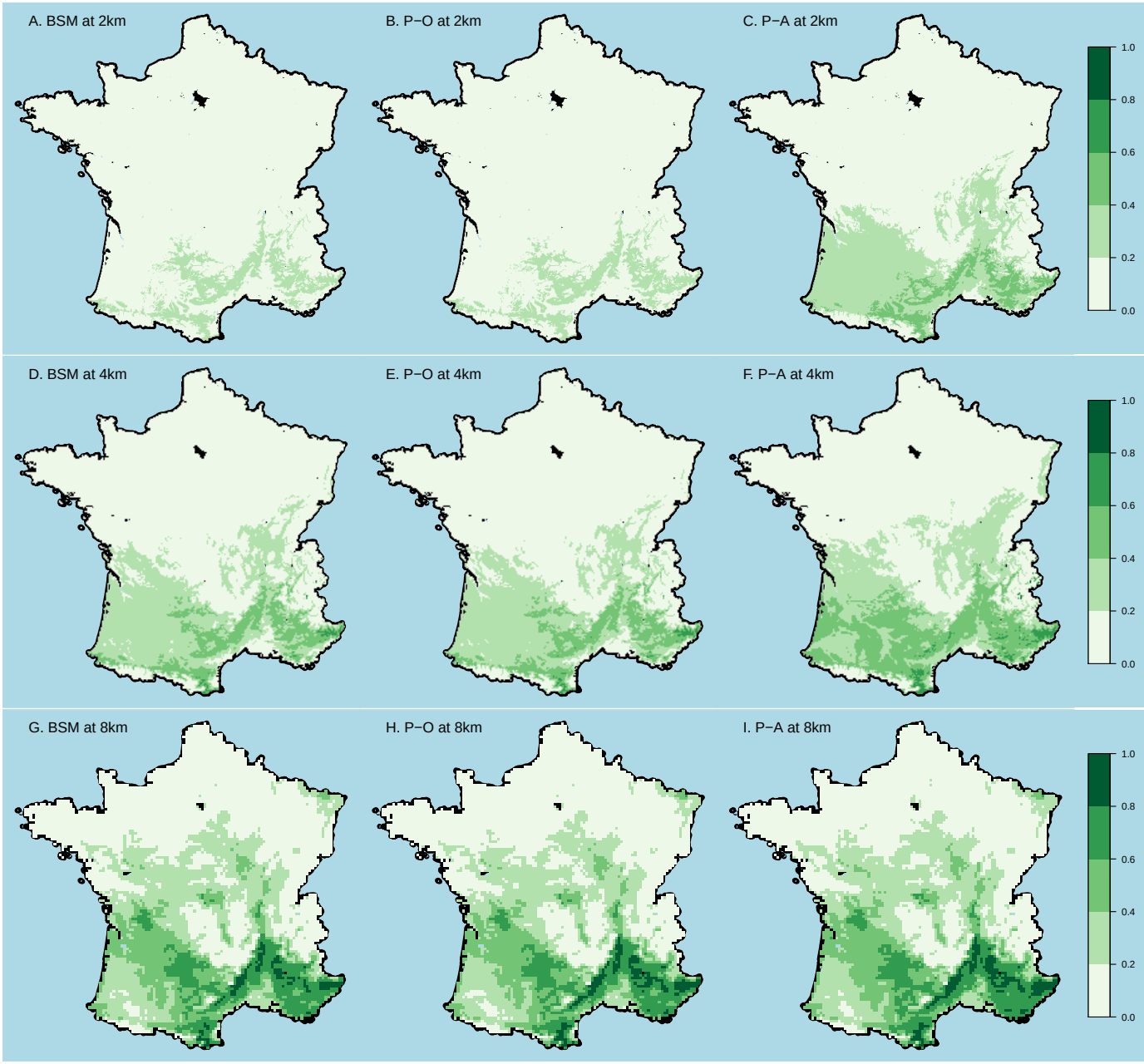


Figure 13: **Maps of