Tesis Doctoral

# Muestreo de Pseudo–ausencias en Modelos de Distribución de Especies y Transferibilidad en Condiciones de Cambio Climático

Doctorado en Agrobiología Ambiental

Departamento de Biología Vegetal y Ecología

Presentada por
Maialen Iturbide Martínez de Albéniz

bajo la dirección de
Dr. José Manuel Gutiérrez Llorente y Dr. Miriam Pinto Tobalina

Euskal Herriko Unibertsitatea

Mayo de 2017

*Portada: Distribución potencial de la filogenia H7 del roble común en Europa. Contraportada: Distribución potencial de la misma filogenia en condiciones futuras de cambio climático.*

PhD Thesis

# Background sampling in Species Distribution Models and Transferability in Climate Change Conditions

PhD in Environmental Agrobiology

Department of Plant Biology and Ecology

Presented by
Maialen Iturbide Martínez de Albéniz

under the supervision of
Dr. José Manuel Gutiérrez Llorente and Dr. Miriam Pinto Tobalina

Euskal Herriko Unibertsitatea

May 2017

*Cover: Potential distribution of Oak H7 phylogeny in Europe.*

*Back cover: Potential distribution in future climate conditions.*

# Esker ematea
# Acknowledgements
# Agradecimientos

First of all, I want to express my most sincere gratitude to all my colleagues at the Santander Meteorology Group. Specially to my director Dr. José Manuel Gutiérrez for giving me the opportunity to join the group and for his dedication as supervisor of this Thesis. I would also like to thank Dr. Joaquín Bedia for his skillful assistance and for sharing his knowledge and time with me.

I am very thankful to my colleagues at Neiker, specially to my director Dr. Miriam Pinto and to Oscar del Hierro for the trust placed in me and for giving me the opportunity to begin my doctoral studies.

I am grateful to Rémy Petit and François Ehrenmann for providing the phylogenetic distribution of *Quercus*. I acknowledge the fruitful discussions arisen in the the WG1 of the FPS COST Action FP1202 (MaP-FGR, "Strengthening conservation: a key issue for adaptation of marginal/peripheral populations of forest trees to climate change in Europe"). I am also grateful to the ENSEMBLES project (GOCE-CT-2003-505539), supported by the European Commission's

6th Framework Programme for providing publicly the RCM simulations used in this study. This work was carried out at Neiker-Tecnalia and the Institute of Physics of Cantabria, and was supported by the EC-funded project ADAPTA-CLIMA II (INTERREG IVB SUDOE Program) and the EC-funded project INTACT (FP7-SEC-2013-1).

Quiero dedicar mi Tesis a mi familia y amigos, en especial a aquellos que han sido estudiantes de doctorado a la vez que yo: Mi hermana Ane y mis colegas Deiene, Nebai, Aritz, Ania, Maialen, Maite, Urtzi y Damaris.

Esker anitz gurasoei, Donostia, Bilbo, Gasteiz eta Santanderreko lagunei, Kuadrupediari, Damarisi, Nebairi, Aritzi eta Violetari.

Zuek denoi bihotzez, eskerrik beroenak.

Maialen Iturbide Martínez de Albéniz
Santander, 16 de Mayo de 2017

# Contents

# Nomenclature

| | |
|---|---|
| AEMET | *Agencia Estatal de Meteorología* |
| AL | *Alps* |
| AOGCMs | *Atmosphere-Ocean General Circulation Models* |
| BI | *British Isles* |
| BIO | *Bioclimatic variable* |
| C4I | *Community Climate Consortium for Ireland* |
| CART | *Classification and Regression Trees* |
| CERA | *Climate and Environmental Retrieval and Archive* |
| CMIP | *Coupled Model Intercomparison Project* |
| CMIP3 | *Coupled Model Intercomparison Project Phase 3* |
| CMIP5 | *Coupled Model Intercomparison Project Phase 5* |
| CNRM | *Centre National de Recherches Meteorologiques* |
| CORDEX | *Coordinated Regional Downscaling Experiment* |
| DMI | *Danish Meteorological Institute* |

| | |
|---|---|
| E-OBS | *Eurpean Observational climate dataset* |
| EA | *Eastern Europe* |
| EC | *Expert Criteria* |
| ECA&D | *European Climate Assessment & Dataset* |
| ECMWF | *European Centre for Medium-Range Weather Forecasts* |
| EP | *Environmental Profiling* |
| ESGF | *Earth System Grid Federation* |
| ETHZ | *Swiss Institute of Technology* |
| FDR | *False Discovery Rate* |
| FN | *False Negatives* |
| FOR | *False Omission Rate* |
| FP | *False Positives* |
| FR | *France* |
| GARP | *Genetic Algorithm for Rule set Production* |
| GBIF | *Global Biodiversity Information Facility* |
| GCM | *Global Climate Model* |
| GD$^2$ | *Georeferenced Database of Genetic Diversity* |
| GIS | *Geographic Information System* |
| GLM | *Generalized Linear Models* |
| ICTP | *Abdus Salam International Centre for Theoretical Physics* |
| IP | *Iberian Peninsula* |

| | |
|---|---|
| IPCC | *Intergovernmental Panel on Climate Change* |
| IPCC-SRES | *Intergovernmental Panel on Climate Change, special report on emission scenarios* |
| KNMI | *Koninklijk Nederlands Meteorologisch Instituut* |
| MARS | *Multivariate Adaptive Regression Splines* |
| MAXENT | *Maximum Entropy* |
| MD | *Mediterranean* |
| ME | *Mid-Europe* |
| Met.NO | *The Norwegian Meteorological Institute* |
| MetoHC | *Hadley Center/UK Met Office* |
| MPI-M | *Max Planck Institute for Meteorology* |
| OCSVM | *One Class Support Vector Machines* |
| RCBC | *Regional Climate of the Basque Country* |
| RCM | *Regional Climate Model* |
| RCP | *Representative Concentration Pathway* |
| RF | *Random Forest* |
| RS | *Random Sampling of the whole study domain* |
| RSEP | *Random Sampling + Environmental profiling* |
| SC | *Scandinavia* |
| SDM | *Species Distribution Model* |
| SMHI | *Swedish Meteorological and Hydrological Institute* |
| SVM | *Support Vector Machines* |

| | |
|---|---|
| TG | *Target Group background method* |
| TN | *True Negatives* |
| TNR | *True Negative Rate* |
| TP | *True Positives* |
| TPR | *True Positive Rate* |
| TS | *Three–step method* |
| UCLM | *Universidad de Castilla la Mancha* |
| VIA | *Vulnerability, Impact and Adaptation Community* |
| WC | *WorldClim* |

# List of Figures

# List of Tables

# Part I

# Introduction

# CHAPTER 1

# Species Distribution Modeling

Species Distribution Models (SDMs), also known as Environmental Niche Models (ENMs), are statistical tools used for the generation of probabilistic predictions of the presence of biological entities in the geographical space (Guisan & Zimmermann, 2000; Elith & et al, 2006). SDMs operate through the establishment of an empirical link between known presence/absence locations (predictand) and the physical characteristics of their environment (predictors). A popular application of these models is the future projection of species distributions —from future climate projections— in order to assess key topics in environmental conservation such as monitoring biological responses to climate change (Hamann & Wang, 2006), species invasions (Jeschke & Strayer, 2008) or disease transmission (Drake & Beier, 2014) among others.

SDMs have become a valuable tool for the vulnerability and impact assessment community, as a means of estimating distribution shifts due to climate variations, a problem of current interest in environmental conservation studies (see e.g.: Araújo *et al.*, 2004; Hamann & Wang, 2006; Jeschke & Strayer, 2008; Felicísimo *et al.*, 2011). However, there are important sources of uncertainty that affect the credibility of future distribution estimates, such as SDM predictive ability outside the training period/spatial extent (known as SDM transferability in time/space; Fronzek *et al.*, 2011), uncertainties regarding the training data (Mateo *et al.*, 2010b; Bedia *et al.*, 2013), the assumptions underlying the different emission scenarios (Nakićenović, 2000), the global/regional climate model (GCM/RCM) biases (Turco *et al.*, 2013) and others (see e.g.: Falloon *et al.*, 2014, for an overview).

Therefore it is crucial to analyze the contribution of each source of uncertainty in future SDM projections in order to provide reliable estimates of species distributions under climate change conditions.

## 1.1 Species Distribution Models (SDMs)

In this Thesis we use the acronym SDM (Species Distribution Model) to refer to the modeling technique or algorithm used to characterize the ecological niche of a species population as a function of the presence/absence data (predictand, Sections 1.3 and 1.4) and a set of explanatory variables that characterize the environment of the species population (predictors, Section 1.5, Fig. 1.1). Two types of spatial data are required for model calibration: (1) occurrence data documenting presences (and sometimes absences) of a species population and (2) gridded data of the environmental variables (e.g. raster-format GIS layers). The spatial distribution of the environments suitable for the modeled population (a.k.a. suitability maps, Fig. 1.1) are then estimated by projecting (predicting) the built SDMs into the environmental data used for model calibration (reference suitability maps) or into an unsampled environment from other spatial domain (e.g. for estimating potential areas of species invasions, Jeschke & Strayer, 2008) or time period (e.g. for estimating habitat shifts due to climate change, Hamann & Wang, 2006). Depending on the modeling approach used, the resulting suitability maps can be probabilistic or deterministic predictions, this is, predictions of the probability of occurrence of a species population (values ranging from 0 to 1) versus those that directly predict suitable and

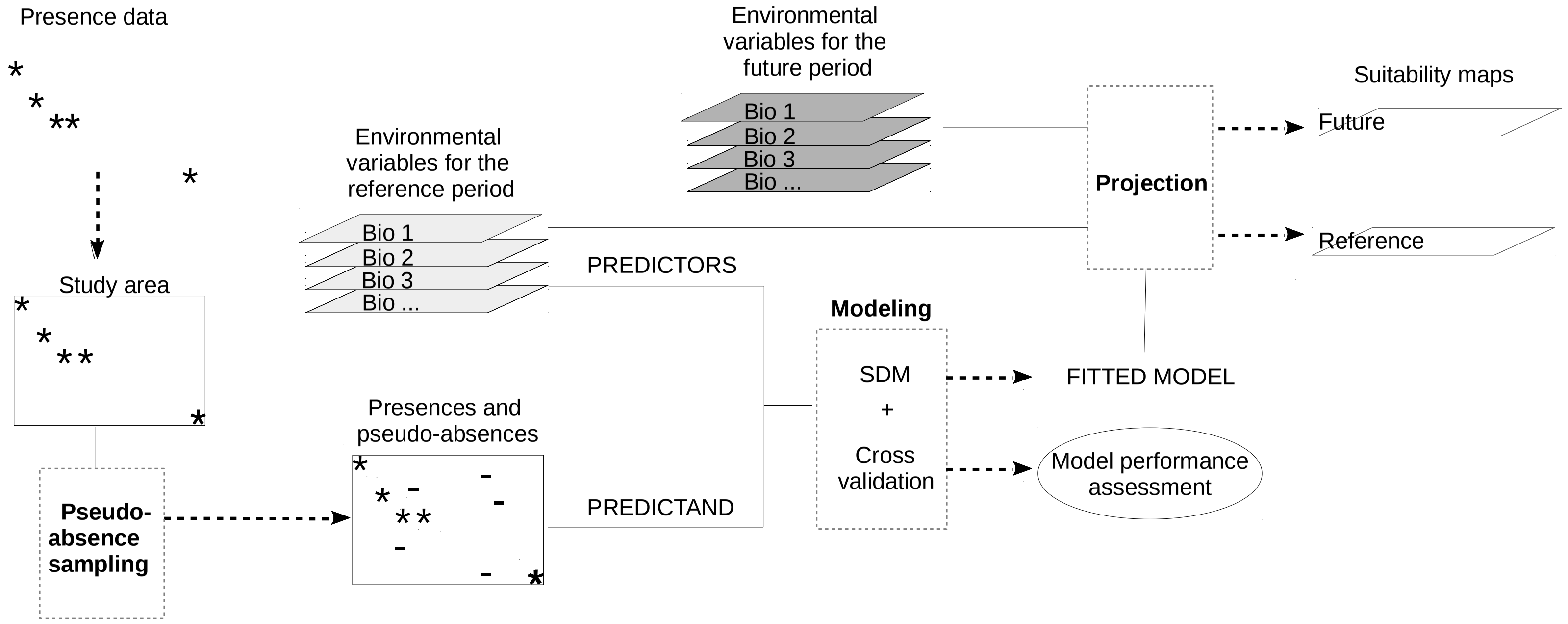

**Figure 1.1:** Conceptual diagram of the process for species distribution modeling and projection with pseudo–absence data.

unsuitable areas (1 and 0). The representation of this information in the form of geographical maps (distribution of the probabilities, Guisan & Thuiller, 2005), constitutes a clear advantage for planning territorial uses or for the management and conservation of species.

Probabilistic predictions have a number of advantages over deterministic predictions. The main advantage is that a probability of occurrence is a relevant information from the ecological point of view, since it allows to quantitatively evaluate the degree of suitability of a given habitat to house a particular species population. Additionally, for this type of applications, it is possible to generate different deterministic outputs of the models using different probability thresholds that best fit the pursued objectives (see e.g. Freeman & Moisen, 2008; Gude *et al.*, 2009).

SDM techniques can be broadly classified into two types: *profile* and *group discrimination* techniques. The first group refers to those modeling approaches that rely solely on known presences to infer the potential distribution of the species (a.k.a. presence–only algorithms), while group discrimination techniques require information of the environmental range where the species do not occur, that is, absence data. Group discrimination techniques have gained popularity in recent years, as they have been reported to yield better results than profile techniques (Engler *et al.*, 2004; Chefaoui & Lobo, 2008; Elith & et al, 2006; Mateo *et al.*, 2010a).

However, in part due to the great effort involved in true absence sampling, most of the available biodiversity datasets for predictive modeling (generally natural history collections, see. e.g. Araújo & Williams, 2000) are lacking explicit absence data. Thus, in most cases discrimination techniques are used, requiring the environmental characterization of the sites of presence in front of a background sample —also known as pseudo–absence data— that characterizes the available environment in the study region (see Section 1.4).

## 1.2 Common SDM Techniques

A number of techniques used for binomial regression and classification constitute the benchmark for modeling species distributions (Muñoz & Felicísimo, 2004; Terribile *et al.*, 2010; Mateo *et al.*, 2011) and building ensembles of the results derived from multiple SDMs (Araújo & New, 2007), ranging from simple and parsimonius Generalized Lineas Models (GLMs) to more complex nonlinear techniques, such as MARS or Random Forest. The most commonly used techniques are described below.

### 1.2.1 Generalized Lineas Models (GLMs)

Generalized Lineas Models (GLMs) build the probabilistic prediction of the occurrence of an event $y$ (presence/absence of the species population in this case) fitting the data to the following formula:

$$y = f(\sum_{k=1}^{m} \alpha_k x_k), \tag{1.1}$$

where $\mathbf{X} = \{X_1, \ldots, X_m\}$ is a set of predictors (in this case the variables used to characterize te environment of the species population) and $f(z) = 1/(1 + exp(-z))$ is the sigmoidal type logistic function (or *logit*) with a bounded output in the $[0, 1]$ range. The unknown parameters $\alpha_k$ are usually estimated through maximum verisimilitude, resulting in a simple optimization problem.

GLMs have been widely used in species distribution modeling (see Guisan *et al.*, 2002, for a description and analysis of its application in ecology).

### 1.2.2 MAXENT

In essence, maximum entropy-based techniques (MAXENT, Phillips *et al.*, 2006) estimate the distribution of a given variable by calculating the distribution with maximum entropy (i.e. the most uniform), subject to the condition that the expected value under this estimated distribution coincides with its empirical mean. Let $\{X_1, \cdots, X_n\}$ be a set of independent observations taken from a region of $X$, according to a certain probability of distribution $\pi$ (in this case the localities of known occurrence of a particular species population). The problem is to construct the distribution of estimated probabilities $\hat{\pi}$ that are closer to $\pi$, using a set of explanatory variables $\{f_1, \cdots, f_n\}$ (predictors, in this case variables characterizing the environment) that act as constraints. The principle of maximum entropy suggests that, among all possible distributions satisfying these constraints, the appropriate is the one that is closest to uniformity (i.e. the one with the greatest entropy). Entropy is here defined as:

$$H(p) = -\sum_{x \epsilon X} p(x) \ln p(x) \tag{1.2}$$

According to DellaPietra *et al.* (1997), this equates to finding the Gibbs distribution of maximum likelihood (i.e. the distribution that is exponential in a linear combination of variables) of the shape:

$$q_{\lambda(x)} = e^{\lambda \times f(x)} / Z_\lambda \tag{1.3}$$

where

$$Z_\lambda = \sum_{x \epsilon X} e^{\lambda \times f(x)} \tag{1.4}$$

and $\lambda \; \epsilon \; \mathbb{R}^n$.