predicted GLM probabilities for *F.sylvatica***

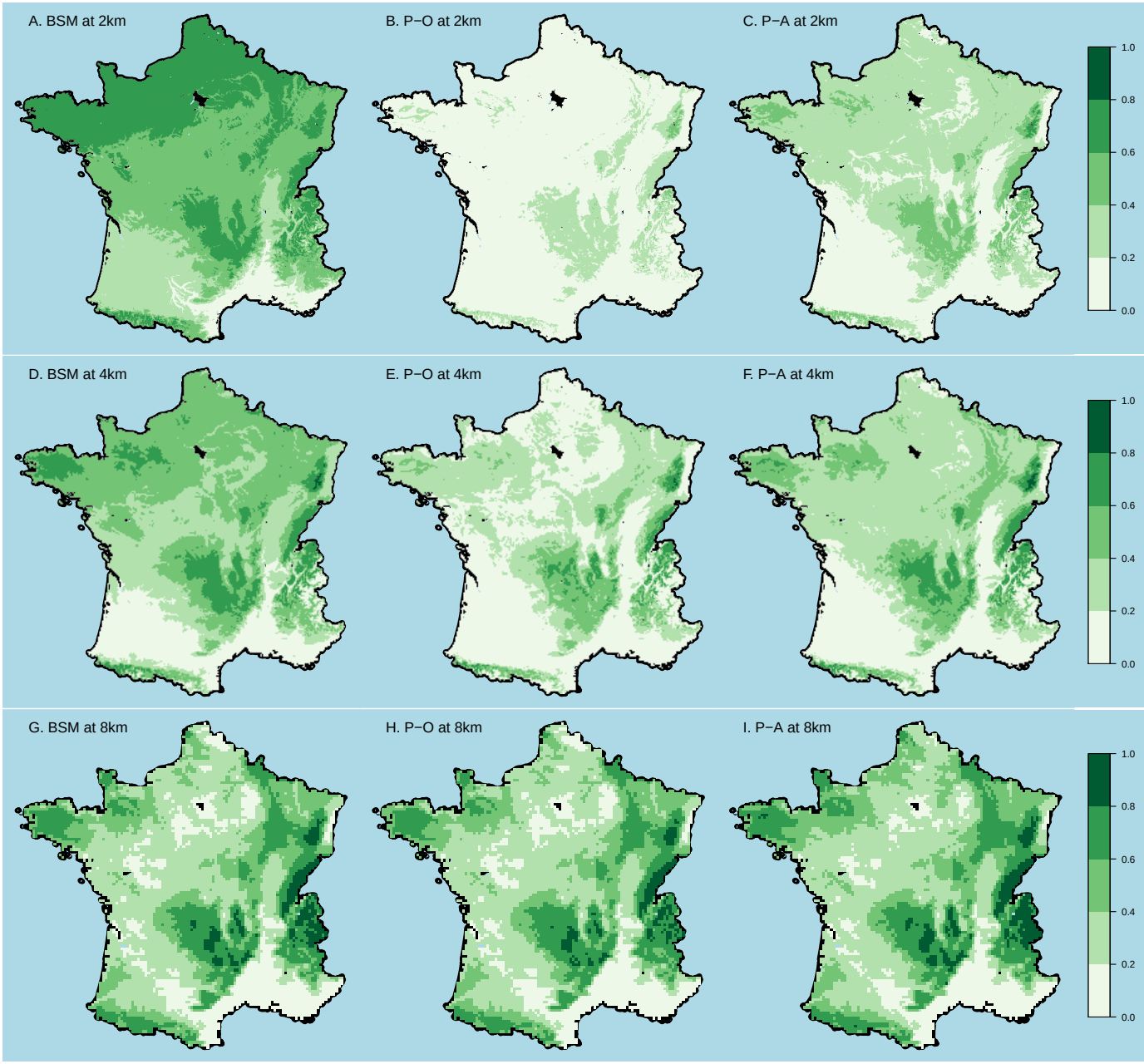
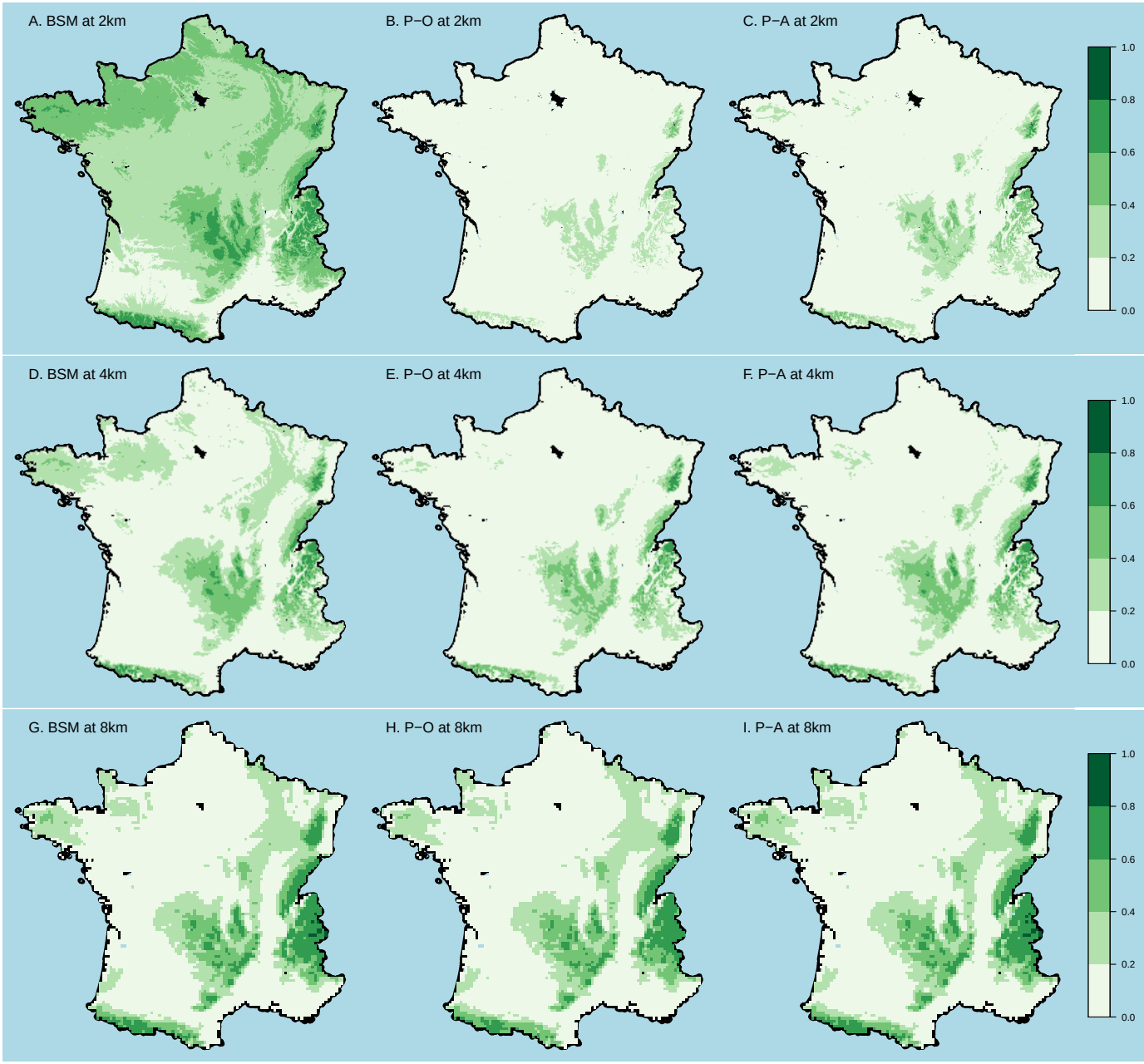