Subsequently, a regularization process is applied in order to avoid over-adjustment (Phillips *et al.*, 2004).

Maximum entropy techniques (MAXENT) have been used more frequently in the last two decades in different fields of research, such as natural language processing (Berger *et al.*, 1996) or spatial physics (Chu & Dowsett, 1997). In the field of ecology and species distribution modeling, MAXENT was introduced as a presence–only technique (Phillips *et al.*, 2004, 2006), for efficiently modeling occurrences lacking absence data. Since then, MAXENT has demonstrated its great performance in comparison with other profiling techniques, such as GARP (garp Genetic Algorithm for Rule set Production, Phillips *et al.*, 2004), or a battery of benchmark algorithms for modeling different species in different geographic areas of the world, (Elith & et al, 2006).

However, the reference to MAXENT as a presence-only method is misleading —as well as for GARP—, since actually do require the use of background or pseudo–absence data (Barbet-Massin *et al.*, 2012; Jiménez-Valverde, 2012), this is, data about a random sample of locations with no information about the presence of the species (see Section 1.4).

### 1.2.3 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression method developed at the beginning of the 90s by Friedman (1991). In essence, MARS allows the approximation of the underlying function using a series of linear regressions by sections —known as *base functions*— as follows:

$$y = \alpha_o + \sum_{k=1}^{K} \alpha_k b_k(\mathbf{x}), \tag{1.5}$$

The slope of these base functions can change in a series of *nodes* $\mathbf{Z}_{ki} = \mathbf{z}_{ki}$, $i = 1, \ldots, m$ con $\mathbf{Z}_{ki} \subset \mathbf{X}$. The popularity of this technique is primarily due to the efficiency of the optimization algorithm that is used for the iterative search of the base functions and the nodes.

In the context of SDMs, MARS has been shown to outperform GLMs in terms of model performance (e.g. Muñoz & Felicísimo, 2004).

### 1.2.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) are classification and regression methods recently developed in the field of artificial intelligence (Scholkopf & Smola, 2001). This technique consists in projecting the input vectors into a multidimensional space in which a hyperplane of maximum separation is constructed, using a metric that is insensitive to $\varepsilon$, by which the (absolute) errors less than $\varepsilon$ are minimized to zero.

The approximation function can be defined as follows:

$$y =< \mathbf{w}; \mathbf{x} > +b \tag{1.6}$$

where $<;>$ denotes the scalar product (in the linear case) or a kernel function (e.g. Gaussian kernel) in the general case of non-linear classifiers. Parameters are obtained from the data by solving the following optimization problem:

$$\text{minimize} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{1.7}$$

$$\text{conditioned to} \quad \begin{cases} y_i - < w; x_i > \leq \epsilon + \xi_i \\ < w; x_i > +b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{1.8}$$

SVMs have recently been developed as a supervised learning technique used for regression and classification, as well as for probabilistic estimation. From the ecological point of view, can be conceptually assimilated into the classic definition by Hutchinson (1957) of the ecological niche, this is, the multidimensional environmental space in which a species is developed (Drake *et al.*, 2006).

Although its application in species distribution modeling is still rare, has been shown to be a potentially useful tool in ecological studies, for example in the prediction of the Zebra Mussel (Dreissena polymorpha) invasion in freshwater systems of North America (Drake & Bossenbroek, 2009).

### *1.2.5 Random Forests (RF)*

Random Forest (RF, Breiman, 2001), are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training data, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging (Breiman, 1996), to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, and being $b = 1, ..., B$ selects a random sample $X_b$ and $Y_b$ with replacement and trains a decision or regression tree $f_b$ on $X_b$, $Y_b$, $B$ times. After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$ as follows:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{1.9}$$

RF is an algorithm that developed out of CART (see below) and bagging approaches and its application in species distribution modeling has been studied by Evans *et al.* (2011). This modeling technique is gaining prominence in remote sensing (Lawrence *et al.*, 2006), forestry (Falkowski *et al.*, 2009), ecology (Cutler *et al.*, 2007), and climate change (Prasad *et al.*, 2006).

### *1.2.6 Classification and Regression Trees (CART)*

Models based in Classification and Regression Trees (CART) have shown a better performance than GLMs to predict the distribution of three species of the Californian oak, as well as to offer interesting properties such as their easy implementation and interpretation of the results, by producing a multidimensional space of variables fully described by a single tree (Hastie *et al.*, 2010). However, some authors have shown their worst behavior against, for instance, GLMs constructed by introducing interactions between variables with simulated species (Santika & Hutchinson, 2009).

## 1.3 Presence Data

Presence data refers to the point localities in the geographical space where individuals of a species population have been observed (defined by x and y coordinates, see Fig. 1.1). The Global Biodiversity Information Facility (GBIF, `http://data.gbif.org`) is a widely used database that collects this information for more than 1.6 million species, shared freely by hundreds of institutions worldwide, including natural history collections and current observations from scientists, researchers and automated monitoring programs.

Another source for obtaining presence data is the Georeferenced Database of Genetic Diversity (GD$^2$, Ehrenmann *et al.*, 2016, `http://gd2.pierroton.inra.fr/gd2/home`), which contains georeferenced data of natural tree population phylogenies. The level of information that provides the GD$^2$ constitutes an added value in the context of species distribution modeling, since experimental evidence suggests that conventional SDMs are not able to properly capture the climatic response of species by treating them as homogeneous units (Pearman *et al.*, 2010; Beierkuhnlein *et al.*, 2011), in fact, the term "species" is a taxonomic designation, and may not necessarily refer to an ecologically homogeneous group of organisms, specially when different ecotypes occur within the study area (Oney *et al.*, 2013). With this regard, Hernández *et al.* (2006) suggested that research in environmental niche modeling should focus on broad distributional sub-units based on distinct genetic linages. This is particularly relevant in climate change studies, because these sub-specific units have differentiated niches (Serra-Varela *et al.*, 2015) and thus, a different response to climate change can be expected (D'Amen *et al.*, 2013). Moreover, González *et al.* (2011) demonstrated that omission error (False Omission Rate, see Section 1.6) is reduced when "biologically meaningful" data (in reference to genetically distinct populations of the same species) are modeled. Therefore, in this Thesis we modeled the distribution of different *Quercus sp* phylogenies (Petit *et al.*, 2002a,b,c), from the $GD^2$ database (Ehrenmann *et al.*, 2016).

## 1.4 Methods for Pseudo-absence Data Generation

Pseudo-absence data is generated by sampling the background area of the study domain from which presence records have not been collected (see Fig. 1.1), assuming that the species is missing in those sites, although they may include presences (i.e. false absences). Consequently, pseudo–absences may represent biased or arbitrary data, and the resulting SDMs may be unreliable (Mateo *et al.*, 2010a).

Pseudo–absence generation process has been shown to have a strong influence on the results obtained. There are two basic questions involved in the generation of pseudo–absences: (1)how and (2)how many. Regarding the second, Barbet-Massin *et al.* (2012) provided different recommendations depending on the SDM used. In this sense, a larger proportion of pseudo–absences against presences can affect model performance positively or negatively, introducing biases in model inter-comparisons, for which an intermediate level of prevalence (proportion of presences *vs* pseudo–absences) should be kept (McPherson *et al.*, 2004; Liu *et al.*, 2005). Alternatively, when working with different proportions, prevalence can be balanced if model fitting is performed with equal weighting of presences vs pseudo–absences (i.e. the total weight of all presences is the same as the total weight of all pseudo–absences).

With respect to how pseudo–absences are generated, comparative analyses addressing the suitability of different methods, some of them quite novel, are scarce in the literature (Zaniewski *et al.*, 2002; Phillips *et al.*, 2009; Lobo *et al.*, 2010), and there is not a consensus on the way in which pseudo–absences should be generated. In fact, several previous studies addressing this issue (e.g. Hengl *et al.*, 2009; Wisz & Guisan, 2009; Stokland *et al.*, 2011; Senay *et al.*, 2013) propose contradictory solutions. As such, the inclusion of reliable pseudo–absences in model calibration remains an open issue.

The most widely applied method of generating pseudo–absences is random selection of the entire study area (RS method, e.g., Gastón & García-Viñas, 2011; Hanspach *et al.*, 2011; Domisch *et al.*, 2013), however, this rises the risk of introducing false absences into the model from locations that are suitable for the species, leading to underestimates of its fundamental niche and potential distribution (Anderson & Raza, 2010). This occurs naturally due to biotic interactions and dispersal limitations that do not allow the species to inhabit, and also very often as a result of sampling biases in the presence–data collections. Faced with this problem, it is common practice to set a buffer distance from known presence localities (exclusion buffer hereafter) in order to minimize the false negative rate (e.g., Mateo *et al.*, 2010a; Bedia *et al.*, 2013).

More elaborated approaches apply a geographically weighted exclusion, which keeps pseudo–absences out from presences using distance maps (Hirzel *et al.*, 2001; Barbet-Massin *et al.*, 2012; Norris *et al.*, 2011; Hengl *et al.*, 2009) or employ a profile technique (presence–only algorithm) as a preliminary step to exclude the background areas classified as suitable, so that pseudo–absences are moved away in the environmental space (RSEP method e.g. Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Barbet-Massin *et al.*, 2012; Liu *et al.*, 2013). These strategies are intended to reduce the background data to those areas where false absences are less likely to occur, while the target group background method (TG method) has been posited as a solution to remove some of the bias in presence–data collections, using the presence localities of other species as biased background data (Phillips *et al.*, 2009).

Another critical matter regarding pseudo–absence data is the extent from which background is sampled. In fact, the available data in the background is usually much larger than the data characterized by presence localities (Anderson & Raza, 2010). A constrained distribution of pseudo–absences around presence locations can lead to misleading models, while unconstrained sampling can artificially inflate test statistics of model performance (see Section 1.6), as well as the weight of less informative predictor variables (Van der Wal & Shoo, 2009). With this regard, Senay *et al.* (2013) limited the background data using a variable importance change criterion based on principal component analysis, and proposed the three–step method (TS method) as an adequate approach to overcome these limitations, envisaged to define the extent and the environmental range of the background from which pseudo–absences are sampled. However, variable importance may not always vary significantly for the whole range of distances tested in a certain background, thus it is not a generalizable method.

### 1.4.1 Overview of Usage of the Different SDMs

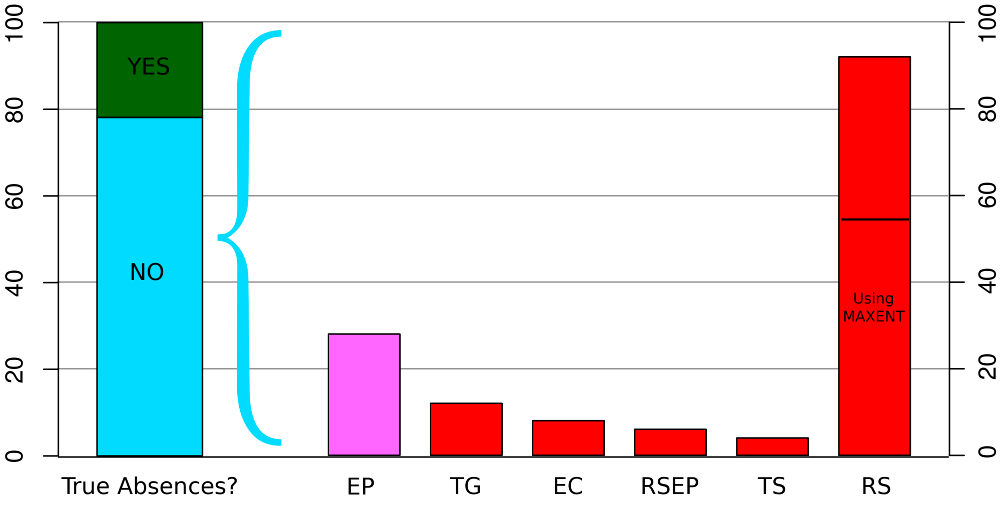

**Figure 1.2:** Percentages of the strategies, regarding absence or pseudo–absence data, used in 64 articles of the first quartile and the topic “environmental sciences” resulting from a search in the SCOPUS database containing the terms “habitat suitability”, “niche modeling” and “background data”, “pseudo–absence” or “presence-only”, for the period 2009–july 2014. The first bar shows the percentages of true absence data availability. Acronym EP (pink bar) refers to the use of profile modeling techniques. Red bars refer to different pseudo–absence generation methods, these are: TG (Target Group), EC (Expert Criteria), RSEP (Randfom Sampling + Environmental Profiling), TS (as RSEP but adding background distance limits) and RS (Random Sampling of the entire background).

In order to have an approximate estimation of the frequency of use of different methods for pseudo–absence data generation, we carried out a search in the SCOPUS database containing the terms “habitat suitability”, “niche modeling” and “background data”, “pseudo–absence” or “presence-only”. The amount of resulting articles were narrowed to the journals of the first quartile and the topic “environmental sciences” for the period 2009–july 2014, yielding a total of 64 articles from which roughly 80% used presence–only datasets, that is, they were lacking true absence data (Fig. 1.2). Of them, the 92% used randomly generated pseudo–absences by considering the entire background for sampling (RS), either explicitly (38%), or implicitly (54%) via the MAXENT algorithm (see e.g.: Barbet-Massin *et al.*, 2012; Jiménez-Valverde, 2012, for details), other 28% used profile techniques (EP, i.e. pseudo–absences are not used) and a 12% used target group background (TG). Percentages under 10% correspond to the novel approaches analyzed in this Thesis (RSEP and TS).

Note that some of the articles analyzed used more than one type of technique, and therefore percentages do not sum up to 100%.

## 1.5 Environmental Data

The environmental conditions at locations of presences and (pseudo-)absences constitute the explanatory variables (predictors) used to characterize the niche of a species population (*Environmental variables for the reference period* in Fig. 1.1). A particular case of SDM application is the characterization of the climatic conditions where a species can potentially live, for which specific climatic variables are used as predictors, typically in the form of bioclimatic variables (Nix, 1986; Busby, 1991). The set of climate predictors used to calibrate SDMs constitute the reference or baseline climate.

Bioclimatic variables (Table 1.1) are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables, representing annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters).

Using climate change data, SDMs can project habitat shifts back in time

**Table 1.1:** The standard set of 19 bioclimatic variables for modeling species distributions. Source: `http://www.worldclim.org/bioclim`.

| ID | Variable definition |
|---|---|
| BIO1 | Annual Mean Temperature |
| BIO2 | Mean Diurnal Range |
| BIO3 | Isothermality |
| BIO4 | Temperature Seasonality |
| BIO5 | Max Temperature of Warmest Month |
| BIO6 | Min Temperature of Coldest Month |
| BIO7 | Temperature Annual Range |
| BIO8 | Mean Temperature of Wettest Quarter |
| BIO9 | Mean Temperature of Driest Quarter |
| BIO10 | Mean Temperature of Warmest Quarter |
| BIO11 | Mean Temperature of Coldest Quarter |
| BIO12 | Annual Precipitation |
| BIO13 | Precipitation of Wettest Month |
| BIO14 | Precipitation of Driest Month |
| BIO15 | Precipitation Seasonality |
| BIO16 | Precipitation of Wettest Quarter |
| BIO17 | Precipitation of Driest Quarter |
| BIO18 | Precipitation of Warmest Quarter |
| BIO19 | Precipitation of Coldest Quarter |

(e.g. Maiorano *et al.*, 2013) or to the future (e.g. Engler *et al.*, 2009). In this sense, the possibility of building predictive models that are able to extrapolate across time (and space) are contingent on the choice of appropriate predictors (Peterson, 2011; Rödder *et al.*, 2009). This includes the choice of the baseline climate dataset and the strategy for variable selection (Peterson & Nakazawa, 2008; Pliscoff *et al.*, 2014; Baker *et al.*, 2016).

### 1.5.1 Baseline Climate Data

An important barrier for SDM development is climate data retrieval and preparation. Gridded datasets of baseline climate are built from historical observations. The numerous climate databases available are scattered across many different repositories with various file formats, variable naming conventions, etc., sometimes requiring relatively complex, time-consuming data downloads and error-prone processing steps prior to SDM development. This is also a major barrier for research reproducibility and data exchange.

As a result, there is an increasing demand of climate products to produce models at an adequate spatial resolution and varying geographical extents –up to global–. The recent development of new high-resolution bioclimatic datasets has broadened the scope of SDMs, including its application in climate change impact studies (Peterson *et al.*, 2002; Hijmans & Graham, 2006). In this context, some authors have highlighted the need for high-resolution data, given the inability of coarse resolution climate models (see Section 2) to represent local refugia (Randin *et al.*, 2009; Franklin *et al.*, 2013). One of the most popular global bioclimatic products is the WorldClim dataset (Hijmans *et al.*, 2005), which is widely used because it is easily available and offers high resolution data worldwide. Other new global products of similar characteristics have recently appeared in the literature (e.g., the new data set by Climond Kriticos *et al.*, 2012), which is based partly on WorldClim data), indicating the high demand of this type of products for SDM applications. However, these global datasets have not been rigorously tested in smaller regions, and their use in regional studies may pose problems due to their poor representation of local climate features (Bedia *et al.*, 2013). Moreover, this problem may be aggravated when predicting potential distributions in the future as a consequence of the uncertainty derived from the future altered climate scenarios (see Sections 1.7 and 2.3). Faced with this problem The Regional Baseline Climate of the Basque Country (**RCBC**) was developed as an alternative to existing public products.

In the following, the main characteristics of the baseline climate datasets used in this Thesis are introduced. The interested reader is referred to the published documentation of these datasets for further details on their construction.

### *RCBC*

The Regional Climate of the Basque Country (**RCBC**) was generated in the frame of the ADAPTACLIMA project (http://www.adaptaclima.eu/). This gridded dataset is based on AEMET (Spanish Meteorology Agency) stations distributed across the Basque Country and surrounding areas. After a process of data quality control within the period 1950–2007, a subset of stations was selected for the period 1971-2000, based on the available percentage of data, the homogeneity of the series and the spatial distribution of the station network. As a result, almost all the stations selected have more than the 50% of the data and the number of stations with at least a 75% of the data is constant through the whole period.

Regarding interpolation and regression of station data, the methodology for building the high resolution climate grid of Cantabria (UC, Gutiérrez *et al.*, 2010) was followed. The performance of different techniques was tested, namely thin-plate splines, angular distance weighting and kriging (Krige, 1951), obtaining best results with the latter one, which has been widely used in climate research (Atkinson & Lloyd, 1998; Biau *et al.*, 1999; Haylock *et al.*, 2008). For precipitation, a two-step interpolation process was conducted: first, precipitation occurrence was interpolated using indicator kriging (Juang & Lee, 1998); then, the amount of precipitation was interpolated using ordinary kriging, assigning values of 0 to all 'dry' points. Thus, the frequency distribution of precipitation for both occurrence and amount was optimally fit. The final 1 km-resolution grid was obtained by regression-kriging (Hengl *et al.*, 2007), introducing a set of basic covariates describing terrain chacteristics including, elevation, distance to coastline, and topographic blocking effects (Bedia *et al.*, 2013).

### *WorldClim*

WorldClim (**WC**, Hijmans *et al.*, 2005) is a global temperature and precipitation dataset available at different spatial resolutions, from 10 arc minutes (≈ 20 km) to 30 arc seconds (≈ 1 km), obtained by applying a thin-plate spline smoothing interpolation algorithm to a large number of weather stations throughout the world, covering most of Earth for approximately 50 years (1950–2000). A set of standard bioclimatic variables (Hijmans *et al.*, 2005) for modeling is freely available for download from the internet (http://www.worldclim.org), —including future Climate Change projections— therefore, WorldClim has been widely used in SDM studies (e.g. Barredo *et al.*, 2015; Mellert *et al.*, 2015; Curtis & Bradley, 2016),

### *E-OBS*

The **E-OBS** dataset (Haylock *et al.*, 2008, v14) is a European daily high-resolution (0.25° ≈ 30km) gridded dataset for precipitation, mean, maximum and minimum temperature for the period 1950-2012, developed in the frame of EU-ENSEMBLES project (van der Linden & Mitchell, 2009, `http://www.ensembles-eu.org`) with the aim of using it for validation of Regional Climate Models and for climate change studies. It was constructed through interpolation of The European Climate Assessment & Dataset (ECA&D, http://eca.knmi.nl/) station data, the most complete collection of station data over Europe. The E-OBS dataset was obtained applying a three stage process: monthly mean values of temperature and precipitation were first interpolated to a rotated pole 0.1° grid using three dimensional thin plate splines; daily anomalies (departure from the monthly mean) were interpolated on the same grid and combined with the monthly mean grid (interpolation was performed applying the kriging method); Finally, the 0.1° grid values were used to compute area-average values at the E-OBS grid resolution.

### 1.5.2 Strategy for Variable Selection

The are three basic properties in a set of explanatory variables or predictors that need to be considered, these are (1) proximality, (2) multicolinearity and (3) dimensionality:

*Proximality* is the degree in which a set of variables can define the physiological limits of a species population. Proximal variables are expected to bring the model closer to the real requirements of the species, thus allowing more robust predictions (Rödder *et al.*, 2009; Petitpierre *et al.*, 2016).

*Multicollinearity* is the high correlation between two or more variables and can affect model performance negatively if these correlation varies between the environmental subset used for calibration and the projection environment (Dormann *et al.*, 2008).

*Dimensionality* is the number of variables relative to the available observations. Building SDMs with too many predictors leads to over–parameterization, potentially reducing model transferability (Warren & Seifert, 2011).

Obtaining proximal predictors is a difficult task, given that involves previous knowledge of the species ecological requirements and the availability of the corresponding spatial data objects. Moreover, proximality could be confounded with highly correlated variables.

Regarding multicollinearity and dimensionality, there are different strategies for variable selection aimed at reducing both properties (Petitpierre *et al.*, 2016), such as removing highly correlated variables, using statistical algorithms to select the most relevant variables (e.g. stepwise selection) or using the first principal components (PCs) of the whole set of variables (see Chapter 6).

*The stepwise procedure* automates the selection of significant explanatory variables through three alternative approaches: forward selection, backward selection and forward–backward selection. In forward selection, the model initially contains no variables, and variables are added sequentially until a final model is obtained. In backward selection, all variables are included in the initial model, and these are then removed sequentially until a final model is produced. Forward–backward selection is a variation on forward selection, in which each forward step is followed by a backward step to remove variables in the model that are no longer significantly related to the response (Pearce & Ferrier, 2000).

*Principal Component Analysis (PCA)* can be used to reduce the number of variables (dimensionality) by selecting the first components. Collinearity is also reduced, because components are orthogonal (see e.g. Townsend Peterson *et al.*, 2007; Zhang & Zhang, 2012).

## 1.6 Model Evaluation

Models are evaluated based in the level of agreement between observed presences/absences and the predicted values for the occurrence data used to built SDMs (*Model performance assessment* in Fig. 1.1). This is known as model accuracy, performance or goodness.

In this Section, different procedures to evaluate and compare the models are described. First, we will discuss the different numerical indexes that are used to assess the predictive goodness of the models, reviewing their use in previous ecological studies and their advantages and limitations. Second, other alternative evaluation techniques —not generally applied in ecological studies— are described.

### 1.6.1 Model Performance Assessment

#### Common numerical indices

There is no single index or metric for SDM performance assessment, since different metrics provide information on different aspects of the relationship between predicted and observed values; The situation is more complex in the case of probabilistic predictions (see e.g. Jolliffe & Stephenson, 2003).

Regarding deterministic binary predictions, there are two error sources: false positives (FP, or error type I), which occur when the model predicts a positive case (presence) when in fact one negative is observed (absence), and, on the other hand, false negatives (FN, or error type II), when the model misses the prediction of a positive case by predicting a negative one.

These values are typically arranged in a table (Fig. 1.2), together with the other two cases left (i.e. the positive and the true negatives, TP and TN respectively), in what is known as the *confusion matrix* (Fielding & Bell, 1997). From this table, a series of measures of the goodness and/or error of prediction are calculated, for instance:

**Table 1.2:** Confusion matrix. Error types I and II (i.e. false positives and false negatives respectively) are written in red. Well classified cases (i.e. true positives and true negatives) are written in green.

| | | PREDICTED | |
|---|---|---|---|
| | | positive | negative |
| OBSERVED | positive | TP | FN |
| | negative | FP | TN |

*Sensitivity* or the True Positive Rate (TPR), is the proportion of positives that are correctly predicted as such:

$$TPR = \frac{TP}{TP + FP} \tag{1.10}$$

*Specificity* or the True Negative Rate (TNR), is the proportion of negatives that are correctly predicted as such:

$$TNR = \frac{TN}{TN + FN} \tag{1.11}$$

*False Discovery Rate (FDR)* is the proportion of incorrectly predicted positives:

$$FDR = \frac{FP}{FP + TP} \tag{1.12}$$

*False Omission Rate (FOR)* is the proportion of incorrectly predicted negatives:

$$FOR = \frac{FN}{FN + TN} \tag{1.13}$$

Regarding probabilistic predictions, a graph called the ROC (Receiver Operating Characteristics, Hanley & McNeil, 1982) curve is widely used, which is constructed by plotting the values of $sensitivity(u)$ versus $1-specificity(u)$ of a deterministic prediction given for a probability threshold $u$. Probability values below/above $u$ are considered positive/negative (presence/absence). Therefore the ROC curve describes the predictive ability of the system for the entire range of probabilities, that is quantitatively assessed by the area it encloses, this is the AUC (area under the curve).

*The AUC* provides an overall measure of the system and ranges from 1 (perfect prediction) to 0 (random prediction). Previous ecological studies have shown that the AUC is independent of the prevalence (Manel *et al.*, 2001; Allouche *et al.*, 2006), and is an appropriate measure when the aim is to compare the performance of different SDMs (Fielding & Bell, 1997; Allouche *et al.*, 2006).

In addition, a deterministic prediction is often necessary. In this case, it is necessary to define a probability threshold (cut value) for the separation of positive and negative cases. However, the AUC does not give any information about the threshold to be used, which often depends on the particular objectives of each case, depending on whether the objective is to minimize FN or FP errors, or other conditions imposed by the user (Fielding & Bell, 1997; Freeman & Moisen, 2008). A typical practice is to use prevalence (proportion of presences *vs* absences) as cut value. Alternatively, an optimized probability threshold (OPT) can be computed, for instance, the threshold that maximizes the True Skill Statistic (TSS, see below).

From the defined probability thresholds, the corresponding confusion matrixes are constructed to calculate further evaluation statistics, such as the previously defined Sensitivity and Specificity or the commonly used *Cohen's Kappa* ($\kappa$) and *True Skill Statistic* (TSS):

*Cohen's Kappa ($\kappa$)* measures the level of agreement between the deterministic prediction and the observed value, relative to what would be a prediction obtained by chance. $\kappa$ is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1.14}$$

where $Pr(a)$ is the proportion of correctly classified events and $Pr(e)$ is the hypothetical probability of success due to chance. $Pr(e)$ is defined as follows:

$$Pr(e) = \frac{1}{N}[(TP + TN) \times (TP + FP) + (TN + FN) \times (TN + FP)] \tag{1.15}$$

where $N$ is the total number of observations. The maximum value ($\kappa = 1$) occurs when the coincidence between predicted and observed values is perfect, whereas normally a perfect agreement does not occur, it is expected to be larger than simply by chance, so $1 > \kappa \geq 0$.

*True Skill Statistic* is similar to the Cohen's Kappa, and gives a measure of the goodness of the classifier to separate positive (presence) events from negative ones (absences). TSS is defined as follows:

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{1.16}$$

Both statistics have the advantage of correcting the precision of the models by which they are expected to be due to chance (Fielding & Bell, 1997; Manel *et al.*, 2001). However, TSS has the additional advantage of being independent of the prevalence, whereas $\kappa$ can sometimes distort the performance measure due to its unimodal response to the prevalence (Allouche *et al.*, 2006).

#### *Reliability Diagrams*

The AUC does not report on other important aspects of a predictive system (see e.g. Lobo *et al.*, 2008, for a critical review of this index). For instance, high AUC values (closer to 1) indicate good model discrimination, although this does not necessarily correspond to a high numerical accuracy of the predictions (Bedia *et al.*, 2011). *Reliability diagrams* (also known as *calibration plots*) provide additional information regarding the level of agreement between predicted and observed probabilities of occurrence. This information is displayed in the form of a plot such that the better the agreement, the closer the line is to the diagonal for the whole range of probability values (see e.g. Bedia *et al.*, 2011; Vaughan & Ormerod, 2005, for a wider explanation in the context of SDM assessment).

#### *Boyce Index*

The Boyce Index ($B$) is a presence–only measure that provides information on how observed presences are distributed across the gradient of predicted presences and how this differs from the random expectation in the study area. It is analogous to the Spearman correlation and varies between -1 and 1, with zero meaning no different from random (see Hirzel *et al.*, 2006; Petitpierre *et al.*, 2016).

### *1.6.2 Validation Procedure*

The validation is a fundamental process in evaluating the effectiveness of any predictive model. In the case of SDMs, the ideal validation is to contrast the skill of the constructed model with an independent set of occurrence data collected in the field. However, field data are often scarce and valuable, and collecting new information is expensive in time and effort, or simply not feasible. There are, fortunately, other possibilities that allow robust estimates of SDM performance. Resampling techniques (e.g. bootstrapping, cross-validation, etc.) are simple to implement and effective, allowing the optimization of the available occurrence data and a realistic performance assessment.

#### *Cross-Validation*

Cross-validation techniques (Steyerberg *et al.*, 2010) consists in leaving part of the data outside model calibration to replace truly independent data for model evaluation (see Fig. 1.1), as it is commonplace in ecological studies (e.g. Manel *et al.*, 1999).

In particular, we used a 10-fold cross validation approach to perform all the analysis in this Thesis, given that it is equally efficient in the error estimation as other techniques computationally more demanding like for instance leave-one-out cross validation (Kohavi, 1995).

## *1.7 Illustrative Example: Reference Climate*

This Thesis emerged from the ADAPTACLIMA (`www.adaptaclima.eu`) and K-EGOKITZEN (`http://www.neiker.net/neiker/k-egokitzen/`) projects, where the impact of climate change to different forest species habitats was studied. In this framework, the RCBC baseline climate dataset was developed (see Section 1.5), in view of the need of an appropriate dataset for regional studies in the Basque Country (Northern Iberian Peninsula).

In this section, an example of Species Distribution Modeling application is illustrated, where the RCBC and the WC bioclimatic datasets are compared in a region of complex orography.

Figure 1.3 compares mean climatologies of the minimum temperature of coldest month, maximum temperature of warmest month and the annual precipitation (variables named BIO6, BIO5, BIO12 respectively) among datasets (RCBC and WC) for the reference period 1971-2000. The spatial pattern of temperature (BIO5 and BIO6 in Fig. 1.3) is similar across datasets and strongly controlled by the topography, however, considering RCBC as the reference dataset, maximum temperature (BIO5) of WC is negatively biased and minimum temperature (BIO6) is positively biased. Precipitation (BIO12) is seriously underestimated by WC and the spatial pattern is not well reproduced.

In order to see the influence of the dataset used for building SDMs, in this section two tree species are modeled in the Basque Country, using different baseline climate datasets (RCBC and WC) and four SDM techniques.

We modeled the European beech (*Fagus sylvatica*) whose distribution responds to a high relation with climate conditions, and the Pyrenean oak (*Quercus pyrenaica*) which has a wider distribution (Fig. 1.4). Presence data (see Fig. 1.1) of each species was generated by sampling 1000 locations (points) from The 3rd Spanish National Forest Inventory (IFN3, `http://www.mapama.gob.es/es/biodiversidad/temas/inventarios-nacionales/`). Same number of pseudo–absences (prevalence = 0.5) were randomly sampled from the background areas were presence data is missing (i.e. the RS method was applied), keeping an exclusion buffer of 5 km around presences in order to decrease the false absence ratio (see Section 1.6).