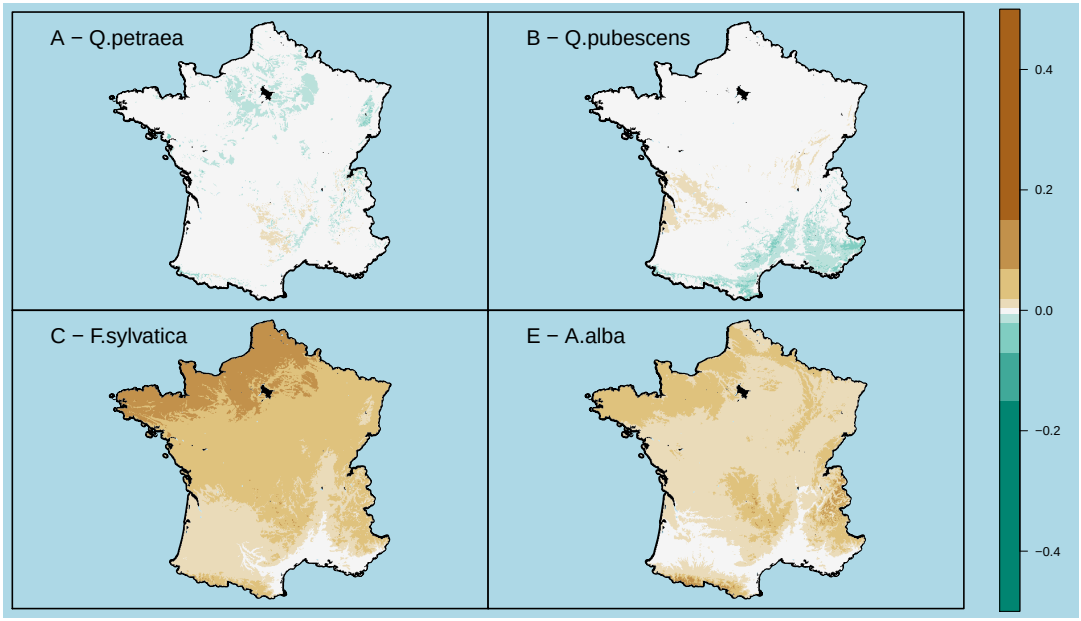
Figure 14: **Maps of predicted GLM probabilities for *A.alba***