In order to reduce dimensionality of the set of predictors, from the 19 standard bioclimatic variables (BIO1-BIO19, see e.g. Hijmans & Graham, 2006), we considered the temperature based BIO1, BIO5 and BIO6 and the precipitation based BIO12, BIO18 and BIO19, for both baseline climate datasets (RCBC and WC).

GLM , MARS, RF and MAXENT modeling techniques were applied for each climate dataset and species, to analyze the discrepancies and the predictive skill in all cases by evaluating the resulting models in the light of their AUC (area under the ROC curve). We performed a k-fold cross–validation of the

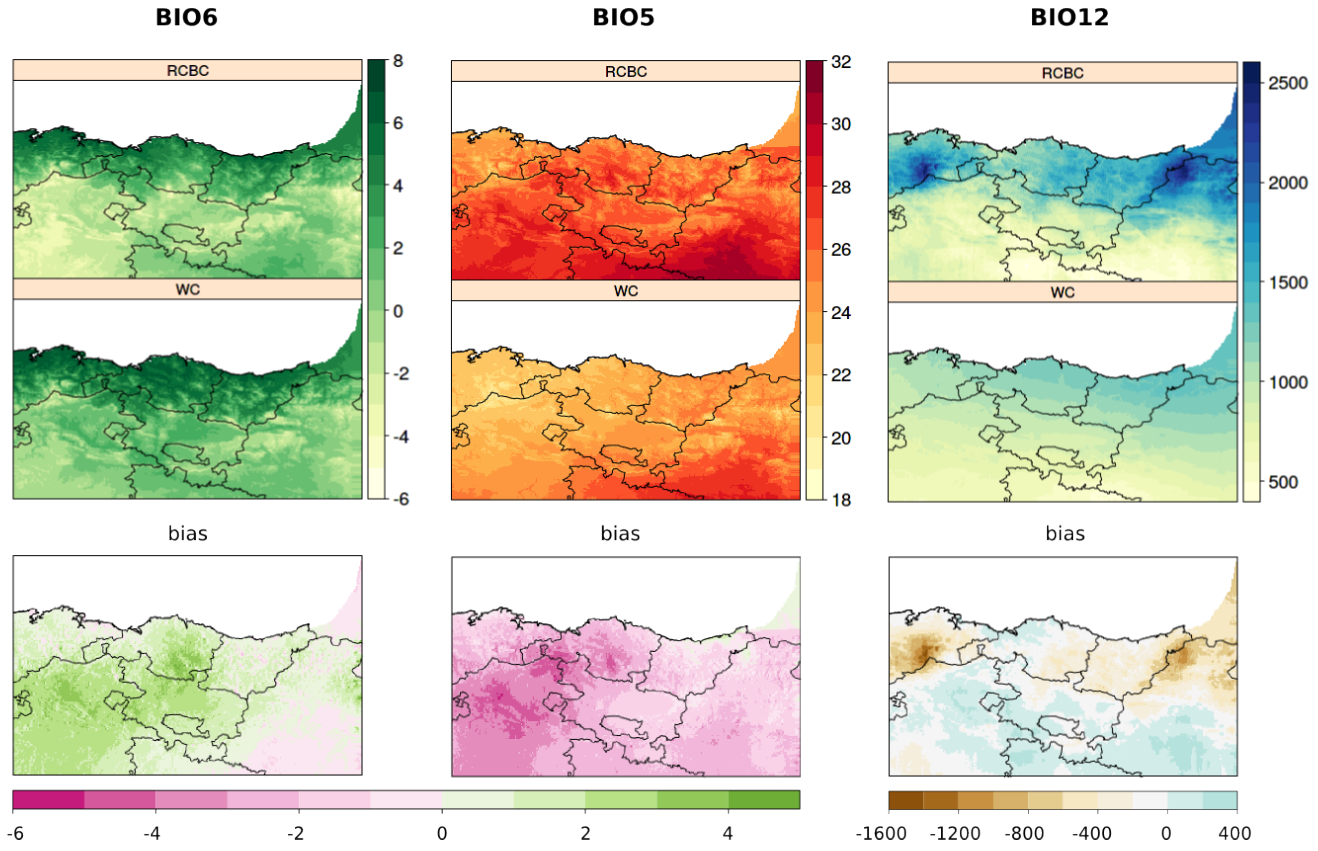

**Figure 1.3:** BIO6 (Min Temperature of Coldest Month), BIO5 (Max Temperature of Warmest Month) and BIO12 (Annual Precipitation) for the RCBC and the WC datasets in the Basque Country. The bias between both datasets are also shown for each bioclimatic variable.

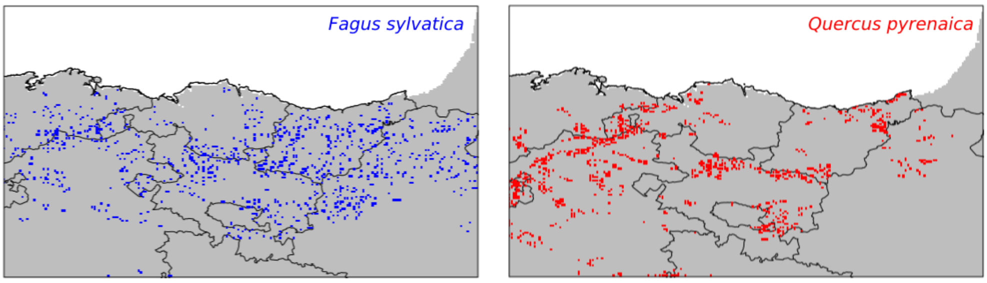

**Figure 1.4:** Presence data for *Fagus sylvatica* (left) and *Quercus pyrenaica* (right).

models, with k=10 strated randomly splitted subsets of presence/absence.

Built SDMs were projected into reference climate, this is, into the same dataset, period and variables used as predictors for model calibration, thus obtaining the reference suitability maps, i.e. the probabilities (ranged from 0 to 1) of the species habitat suitability in reference climate. We used a threshold of 0.5 (the prevalence value) in order to transform probability maps to presence/absence deterministic maps of the predicted species distributions.

Model performance in terms of AUC was higher for *Fagus sylvatica* than for *Quercus pyrenaica* (Fig. 1.5), indicating that the distribution of the first is better explained by the climate variables used as predictors. Both species were modeled with higher accuracy by the non-linear techniques (MARS, RF and MAXENT). Regarding climate datasets, WC achieved higher AUCs for all modeling techniques, specially for *Quercus pyrenaica*, while the scores of RCBC and WC were very similar for *Fagus sylvatica*.

On the contrary, predicted probabilities (suitability maps) were dissimilar among datasets for both species as depicted by Figure 1.6, where the multi-model mean projections (ensemble mean of all SDMs) and the map of the bias are shown. Given that projected probabilities are ranged from 0 to 1, opposite SDM projections would produce a bias map of value 1 for all grid cells. Therefore, the proportion of the bias relative to the bias of hypothetical opposite suitabilities, is given by the mean bias of the projection domain as shown in Equation 1.17, where $p_{x,i}$ and $p_{y,i}$ are the suitability scores (or probabilities)

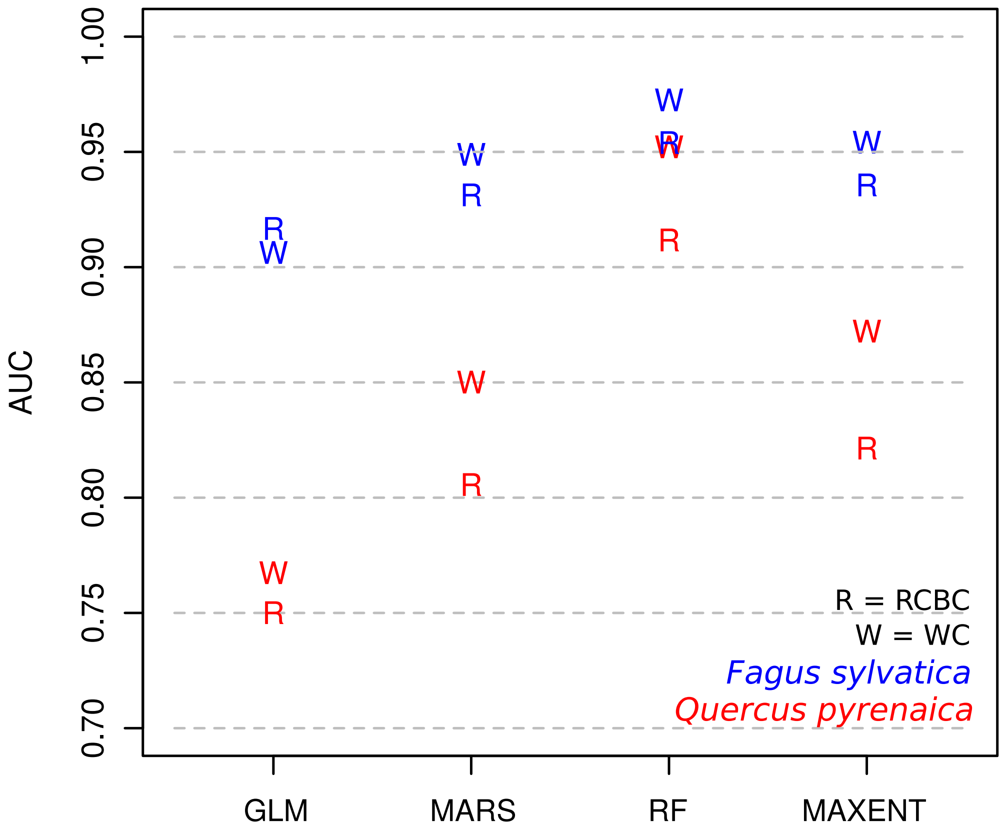

**Figure 1.5:** AUC scores (y axis) corresponding to different baseline climate datasets (R: RCBC and W: WC), tree species (blue: *Fagus sylvatica* and red: *Quercus pyrenaica*) and SDMs (x axis; GLM, MARS, RF and MAXENT). The legend is displayed in the bottom right corner.

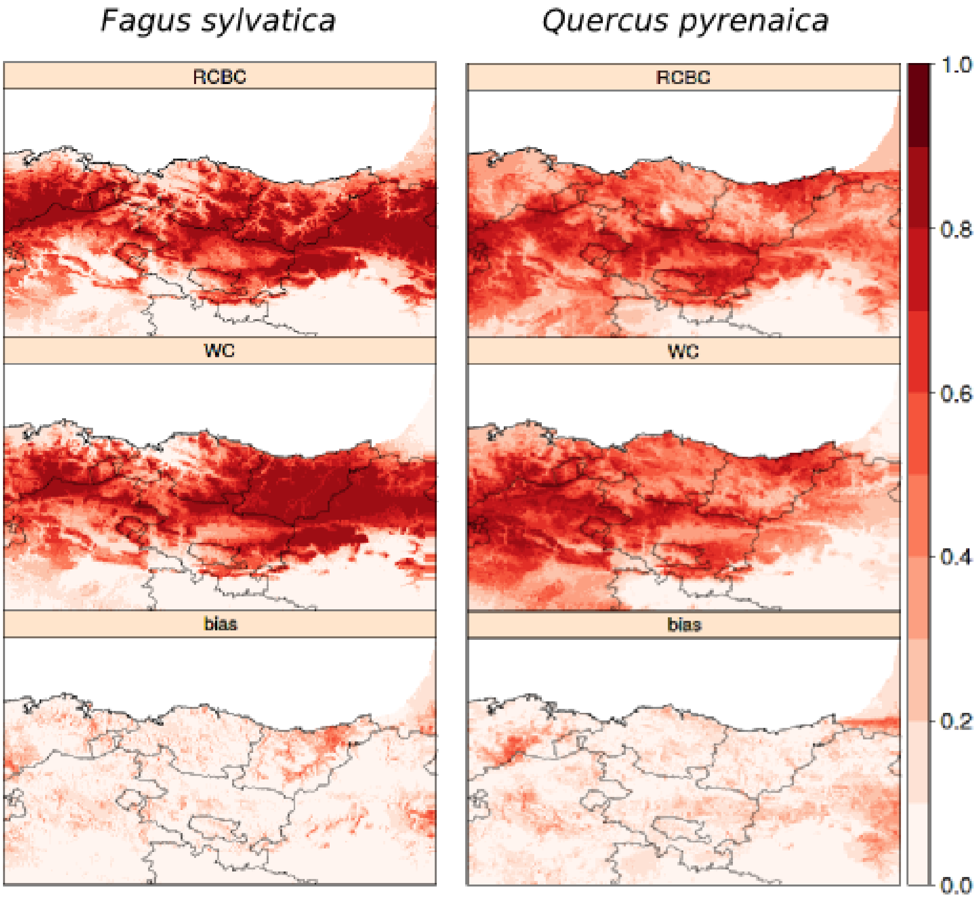

**Figure 1.6:** Reference suitability maps of the SDM ensemble mean, corresponding to different baseline climate datasets (RCBC and WC) and tree species (*Fagus sylvatica* and *Quercus pyrenaica*). The absolute bias between both datasets is also shown (bias) for each tree species.

for dataset $X$ and $Y$ in grid cell $i$. Percentages of the resulting bias in present conditions were 8.9 % and 11.3 % for *Fagus sylvatica* and *Quercus pyrenaica* respectively.

$$B(p_x, p_y) = \frac{\sum_{i=1}^{n} |p_{x,i} - p_{y,i}|}{n} \times 100, \tag{1.17}$$

Figure 1.7 shows the deterministic maps of predicted presence/absence resulting from applying the probability threshold of 0.5 as cut value to classify the maps of Figure 1.6. Here, significant differences can be noted between datasets for both species (e.g. the coast of Gipuzkoa for *Fagus sylvatica* and

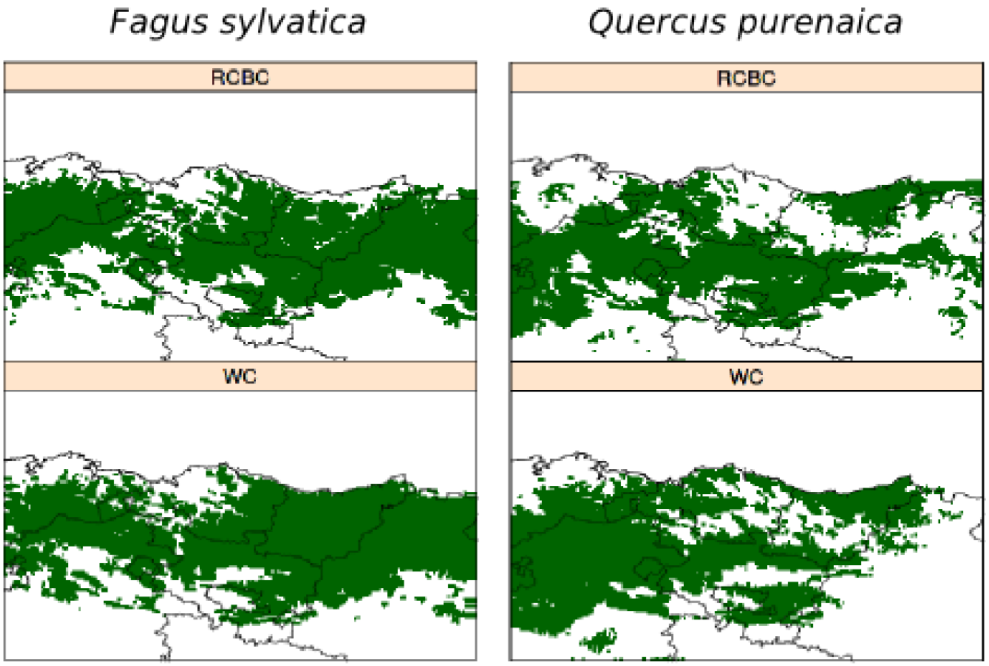

**Figure 1.7:** Reference deterministic maps of predicted presence (green areas) and absence (white areas) of the SDM ensemble mean, corresponding to different baseline climate datasets (RCBC and WC) and tree species (*Fagus sylvatica* and *Quercus pyrenaica*).

the western half for *Quercus pyrenaica*).

This example of species distribution modeling application continues in Section 2.3.

# CHAPTER 2

# Future Projections of Species Distributions

A popular application of SDMs is the future projection of species distributions in order to assess key topics in environmental conservation such as monitoring biological responses to climate change (Hamann & Wang, 2006), species invasions (Jeschke & Strayer, 2008), natural reserve planning (Araújo *et al.*, 2004) or disease transmission (Drake & Beier, 2014) among others. These projections are being increasingly used by the vulnerability, impacts and adaptation (VIA) community, so communicating limitations, credibility and uncertainty in a comprehensive form is crucial for informing decision making processes (Gould *et al.*, 2014; Urban, 2015; Zhang *et al.*, 2015).

These projections are obtained using the climate data provided by global and regional climate change projections. This information is periodically generated by the climate modeling community as an international effort framed under the initiatives of the Intergovernmental Panel on Climate Change (IPCC) considering a cascade of uncertainties: 1) different socio-economic and demographic future pathways and their translation into concentrations of atmospheric greenhouse gas concentrations (*emission scenarios*), 2) global projections of future climate obtained using global climate models forced by the different emission scenarios (*Global Climate Models, GCMs*), 3) regional future projections, obtained using regional models forced with the global climate projections (*Regional Climate Models, RCMs*). The resulting ensemble of regional climate change projections constitute the basis to obtain actionable information at the scale needed to analyze local impacts on human and natural systems (Wilby & Dessai, 2010). For instance, as we show below, SDM projections are typically obtained considering the mean of the ensemble of climate projections, applying change factors (also called the delta rule) to modify the baseline climate according to the changing climate conditions.

In Section 2.1 we analyze each of these components, with special emphasis on the underlying uncertainties which need to be considered in impact studies. In Section 2.1.4 we describe the ENSEMBLES regional projections dataset, which is used in this Thesis. Then, we describe the standard methodology followed by the niche modeling community to use these results in the framework of species distribution models to obtain future projections of species distributions (Section 2.2.2). Finally, in Section 2.3, the different concepts introduced in this chapter are illustrated using the case study introduced in Section 1.7.

## 2.1 Future Climate Projections

Future climate projections are plausible descriptions of the future climate as simulated by both global and regional climate models (GCMs and RCMs, respectively) from different scenarios of greenhouse gas emission, which define the radiative forcing of the climate system for the next decades (e.g. for the 21st century).

### 2.1.1 Emission Scenarios

Climate change emission scenarios are plausible estimations of future pathways for the emission of greenhouse gases resulting from different estimations of future socioeconomic and demographic change, including population levels, economic activity, patterns of technological change, etc. (IPCC, 2000; Nakićenović, 2000). The IPCC is the leading international body for the assessment of climate change. It was established by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO) in 1988 to provide the world with a clear scientific view on the current state of knowledge in climate change and its potential environmental and socio-economic impacts. The IPCC-SRES (Special Report on Emissions Scenarios) scenarios were constructed building on different storylines characterizing plausible future development pathways, determined by driving forces such as demographic growth, socio-economic development, and technological change, and focusing on the production of greenhouse gases (IPCC, 2000, Fig. 2.1). Figure 2.1 shows a schematic illustration of the IPCC-SRES emission scenarios, including the evolution of carbon dioxide concentrations under three illustrative scenarios commonly considered to represent the range of uncertainty due to the scenario (B1, A1B and A2). Besides these future scenarios, there is also a historical (or control) one, considering the historical gas emissions estimated for the 20 century (scenario 20C3M hereafter). This scenario is used to reproduce and validate historical climate conditions with the climate models.

As we show in the next section, the SRES-IPCC scenarios have been used by the different global Climate Model Intercomparisson Projects (CMIP) to produce climate change projections according to the different scenarios. These projections provide detailed information about the future evolution of key climate variables for environmental studies (e.g. temperature and precipitation).

These scenarios were revised in the last IPCC Assessment Report (IPCC-AR5), including a new methodology building on representative emissions (and greenhouse gas concentrations) as given by the Representative Concentration Pathways (RCPs, Moss *et al.*, 2010) (see more details at `http://sedac.ipcc-data.org/ddc/ar5_scenario_process/RCPs.html`). These new scenarios have fed the new generation of CMIP global change projections

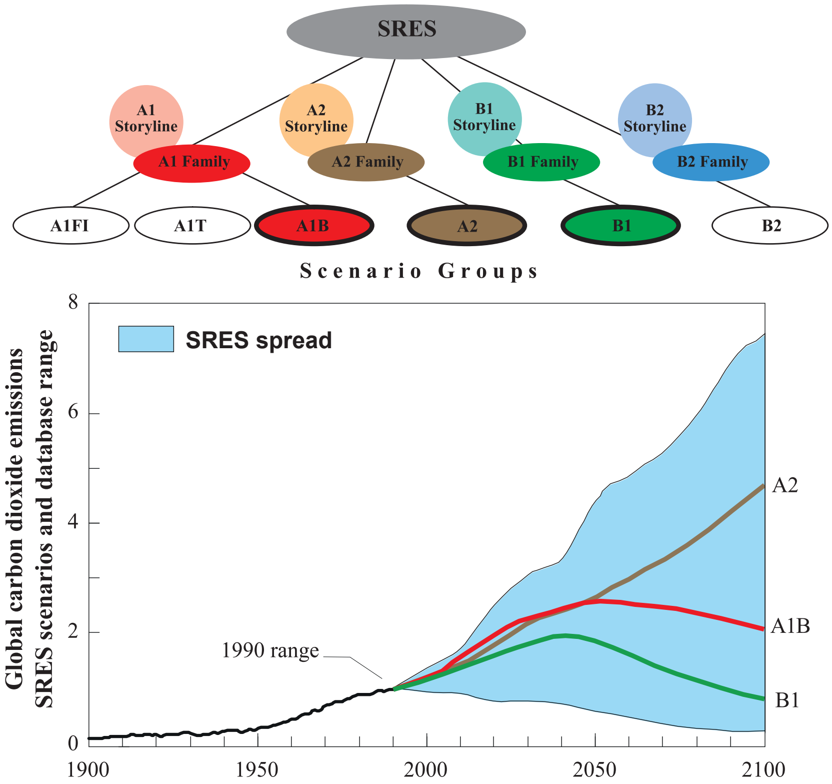

**Figure 2.1:** (top) Schematic illustration of SRES-IPCC scenarios. Four qualitative storylines yield four sets of scenario families: A1, A2, B1, and B2. The A1 family is characterized by alternative developments of energy technologies: A1FI (fossil fuel intensive), A1B (balanced), and A1T (predominantly non-fossil fuel). The B1 scenario family describes a convergent world with rapid changes in economic structures toward a service and information economy, with reductions in material intensity, and the introduction of clean and resource-efficient technologies. The A2 family describes "business as usual" conditions. (bottom) Evolution of carbon dioxide concentrations along the 21st century as given by three illustrative scenarios B1 (optimistic), A1B (intermediate) and B2 (pessimistic). Source: Adapted from IPCC (2000).

(CMIP5), used in the latest IPCC-AR5 report. However, in this Thesis we consider products derived from IPCC-SRES, in particular the projections developed in the framework of the ENSEMBLES regional climate change initiative, building on CMIP3 models (IPCC-AR4). These projections have undergone an exhaustive quality control and assessment process by the different Vulnerability, Impacts and Adaptation (VIA) communities (e.g. detection of ill-performing models). Therefore, they constitute a consolidated reliable dataset suitable for climate change applications.

### 2.1.2 Global Climate Models (GCMs)

The primary source of information for projecting future climate are the simulations produced using Global Climate Models (GCMs), which simulate the global dynamics of the components of the climate system (i.e. the atmosphere, the oceans, the land surface, and the cryosphere, as well as the interactions between them) for different future emission scenarios (Räisänen, 2007). For instance, the dynamics of the atmosphere is primarily governed by three fundamental physical principles: conservation of mass, conservation of momentum and conservation of energy, represented by a system of equations. These equations are solved using sophisticated mathematical methods, which iterative evolve the state of the system starting from an initial condition. This process is carried out numerically (using supercomputers), considering 3D discretized grids covering the globe and the different levels of the atmosphere with a prescribed resolution. Besides the dynamical equations which are numerically solved, other terms need to be approximated from the system's variables (or *parametrized*), in order to keep the system stable and balanced, considering small-scale process occurring at resolutions not resolved by the model. These parametrizations are most often empirically calibrated and therefore, they constitute one of the major sources of uncertainty of GCM simulations. An schematic illustration of this modeling process is shown in Figure 2.2.

A second source of uncertainty in climate change projections is structural model uncertainty, arising from the fact that not all relevant processes are well represented in the different GCMs. Multi-model ensembles are commonly used as a pragmatic approach to characterize model uncertainty. This idea is using several GCMs (with different dynamical cores and parameterizations) to simulate the future climate conditions under the different forcing scenarios,

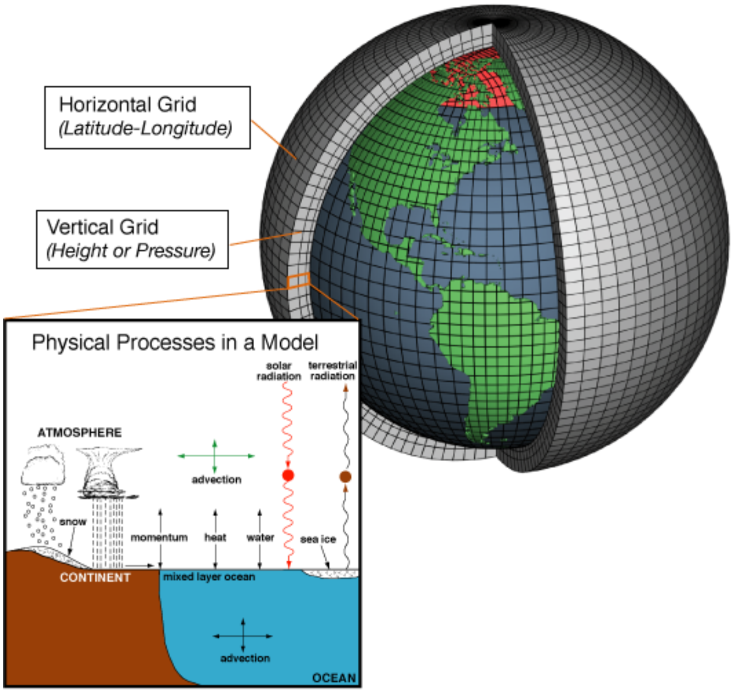

**Figure 2.2:** A schematic image of a Global Climate Model (GCM) dividing the planet into a 3-dimensional grid to solve the basic equations, calculating winds, heat transfer, radiation, relative humidity, etc. within each grid and evaluating interactions with neighboring points. Image source: NOAA 200th Celebration `http://celebrating200years.noaa.gov/breakthroughs/climate_model/modeling_schematic.html`.

thus producing and ensemble of projections which are considered equiprobable and are used to assess all these uncertainties in future climate projections. This is the approach followed by the CMIP international initiative, where several participating models (nearly 30 in the latest generation, CMIP5) are run in the same experimental conditions and driven by the same scenarios (4 in CMIP5), thus producing an enormous amount of information (120 members of the ensemble) characterizing the projected future climate for several scenarios. This poses several data access and computation problems for impact studies, which need to run their models/assessments for each specific member and evaluate their outputs afterwards in order to properly characterize the uncertainty. In practise, this may involve taking a selection of models based on performance evaluations and/or problem expertise.

Table 2.1 shows a summary of the GCMs used in the ENSEMBLES project, an European initiative contributing to CMIP, which is the dataset considered in this Thesis (in particular, the models used to drive regional climate change projections, as described in the next section, are boldfaced). The data were obtained from the CERA-database of the World Data Center for Climate (http://cera-www.dkrz.de/CERA/). The stream 1 (S1) models were used for the Fourth Assessment Report of the International Panel of Climate Change (IPCC-AR4), whereas the stream 2 (S2) models were special simulations developed within the ENSEMBLES project.

**Table 2.1:** Summary of the GCMs from the two streams (Str) of the ENSEMBLES project. Stream 1 corresponds to the IPCC-AR4 model versions (S1), whereas S2 indicates new versions developed within the ENSEMBLES project.

| GCM name | Acronym | Str | Institution | Information |
|---|---|---|---|---|
| **BCCR-BCM2** | BCM2 | S1 | Bjerknes Institute of Climate Res. | Drange (2006) |
| **CNCM-CM3** | CNCM3 | S1 | Centre National de Recher. Mét. | Royer (2006) |
| ECHO-G | EGMAM | S1 | Freie Universität Berlin | Niehörster (2008) |
| IPSL-CM4 | IPCM4 | S1 | Institute Pierre Simon Laplace | Dufresne (2007) |
| METO-HC-HadGEM | HADGEM | S1 | Hadley Centre | Johns (2008) |
| **METO-HC-HadCM3** | HADCM3 | S1 | Hadley Centre | Johns (2009a) |
| **MPI-ECHAM5** | MPEH5 | S1 | Max Planck Institut | Roeckner (2007) |
| CNCM-CM33 | CNCM3 | S2 | Centre National de Recher. Mét. | Royer (2008) |
| ECHO-G2 | EGMAM2 | S2 | Freie Universität Berlin | Huebener & Koerper (2008) |
| IPSL-CM4v2 | IPCM4V2 | S2 | Institute Pierre Simon Laplace | Dufresne (2009) |
| METO-HC-HadCM3C | HADCM3C | S2 | Hadley Centre | Johns (2009a) |
| METO-HC-HadGEM2 | HADGEM2 | S2 | Hadley Centre | Johns (2009b) |
| MPI-ECHAM5C | MPEH5C | S2 | Max Planck Institut | Roeckner (2008) |

The typical resolution of these global simulations is 150-300 kms, mainly constrained by the high computational cost required to undertake these simulations —increasing the model resolution by a factor 2 implies increasing the computational requirements by a factor 16.— Therefore, although each new generation of CMIP projections improve the resolution (typically a factor 2) aligned with the advances of high performance computing, they are still too coarse for impact studies in different sectors. Therefore, some sort of regionalization is needed in order to cope with local characteristics and to provide actionable information for impact studies, e.g. for the niche modeling community.

### 2.1.3 Regional Climate Models (RCMs)

Several factors prevent from the direct application of GCM outputs to local climate studies. In particular, their coarse horizontal resolution (hundreds of kilometers) is unable to represent local climate features. In order to bridge the gap between the large-scale variables provided by the GCMs and the local surface variables of interest, for instance the typical bioclimatic variables used in niche modeling (see Section 1.5), different *downscaling* (also known as *regionalization*) techniques have been developed in the last decades. Dynamical downscaling methods are based on Regional Climate Models (RCMs), which simulate regional features of the climate at a higher resolution over a limited area, driven at the boundaries by the GCM outputs (see Fig. 2.3, and Giorgi & Mearns, 1999). RCMs are physically consistent and provide a large number of variables describing the state of the atmosphere. The resulting regional/local *scenarios* are regarded as plausible descriptions of the future climate that reflect the influence of local topography and/or land-sea effects, and their interactions with changing synoptic-scale weather patterns under rising concentrations of greenhouse gases (Wilby & Dessai, 2010).

RCM projections cannot be directly used in impact studies, since they may contain significant biases (Christensen *et al.*, 2008b) inherited from the driving GCMs and also resulting from different physics and parametrizations involved in the formulation of the models. Thus, a bias adjustment/calibration process is necessary before using these data in real applications. This process requires the availability of historical data over the variables of interest, in order to calibrate the model outputs in a particular region. However, although several

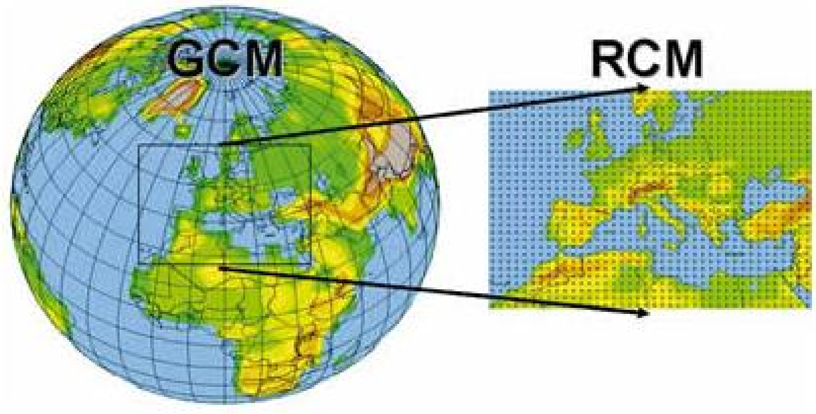

**Figure 2.3:** Schematic representation of the dynamical downscaling approach, based on a Regional Climate Model (RCM) embedded in a GCM grid. Image source: F. Giorgi, WMO Bulletin 52(2), April 2008.

bias adjustment methods have been recently proposed and have quickly became very popular (Déqué *et al.*, 2007), to date there is no completely satisfactory bias-correction method (Christensen *et al.*, 2008; Maraun, 2012). The common approach followed by the niche modeling community is change factors (also called *the delta method*, described in Section 2.2.1), which is suitable when only climatological values (e.g. the mean for a 30-years period) are needed.