3.8 Maps of bias in predicted probabilities

Figure 15: **Bias from predicted probabilities of presence from classical SDMs at 2 km.** The bias are computed as the difference between the probability predicted by the BSM and the probability predicted by the classical SDM. Positive bias (green colors) indicates that classical SDMs over-estimate the probability of presence. Negative bias (orange colors) indicates that classical SDMs under-estimate the probability of presence.

(a) Bias from classical PO SDMs



(b) Bias from classical PA SDMs

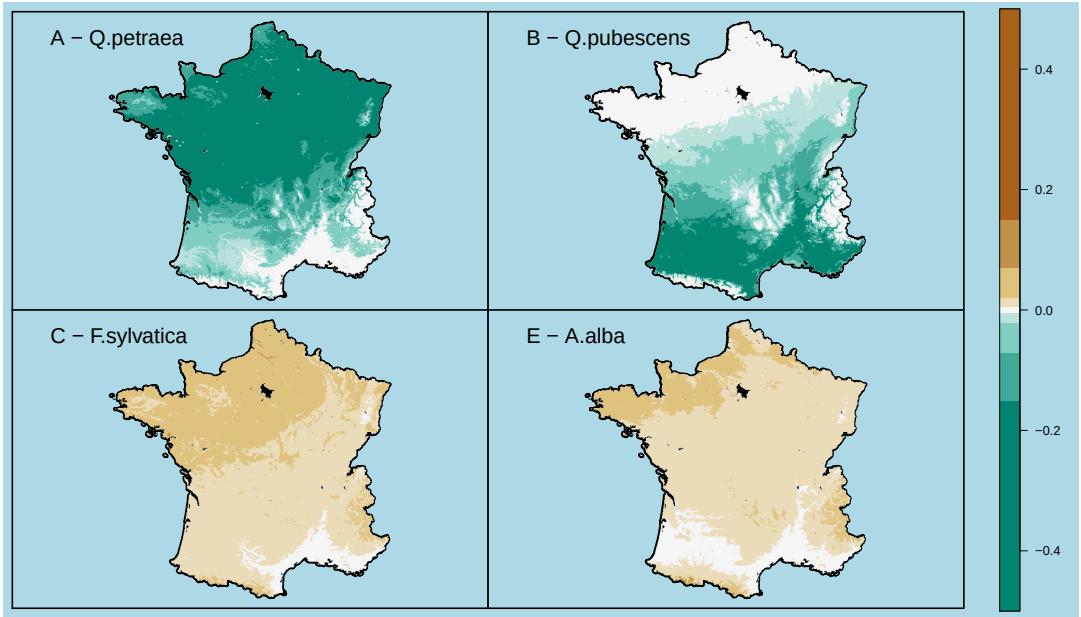
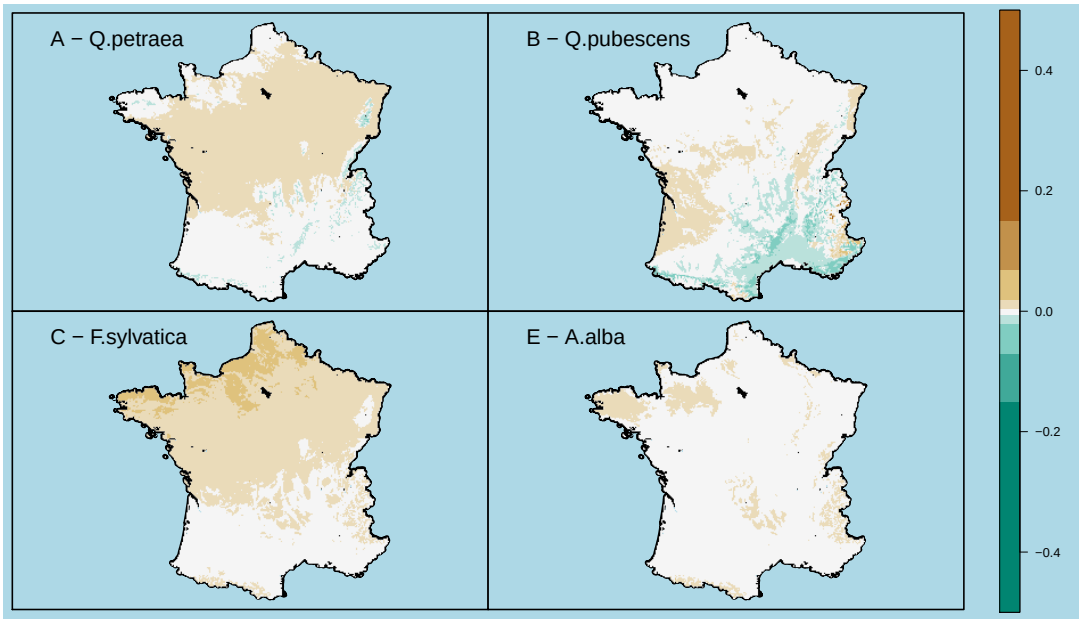


Figure 16: **Bias from predicted probabilities of presence from classical SDMs at 4 km.** The bias are computed as the difference between the probability predicted by the BSM and the probability predicted by the classical SDM. Positive bias (green colors) indicates that classical SDMs over-estimate the probability of presence. Negative bias (orange colors) indicates that classical SDMs under-estimate the probability of presence.