### 2.1.4 The ENSEMBLES Regional Climate Projections Dataset

The regional climate change projections used in this Thesis were obtained from the ENSEMBLES project (van der Linden & Mitchell, 2009; Déqué *et al.*, 2012). ENSEMBLES is the latest in a series of EU-funded projects dealing with regional projection (dynamical downscaling) of large-scale climate simulations over Europe. An ensemble of state-of-the-art European Regional Climate Models (RCMs) was applied to produce regional projections from global climate change scenarios over Europe at 25km resolution. To this aim, the RCMs were forced with different boundary conditions, corresponding to the different historical and future scenarios (a detailed description of the experiments and results achieved in this project is published in a special issue of Climate Research, Christensen *et al.*, 2010).

**Table 2.2:** Summary of the ENSEMBLES RCMs. All ENSEMBLES simulations are publicly available through the DMI repository, in `http://ensemblesrt3.dmi.dk`.

| **Institution** | **Model** | **Reference** |
|---|---|---|
| C4I | RCA3.0 | Kjellström *et al.* (2005) |
| CNRM | RM4.5 | Radu *et al.* (2008) |
| DMI | HIRHAM5 | Christensen *et al.* (2006) |
| ETHZ | CLM | Jaeger *et al.* (2008) |
| KNMI | RACMO2 | van Meijgaard *et al.* (2008) |
| MetoHC | HadRM31 | Collins *et al.* (2006) |
| ICTP | RegCM3 | Pal *et al.* (2007) |
| Met.NO | HIRHAM | Haugen & Haakensatd (2005) |
| MPI-M | REMO | Jacob *et al.* (2001) |
| SMHI | RCA3.0 | Kjellström *et al.* (2005) |
| UCLM | PROMES | Sanchez *et al.* (2004) |

First, the RCMs were driven by different GCMs from the phase 3 of the Coupled Model Intercomparison Project (CMIP3, see Meehl *et al.*, 2007) during the same period (1961-2000), but considering the control twenty century greenhouse gas emission scenario 20C3M. In this scenario the GCMs perform a continuous run encompassing a historical period (approx. 1900-2001), considering the observed concentrations. Then, the RCMs were driven by the same GCMs in the transient period 2011-2050 (some models were ran until 2100), considering future climate conditions from the A1B SRES scenario. Table 2.3 shows the ENSEMBLES GCM/RCM combination matrix with four different GCMs in columns and ten ENSEMBLES RCMs arranged by rows (denoted with acronyms presented in Table 2.2). For each RCM, the red color indicates the "default" GCM used in the simulations (which in most cases correspond to the in-house GCM). The asterisk indicates those runs ending in 2050; the remaining combinations run until 2100.

The results from the historical scenario allow analysing the performance (e.g. the bias or the trends) of the different coupling of RCM-GCM, for a particular region of interest. Thus, each user may decide discarding those RCMs with poor performance (see, e.g. Herrera *et al.*, 2010; Turco *et al.*, 2013, for a comparison of RCMs in Iberia). Note that there is no day-to-day correspondence in these simulations and, hence, they are only expected to reproduce average climate conditions in climatic periods (typically 30 years) and inter-annual trends. Finally, the simulations in future scenarios provide the basis to obtain regional projections for a particular region of interest, after filtering the available data according to the previous validation results. For instance, Figure 2.4 shows the projected changes of total annual precipitation [%] (left) and annual mean temperature [K] (right) for 2071-2100 compared to 1971-2000, for A1B scenario (w.r.t. to 20C3M) as given by the ensemble mean of the ENSEMBLES regional projections dataset (Table 2.3). Besides the average information, this figure shows also the uncertainty obtained from the whole ensemble. Thus, hatched areas indicate regions with robust and/or statistical significant change, as given by the standard deviation of the ensemble. This image shows the change factors (or deltas) which could be used to obtain future climate information (e.g. adding them to the baseline climate).

This data set has been rarely used in SDM applications, presumably because several post-processing steps are necessary to make the data suitable for the modeling process (including the calculation of the bioclimate variables). Niche applications tend to consider special purpose datasets (such as WorldClim), with suitable variables and formats for this community. However, ENSEMBLES (and the follow-on EURO-CORDEX) constitute the state-of-the-art regional climate change projections in Europe and, therefore, this is in principle the most convenient dataset to be used for climate change applications. In this Thesis, we have developed tools to facilitate this task (see Chapter 7).

**Table 2.3:** The ENSEMBLES GCM/RCM combination matrix with four different GCMs in columns and ten ENSEMBLES RCMs aranged by rows (denoted with acronyms presented in Table 2.2). For each RCM, the red color indicates the "default" GCM. The asterisk indicates those runs ending in 2050; the remaining combinations run until 2100.

| GCM<br>RCM | HadCM3Q16 | HadCM3Q0 | ECHAM5-r3 | ARPEGE | BCM |
|---|---|---|---|---|---|
| **HadRM31** | X | X | | | |
| **REMO** | | | X | | |
| **RM4.5** | | | | X | |
| **HIRHAM5** | | | X | X | X |
| **CLM** | | X* | | | |
| **RACMO2** | | | X | | |
| **RegCM3** | | | X | | |
| **RCA3.0** | | | X | | X |
| **PROMES** | | X* | | | |
| **HIRHAM** | | X* | | | X* |
| **RCA3.0** | X | | X* | | |

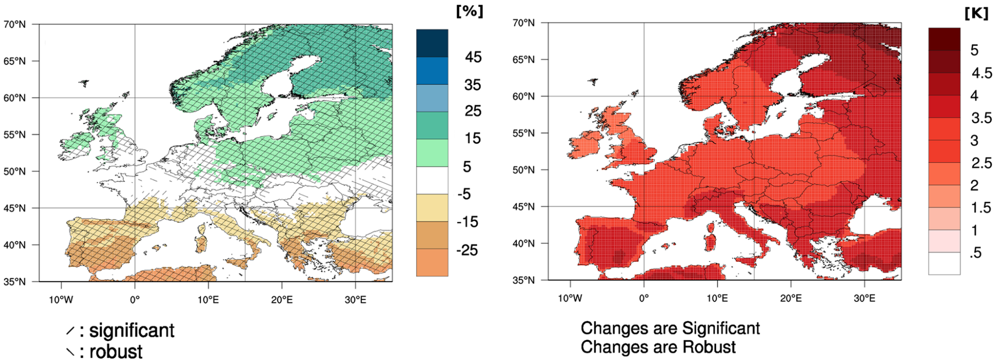

**Figure 2.4:** Projected changes of total annual precipitation [%] (left) and annual mean temperature [K] (right) for 2071-2100 compared to 1971-2000, for A1B scenario (w.r.t. to 20C3M). Hatched areas indicate regions with robust and/or statistical significant change. Image source: Adapted from Jacob *et al.* (2014).

## 2.2 Application for SDM Projections

### 2.2.1 The "delta" Method

The outputs of the GCMs (and/or coupled RCMs) cannot be used directly for impact studies given that they may contain important biases (e.g. Brands *et al.*, 2011). These biases can result from different physics and parameterizations involved in the formulation of the models. Thus, a validation/calibration process is needed before using this data in real applications. This process usually requires the availability of historical data (baseline climatologies, 1.5) to calibrate the model outputs in a particular region of interest.

Alternatively, the "delta" method is often applied in climate research (e.g., Winkler *et al.*, 1997; Zahn & von Storch, 2010) in order to extract the climate change signal ("delta") from model simulations. An advantage of this approach is that as climate change signal is computed relative to the control run of each model, the problem of the different climate model biases are alleviated to a great extent (e.g., Räisänen, 2007). The delta method operates by calculating the difference ("delta") between the GCM/RCM values for a variable of interest in a future period (e.g. 2071-2100) and in a control period (e.g. 1971-2000). Then, the "delta" values are added to the reference/historical climate values. The main objections against this method lie in the assumption of model bias stationarity, which cannot be guaranteed, particularly in the latest decades of the transient period, when model outputs need to be considered with caution (Maraun, 2012). However, it is a suitable method when working with relatively large time periods —for which climatological features are averaged—, as is the case of species distribution modeling. Thus, in this Thesis we applied the "delta" method for building the future climate projections.

### 2.2.2 Model Extrapolation and Transferability

Predicting into new regions or/and alternative climate scenarios raises important difficulties, such as extrapolating beyond the range of environmental conditions over which the model was calibrated. The ability of and SDM (given a set of presences, (pseudo-)absences and predictors) to predict or project the potential distribution of a species population into a non-sampled environment (e.g. future climate change projections or distinct geographic areas) is known as extrapolation capability or transferability.

Future distributions are projected under the assumption that current environmental range will be retained under climate change (Thuiller *et al.*, 2005). Thus, independently from the scenario and the GCM/RCM used, the SDM should be able to correctly reproduce the occupied range in the future. With this respect, over-predictions (underfitting) and over-parameterizations (overfitting) greatly affect models and could explain why two SDMs calibrated in the same species data could produce different projections in the future (Thuiller *et al.*, 2004).

In addition to the extrapolation capability of the modeling algorithm itself, the transferability of an SDM could be significantly affected by other methodological limitations, such as the availability and choice of appropriate predictor variables (Dormann *et al.*, 2008; Petitpierre *et al.*, 2016).

### 2.2.3 Uncertainty of Future Projections

In light of current global change, Species Distribution Models (SDMs) constitute an important tool to assist decision-making in environmental conservation and planning. Nevertheless, a wide range of uncertainties around the SDM projections directly affect their potential value and limitations, remaining their quantification as an ongoing challenge. A common technique to tackle different sources of uncertainty is based on producing ensembles encompassing the whole range of variability by considering the results derived from multiple SDMs, RCM/GCMs, baseline climate datasets, etc. (see, e.g. Araújo & New, 2007; Buisson *et al.*, 2010; Bagchi *et al.*, 2013; Baker *et al.*, 2015).

In particular, the relative contribution of SDMs (GLMs, MARS, MAXENT, etc.) to the total variability of the ensemble projections has shown to be the largest (Buisson *et al.*, 2010; Fronzek *et al.*, 2011; Garcia *et al.*, 2012), since results vary significantly depending on the technique used (GLMs, RF, MARS, etc.) and the model configuration (see, e.g. Araújo *et al.*, 2005; Beaumont *et al.*, 2008; Fronzek *et al.*, 2011). In this sense, the SDM ensemble approach has also limitations, since it assumes that all SDMs are equally transferable to climate change conditions, thus posing the risk of diluting insightful model signals with noise and error from less useful or defective SDMs forming the ensemble (Thuiller *et al.*, 2004; Peterson *et al.*, 2011). However, there is not an objective basis to perform a selection of various alternatives, since a proper validation of future SDM outputs is inherently impossible. Therefore, the provision of new methodologies assessing SDM transferability and helping to narrow the uncertainty range of future ensemble forecasts is of paramount importance.

### 2.2.4 Available Tools

The popularity of the open-source R language (R Core Team, 2015) and its statistical modeling and spatial analysis support has favored the development of specific, well-established and actively maintained packages for SDM construction and analysis, such as **sdm** (Naimi & Araujo, 2016), **biomod2** (Thuiller *et al.*, 2016), **dismo** (Hijmans *et al.*, 2017) and **SDMTools** (VanDerWal *et al.*, 2014), some of them also implementing pseudo–absence data generation and ensemble building utilities. For instance, both **sdm** and **biomod2** implement methods for building ensemble projections based on model performance in the calibration phase —e.g. by discarding or weighting the obtained results—. However, they are not oriented towards the analysis of components that add variability to the projected distributions in non-sampled environmental spaces (e.g. under climate change conditions) that can not be properly evaluated during model calibration, thus it is necessary further development of methods and tools for addressing the problem of SDM transferability and uncertainty in a straightforward manner. Therefore, is this Thesis we have developed a tool focusing on this issues (see Chapter 7).

## 2.3 Illustrative Example: Future projections

This Section is the continuation of the illustrative example shown in Section 1.7, where built SDMs —for each species and baseline climate datasets— are here projected to future climate conditions, considering future period 2011-2041 and the outputs from the MPI regional climate model (RCM). We applied the "delta" method, as illustrated by Figure 2.5, to alleviate the bias linked to the RCM (see Section 2.2.1). The extracted "deltas" where added to each baseline climate in order to obtain future climate projections of the same set of predictors considered in the calibration phase (Section 1.7).

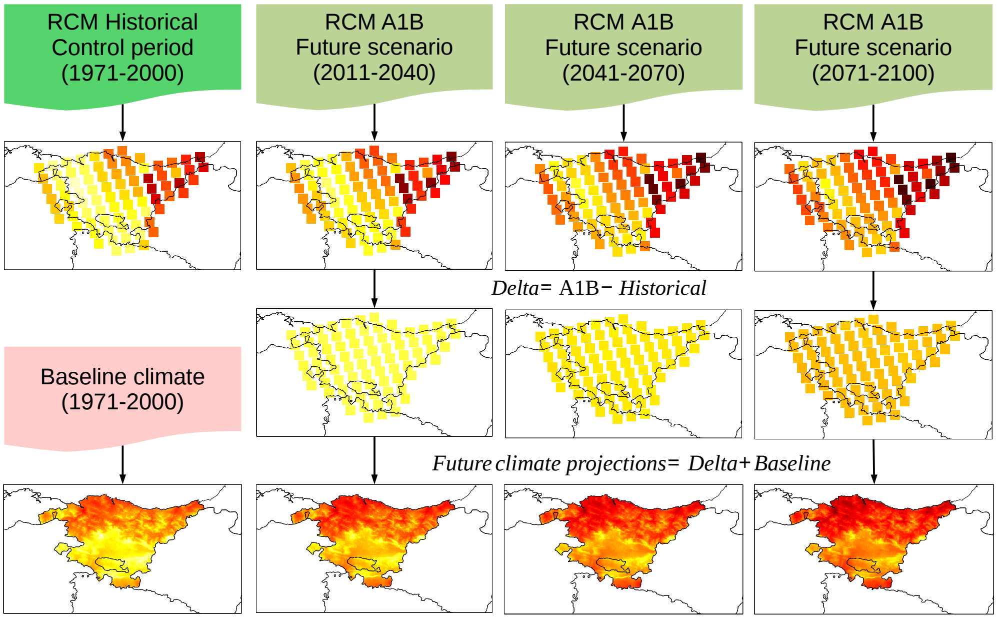
RCM Historical Control period (1971-2000)
RCM A1B Future scenario (2011-2040)
RCM A1B Future scenario (2041-2070)
RCM A1B Future scenario (2071-2100)
Delta= A1B− Historical
Baseline climate (1971-2000)
Future climate projections= Delta+Baseline

**Figure 2.5:** Conceptual diagram of the application of the "delta" method.

As a result, future suitability maps were obtained for each species and baseline climate dataset (Figs. 2.6 and 2.7). Predictions of the habitat suitability derived from each dataset were quite different according to the bias percentages obtained, that increased in future conditions (15.8 % and 21.4 % for *Fagus sylvatica* and *Quercus pyrenaica* respectively). Therefore, the baseline climate dataset constitutes an added source of uncertainty in SDM future projections that can not be assessed by relying on the model performance shown in the

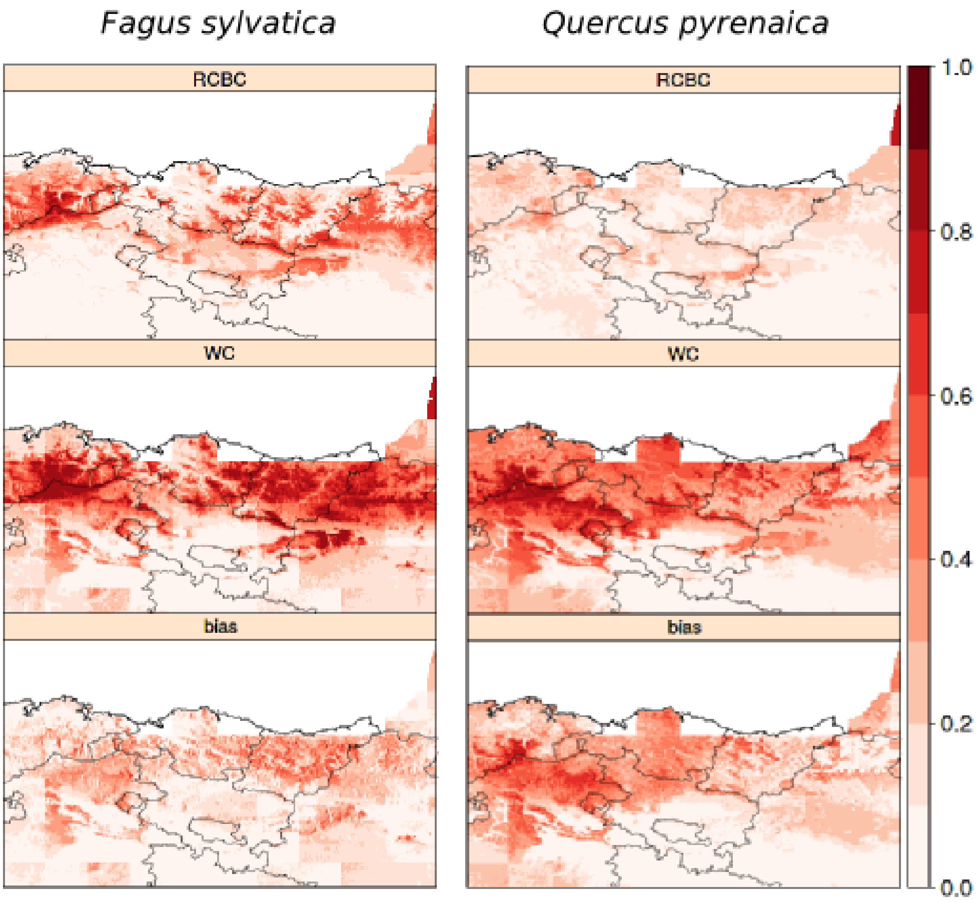

**Figure 2.6:** Future suitability maps of the SDM ensemble mean, corresponding to different baseline climate datasets (RCBC and WC) and tree species (*Fagus sylvatica* and *Quercus pyrenaica*). The absolute bias between both datasets is also shown (bias) for each tree species.

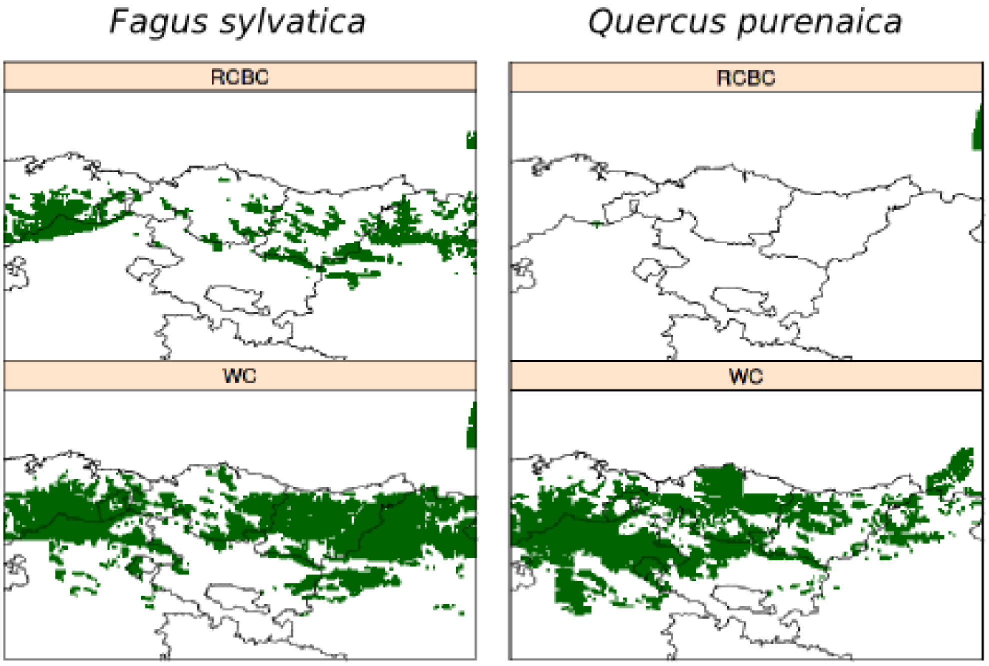

**Figure 2.7:** Future deterministic maps of predicted presence (green areas) and absence (white areas) of the SDM ensemble mean, corresponding to different baseline climate datasets (RCBC and WC) and tree species (*Fagus sylvatica* and *Quercus pyrenaica*).

model calibration phase (Fig. 1.5). In fact, despite reproducing better the climatic features of the region, RCBC showed lower performance values.

SDMs built from RCBC predicted greater habitat loss in the Basque Country. In this sense, using WC does not warn about the threat of habitat loss as a consequence of climate change to the same extent. This example stresses the importance of using quality climate data in regional studies.

# CHAPTER 3

# Objectives and Outline

Species Distribution Models (SDMs) are data-driven techniques widely used by the ecological niche modeling community to model and predict the distribution of biological entities in the geographical space (see Chapter 1). SDMs are based on empirical links established between absence/presence locations and the characteristics of their environment, including historical climate information typically in the form of bioclimatic variables (Guisan & Zimmermann, 2000; Elith & et al, 2006). A popular application of these models is the future projection of species distributions —from future climate projections, see Chapter 2— in order to assess key topics in environmental conservation such as monitoring biological responses to climate change (Hamann & Wang, 2006), species invasions (Jeschke & Strayer, 2008) or disease transmission (Drake & Beier, 2014) among others. These projections are being increasingly used by the vulnerability, impacts and adaptation (VIA) community, so communicating limitations, credibility and uncertainty in a comprehensive form is crucial for informing decision making processes (Gould *et al.*, 2014; Urban, 2015; Zhang *et al.*, 2015).

A number of sensitivity studies have been already performed considering ensembles of Species Distribution Models (SDM) formed by sampling different sources of uncertainty, such as the choice of multiple SDMs, the baseline climate datasets, the future emission scenarios and/or the global/regional (GCMs/RCMs) climate projections (see e.g. Araújo & New, 2007; Garcia *et al.*, 2012; Baker *et al.*, 2016, and references therein). In particular, it has been shown that SDMs have a large contribution to the total variability of the projections, since results vary significantly depending both on the technique used (GLMs, RF, MARS, etc.) and on the particular configuration (Buisson *et al.*, 2010; Fronzek *et al.*, 2011; Garcia *et al.*, 2012). For instance, a particular SDM built with different sets of predictors could project different probability distributions (Porfirio *et al.*, 2014; Pliscoff *et al.*, 2014).

Part of this uncertainty could be the result of diluting insightful SDM signals with noise from non–transferable (e.g. over-parameterized) SDMs with deficient extrapolation capabilities (Thuiller *et al.*, 2004; Peterson *et al.*, 2011). In addition, poor model configurations (e.g. the use of inadequate predictors Petitpierre *et al.*, 2016) could reduce significantly model transferability to different regions and/or changing climate conditions. Thus, in order to provide plausible actionable information to the VIA community it is necessary to narrow the uncertainty which can be attributed to methodological problems, including the above mentioned ones.

With this regard, the lack of reliable absence information poses several methodological problems for SDMs (Varela *et al.*, 2009). The generation of pseudo–absence data (in addition to the available presence one) has been proved to be an useful alternative to calibrate SDMs (Chefaoui & Lobo, 2008; Wisz & Guisan, 2009; Václavík & Meentemeyer, 2009); therefore, this approach is widely applied in SDM studies (see Section 1.4). For this purpose, different methodologies for pseudo–absence data generation have been proposed (e.g. Hengl *et al.*, 2009; Wisz & Guisan, 2009; Stokland *et al.*, 2011; Barbet-Massin *et al.*, 2012; Senay *et al.*, 2013; Iturbide *et al.*, 2015) attending to their performance in a sampled environment (using present climate information) that is typically assessed by measuring the accuracy that results from applying different cross-validation approaches in the calibration phase (see Section 1.6). However, similar accuracy scores can be obtained for dissimilar predicted distributions (Lobo *et al.*, 2010). In this context, if true-absences are missing, the accuracy measures can only indicate how well models discriminate data considered in the training process, but reveals little about their real predictive capability (Václavík & Meentemeyer, 2009). Furthermore, well performing SDMs may fail in extrapolating out-of-sample future climatic values and therefore, may not properly predict future species distributions (Fronzek *et al.*, 2011). However, the sensitivity of different SDMs to the sample of pseudo–absences when projecting on a non-sampled environment (e.g. under climate change conditions) has been neglected until now.

In this context, the following main objectives will be addressed through the Results of this Thesis:

1. To compare and assess the limitations of standard methods for pseudo–absence data generation in terms of model performance, considering a representative set of SDMs. Research will be also conducted for the development of new methods, focusing on new alternatives for the implementation of the background extent restriction.

2. To analyze pseudo–absence sampling as a determinant factor to characterize model stability and transferability in climate change conditions. This will be done by assessing the uncertainty in future ensembles of SDM projections (suitability maps) due to this factor. The interrelationship between predictors and pseudo–absences in this context will be also analyzed.

3. To develop an open-source modeling framework implementing the state-of-the-art SDM techniques, incorporating tools for pseudo–absence data generation and uncertainty analysis, envisaged to yield optimal future

estimates of habitat suitability. Special attention will be paid to the transparent connection with standard climate data repositories, thus helping to bridge the gap between the niche and the climate modeling communities. This package will be develop in R language.

# Part II

# Results

# CHAPTER 4

# Pseudo-absence Data Generation Methods

## 4.1 Introduction

Species distribution models (SDMs) most often require explicit absence information to adequately model the environmental space on which species can potentially inhabit. In the so called *background pseudo–absences* approach, absence locations are simulated in order to obtain a complete sample of the environment. Whilst the commonest approach is random sampling of the entire study region (Section 1.4), in its multiple variants, its performance may not be optimal. Moreover, the method of generation of pseudo–absences is known to have a significant influence on the results obtained. In this chapter we compare five pseudo–absence data generation methods (see Section 1.4), ranging from the classical random sampling of the whole region (RS) and the target group method (TG), to more sophisticated three–step techniques (TS), which limits the extent and the environmental range of the background from which pseudo–absences are sampled.

Regarding background extent restriction, Senay *et al.* (2013) proposed a variable importance change criterion based on principal component analysis, however, this strategy did not appropriately fit to the case studies presented in this Thesis. Therefore, here we propose a new criterion for optimizing background extent selection based on the theoretical properties of model performance as a function of distance to presence locations (Van der Wal & Shoo, 2009).

From an ecological perspective, the uncertainty associated to the presence of a biological entity is a combined effect of separate factors (biotic, abiotic and movement factors), that in turn depend on the environment of a specific site. In this context, the three–step method pursues the estimation of the fundamental distribution (regions of favorable abiotic factors) by the introduction of pseudo–absences within the niche space corresponding to areas of non-presence (outside the realized niche) and where movement factors are likely favorable (accessible geographic areas) but not so the abiotic factors (Peterson *et al.*, 2011). On the opposite, random sampling would produce predictions closer to a realized distribution, since it only excludes the presence locations for pseudo–absence data generation.

Here we consider 11 phylogenetic groups of Oak (*Quercus* sp.) described in Europe. We evaluate the influence of different pseudo–absence types on model performance (area under the ROC curve), calibration (reliability diagrams) and the resulting suitability maps, using a cross–validation approach (see Section 1.6).

The main results of this Chapter have been published in: Iturbide Iturbide, M., J. Bedia, S. Herrera, O. del Hierro, M. Pinto, and J. M. Gutiérrez, 2015: A framework for species distribution modelling with improved pseudo–absence generation. *Ecological Modelling*, **312**, 166–174, 10.1016/j.ecolmodel.2015.05.018

## 4.2 Methods and Materials

### 4.2.1 Presence Data and Study Domain

We consider genetically differenced groups of *Quercus* sp in Europe from the $GD^2$ database. Each group corresponds to a different chloroplast haplotype, determined by PCR analysis on more than 2600 populations of Oaks in Europe (see Petit *et al.*, 2002c,b,a). We considered 11 out of the total 42 Oak haplotypes, attending to the minimum population size needed to build the models ($n > 30$) while attending to the best possible representation of all European *Quercus* linages, excluding only one (linage F) out of five (Petit *et al.*, 2002b, Table 4.1). The study area was divided in 11 parts (in correspondence to each haplotype distribution) by defining a bounding box around the presence points (Fig. 4.1).

**Table 4.1:** Haplotypes considered ordered by decreasing sample size ($n$), and the lineages they belong to, according to the *Quercus* sp Europe database (Ehrenmann *et al.*, 2016). Only one linage (F) out of five was not included in the analyses due to insufficient sample size of all its haplotypes.

| Haplotype | Linage | n |
|---|---|---|
| H7 | A | 734 |
| H10 | B | 651 |
| H1 | C | 490 |
| H12 | B | 466 |
| H11 | B | 283 |
| H5 | A | 250 |
| H17 | E | 67 |
| H4 | A | 53 |
| H6 | A | 41 |
| H15 | E | 36 |
| H27 | D | 31 |

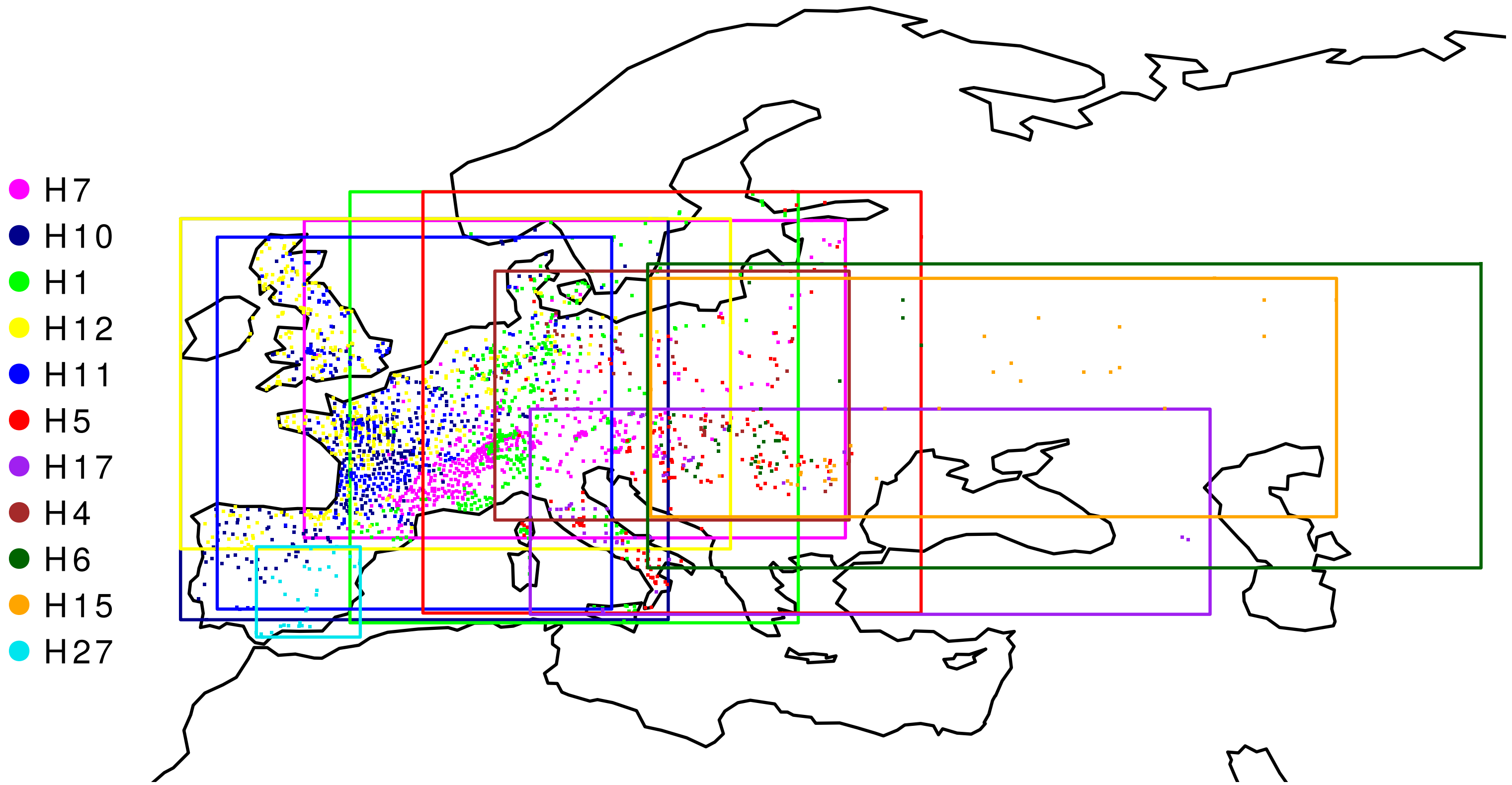

**Figure 4.1:** Phylogenetic distribution of *Quercus* sp in Europe. Oak groups in decreasing sample size order are: H7(n=734), H10(n=651), H1(n=490), H12(n=466), H11(n=283), H5(n=250), H17(n=67), H4(n=53), H6(n=41), H15(n=36) and H27(n=31).