(a) Bias from classical PO SDMs



(b) Bias from classical PA SDMs

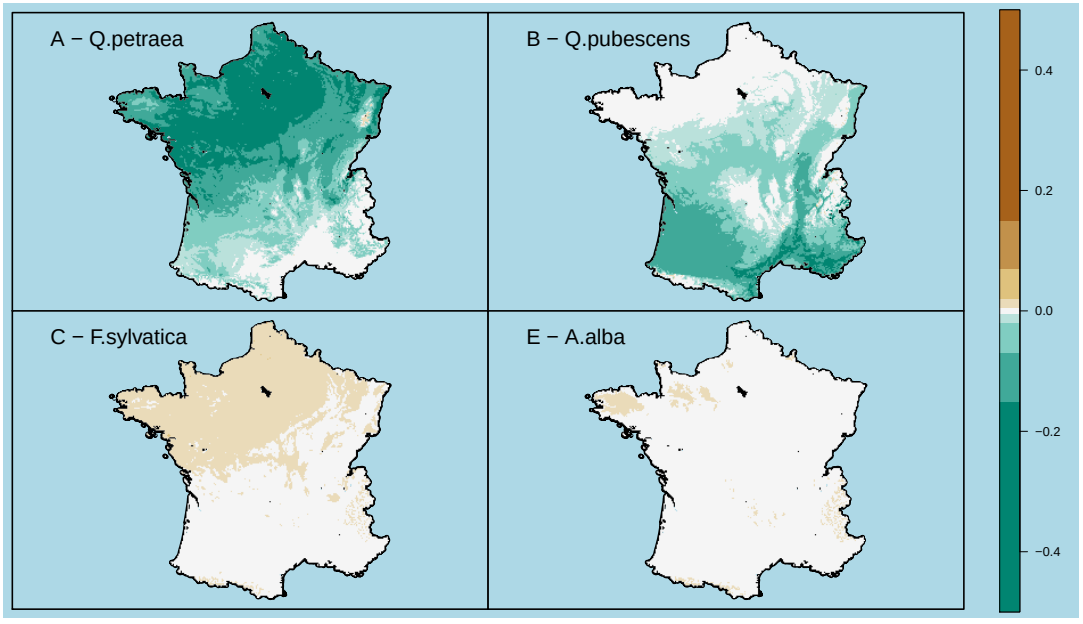
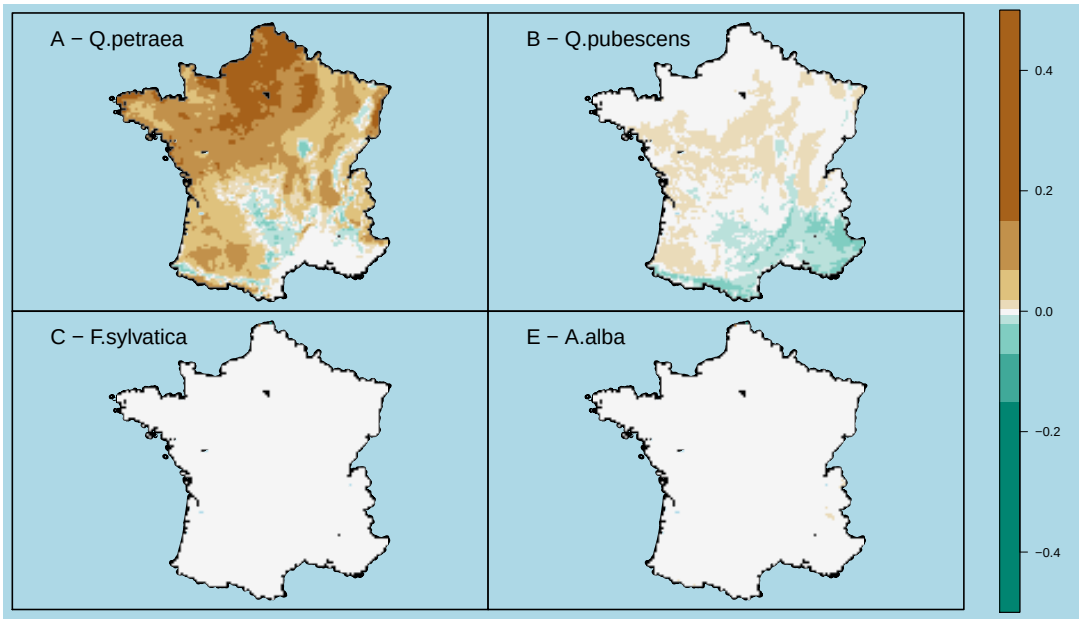
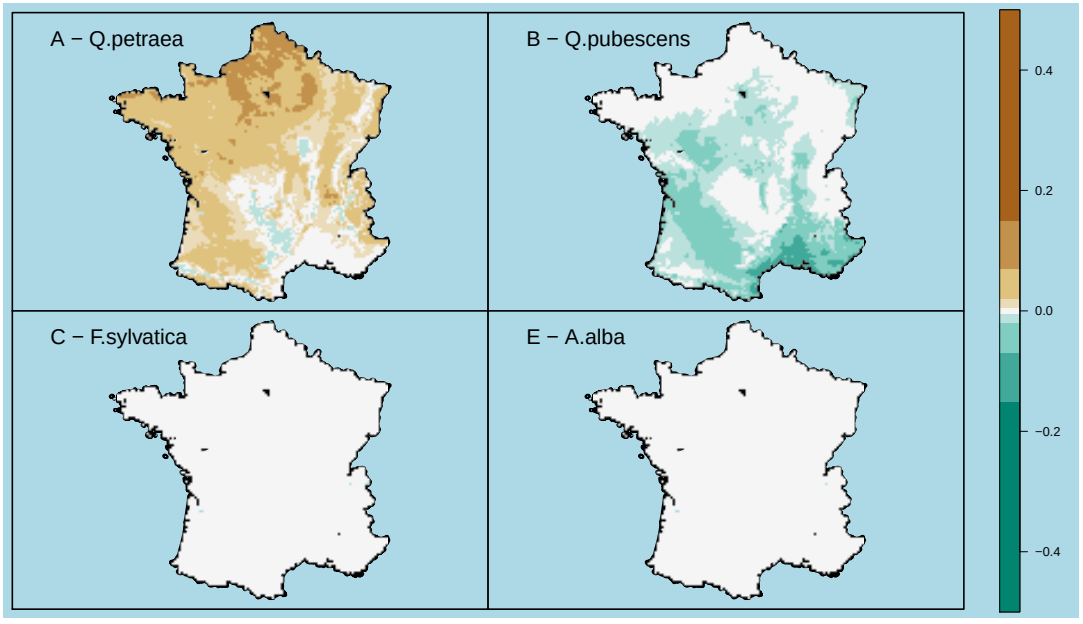


Figure 17: **Bias from predicted probabilities of presence from classical SDMs at 8 km.** The bias are computed as the difference between the probability predicted by the BSM and the probability predicted by the classical SDM. Positive bias (green colors) indicates that classical SDMs over-estimate the probability of presence. Negative bias (orange colors) indicates that classical SDMs under-estimate the probability of presence.

(a) Bias from classical PO SDMs



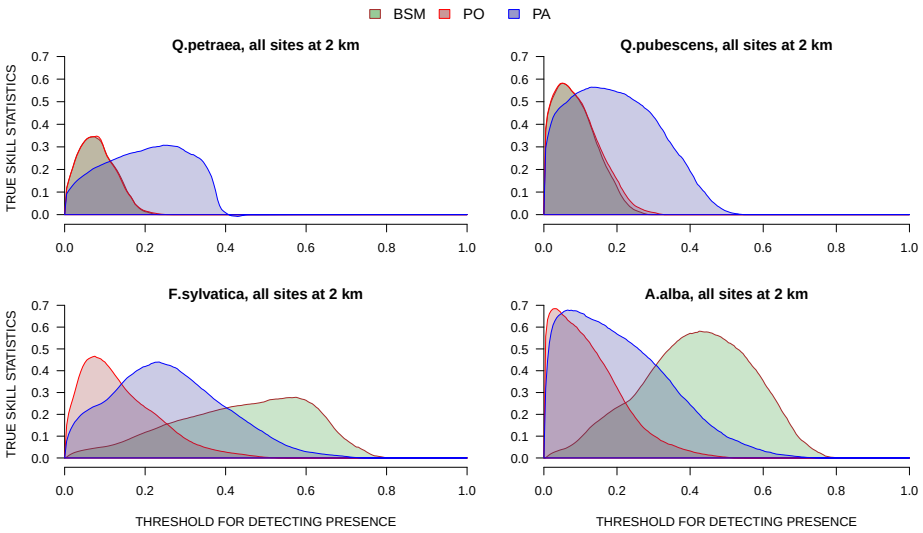
(b) Bias from classical PA SDMs



3.9 Validation of predictions

Figure 18: **True Skill Statistics (TSS) from internal (IFN) and external data (EuroVegMap, EVM).**
TSS takes into account both omission and commission errors, and success as a result of random guessing, and ranges from -1 to +1, where +1 indicates perfect agreement and values of zero or less indicates a performance no better than random.

(a) Predictions' TSS compared to IFN data



(b) Predictions' TSS compared to EVM data