### 4.2.2 Climate Data

We used the climatic variables of the WorldClim dataset (Hijmans *et al.*, 2005) at 10 km resolution as explanatory variables to build the SDMs. The chosen resolution is adequate to the aims of this study, given the 'false precision' provided by the downscaled WorldClim climate surfaces of 1 Km, as highlighted in previous niche modeling studies (Bedia *et al.*, 2013). After a pairwise cross-correlation analysis of the bioclimatic variables (following Bedia *et al.*, 2013), we retained a subset of uncorrelated predictors (BIO2, BIO03, BIO08, BIO13, BIO14 and BIO15, see Table 1.1) rescaled in the range [0,1].

### 4.2.3 SDM Development, Evaluation and Projection

SDMs were built using three different popular techniques, namely maximum entropy (MAXENT, Phillips *et al.*, 2006), generalized linear models (GLMs, Guisan & Zimmermann, 2000) and multivariate adaptive regression splines (MARS Friedman, 1991). Constrained by data availability, we resorted the use of a 10-fold cross validation approach to measure the area under the ROC curve (AUC) as the most widely used metric for model performance assessment. Models were also evaluated by calculating reliability diagrams.

For all methods tested we kept the number of pseudo–absences equal to the number of presences in all cases (prevalence = 0.5, Hengl *et al.*, 2009; Mateo *et al.*, 2010a; Hanspach *et al.*, 2011; Senay *et al.*, 2013). Additionally, a exclusion buffer of 10 km around the occurrence points was set in order to avoid cells containing both presence and pseudo–absence data (Chefaoui & Lobo, 2008). All steps involved in pseudo–absence generation according to the different methods tested are indicated in the diagram of Figure 4.2.

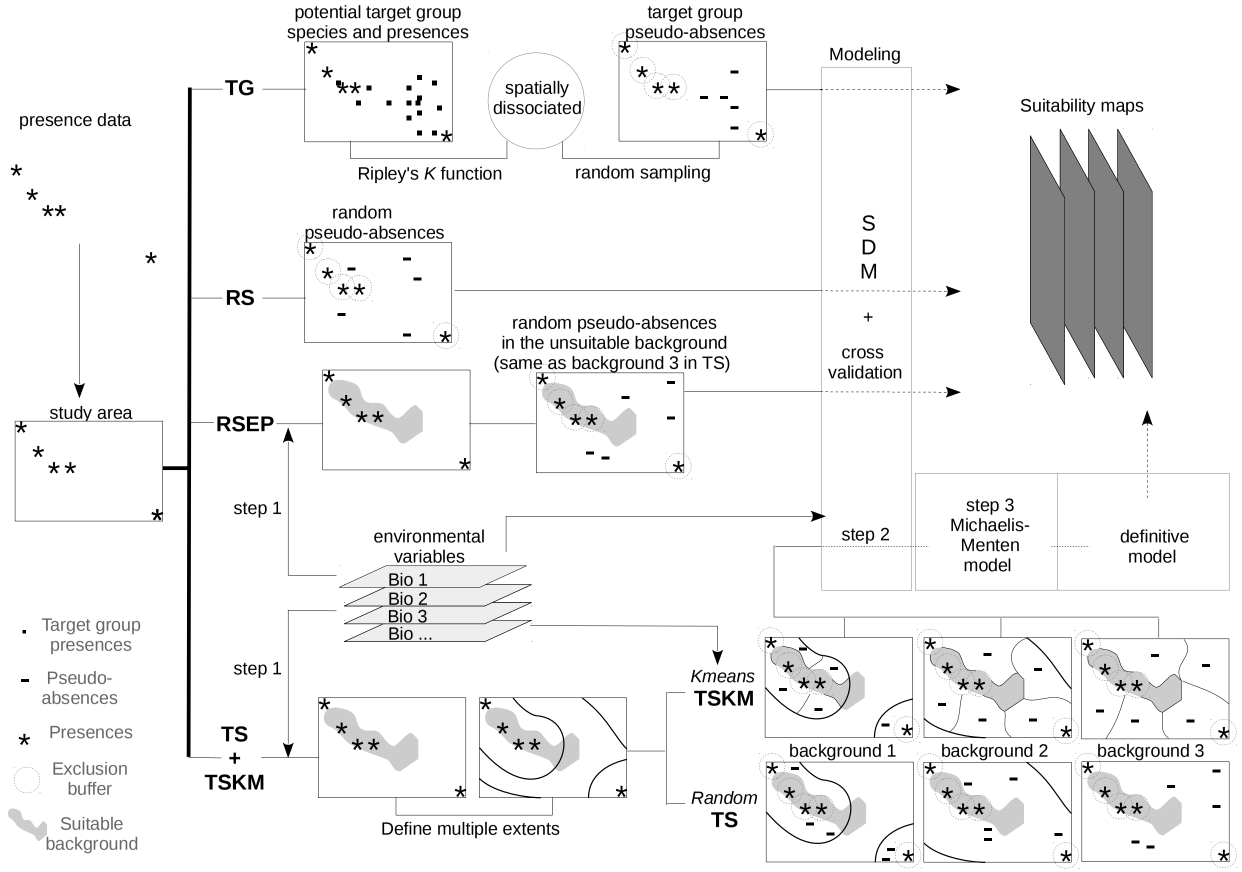

**Figure 4.2:** Conceptual diagram of the methodology used for generating pseudo–absences. Legend is shown in the bottom left corner.

### 4.2.4 Pseudo-absence Data

#### Random Selection (RS)

Pseudo–absences were sampled at random in the whole background, excepting the grid points within the exclusion buffer.

#### Random Selection with Environmental Profiling (RSEP)

The RSEP method is aimed at defining the environmental range of the background from which pseudo–absences are sampled. Environmentally unsuitable areas are defined using a presence–only profiling algorithm. To this aim, we run one–class support vector machines (OCSVM, Scholkopf & Smola, 2001) for each Oak group (see e.g. Drake *et al.*, 2006; Bedia *et al.*, 2011, for specific details on the use of support vector machines in SDM studies). OCSVM has been indicated as the most adequate algorithm for this purpose as it can handle high dimensional data and complex non–linear relationships between predictors (Senay *et al.*, 2013).

#### Three–step Selection (TS)

The TS method adds an additional step to the RSEP method to define the environmental range, and also the extent of the background from which pseudo–absences are sampled (Fig. 4.2). Thus, the first step is the definition of the environmentally unsuitable areas as is done in the RSEP method.

Regarding the limitation of the background extent, we applied a model performance criterion based on the findings of Van der Wal & Shoo (2009), that evaluated the relationship between the geographic extent from which pseudo–absences are taken and model performance, and found that the AUC rapidly increased as background size expanded from 10 to 100 km while subsequent expansions resulted in only minor increases in AUC. We found a similar behavior for all the groups of presence data considered in this Thesis, and concluded that the AUC *vs.* distance curve can be optimally fit to a non-linear asymptotic model. We tested the Michaelis-Menten model

$$v(x) = \frac{ax}{Km + x}, \tag{4.1}$$

the exponential of 2 parameters

$$v(x) = a(1 - e^{-bx}) \tag{4.2}$$

and exponential of 3 parameters

$$v(x) = a - be^{-cx}, \tag{4.3}$$

where $v$ and $x$ represent the AUC and the background extent respectively. $a$ is the asymptotic AUC value achieved by the system and $a - b$ is the intercept. $Km$ is the Michaelis constant (i.e. the extent at which the AUC is half of $a$), and $c$ is the coefficient of the point where the curve is most pronounced (Fig. 4.3).

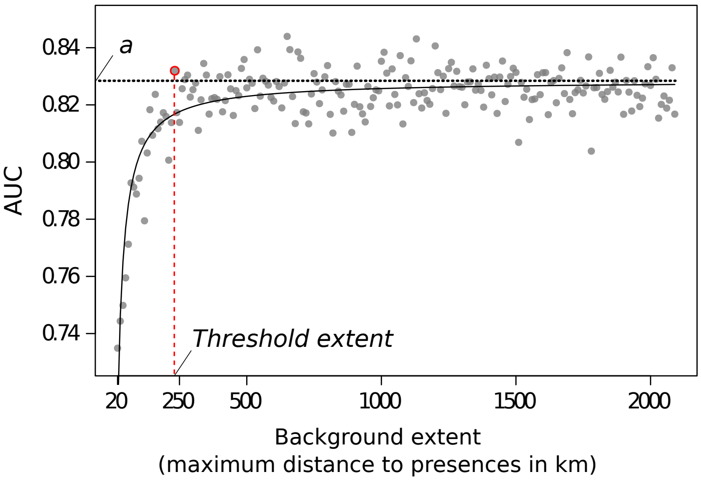

**Figure 4.3:** Relation of the AUC to the background extent for phylogeny H7. The black curve correspond to the fitted Michaelis-Menten model. $a$ represents the maximum AUC achieved by the system. The highlighted point corresponds to the smallest background extent greater than $a$ (i.e., the threshold extent). This relationship is similar to that described in Figure 2 in Van der Wal & Shoo (2009). All Oak groups in the study exhibited the same type of curve.

As a result, in this Thesis we propose a generalizable method to find the threshold extent that minimizes the distance to presences, without penalizing model performance, which constitutes the major novelty in comparison with previous published methodologies for pseudo–absence data generation.

Therefore, in the second step, random pseudo–absences are generated for different spatial extents within the unsuitability background zones defined in the first step. In order to consider all possible extents, we set different maximum *distance thresholds* to each presence location, considering a sequence from 20 km (twice the exclusion buffer) to the length of half diagonal of the bounding box that encloses the background of each Oak phylogeny (i.e. the maximum possible distance between any pair of points within the area).

Finally, in the third step alternative SDMs are built for all possible pseudo–absence configurations generated in step 2. Resulting AUCs and the different background extents tested are fitted to the curve of equations 4.1, 4.2 and 4.3 to extract the theoretical asymptotic AUC value ($a$). Then, the minimum threshold extent $x$ at which $AUC_x > a$ is chosen (Fig. 4.3), and the corresponding fitted SDM is retained to produce the suitability maps for the entire study area.

### *Three–step with k-means Selection (TSKM)*

The difference of TSKM with regard to TS is that, the pseudo–absences are taken from the spatial sub–units defined by a clustering on the background extent in Step 2 (Senay *et al.*, 2013). Instead of using a random selection on the unsuitable areas after Step 1, a k-means clustering is applied on the environmental and geographical space (k being equal to the number of presence points) and the coordinate values of each cluster centroid are retained, thus obtaining a regular distribution of dissimilar points for the study area which constitutes a representative sample of the unsuitable environment. Step 3 is then done as in the TS method. The resulting background extents for the TS and TSKM methods are listed in Table 4.2.

**Table 4.2:** Threshold distances to presences (kilometres) defining the background extents from which pseudo–absences are sampled. Each data in the column $d_{max}$ correspond to the length of the half diagonal of the bounding box that encloses the study area (Fig. 4.1), i.e.: the maximum possible distance between a pair of points within the study area.

| | $\mathbf{d}_{TS}$ | $\mathbf{d}_{TSKM}$ | $\mathbf{d}_{max}$ |
|---|---|---|---|
| H7 | 230 | 290 | 2090 |
| H10 | 500 | 670 | 2100 |
| H1 | 580 | 800 | 2070 |
| H12 | 620 | 620 | 2130 |
| H11 | 390 | 560 | 1800 |
| H5 | 190 | 240 | 2170 |
| H17 | 690 | 830 | 2360 |
| H4 | 150 | 380 | 1440 |
| H6 | 1000 | 1050 | 2950 |
| H15 | 360 | 80 | 2420 |
| H27 | 30 | 70 | 450 |

### *Target Group Selection (TG)*

In order to select a target group for each phylogenetic Oak group we searched for presence records of species not belonging to the *Fagaceae* family in the database of The Global Biodiversity Information Facility (GBIF, `http://data.gbif.org`). To ensure a sufficiently high number of presence points, we focused on species with a widespread distribution in Europe as target group candidates.

For each candidate and Oak group, we computed the cross type of the Ripley's $K$ function (Dixon, 2006) to analyze the spatial behavior of the point pattern. From the estimated Cross K-functions, those showing spatial dissociation of the TG candidate with regard to the Oak group were chosen (see Grantham, 2012, for wider explanation regarding point pattern analysis and Rypley's $K$ function interpretation), resulting in the following target groups: *Ulex europaeus* for groups H3 and H11; *Picea glauca* for groups H1, H2, H4, H5, H6 and H8; *Pinus nigra* for groups H7 and H10; *Pinus strobus* for group H9. TG locations were then randomly sampled to match the number of Oak localities in order to obtain balanced datasets for model training.

### 4.2.5 Implementation and Tools

Bioclimatic variables and MAXENT models were calculated by means of the R package **dismo** (1.0-12, Hijmans *et al.*, 2017). We used the MARS algorithm implementation of the R package **earth** (v4.4.0, Milborrow, 2015). Several raster data operations and representation were done using the **raster** package (v2.3-40, Hijmans, 2015).

## 4.3 Results and Discussion

### 4.3.1 TG Method

TG attained the highest AUCs for almost all the phylogenetic groups (Table 4.3, Fig. 4.4), but in turn it yielded poorly calibrated models (Fig. 4.5), with a strong under-estimation of high probability values. We argue that these results are due to the spatially clustered distribution of targeted group presences used as pseudo–absences, leading to spatially autocorrelated background samples resulting in inflated AUC values (González *et al.*, 2011), and also to an over-estimated suitability for a large proportion of non-sampled areas (Figs. 4.7 and 4.6), as compared to the other methods. Phillips *et al.* (2009) and Mateo *et al.* (2010a) recommended the TG pseudo–absence as the best method for discrimination, resulting in models with the best predictive performance. We find the same result, with TG attaining the highest AUC values, although this comes at the cost of a poor model calibration, and therefore we do not recommend this technique if reliable suitability maps are to be obtained. This stresses the importance of well-distributed presence/absence data across the environmental and geographical space of the study area in order to obtain reliable models (Lobo & Tognelli, 2011).

**Table 4.3:** Multimodel mean AUC values, according to the four pseudo–absence generation methods tested, for each of the Oak groups analyzed. Values for TG method are underlined when they are the best of all methods. Values in bold are the maximum AUC values excluding the TG method.

| | **RS** | **RSEP** | **TS** | **TSKM** | **TG** |
|---|---|---|---|---|---|
| H7 | 0.771 | **0.834** | 0.832 | 0.830 | 0.981 |
| H10 | 0.772 | 0.854 | 0.851 | **0.856** | 0.970 |
| H1 | 0.764 | 0.822 | **0.823** | 0.820 | 0.976 |
| H12 | 0.781 | 0.839 | **0.864** | 0.852 | 0.971 |
| H11 | 0.760 | 0.815 | 0.842 | **0.846** | 0.985 |
| H5 | 0.786 | **0.830** | 0.829 | 0.828 | 0.977 |
| H17 | 0.798 | 0.847 | 0.878 | **0.897** | 0.935 |
| H4 | 0.720 | **0.873** | 0.835 | 0.824 | 0.962 |
| H6 | 0.802 | 0.847 | **0.862** | 0.859 | 0.939 |
| H15 | **0.762** | 0.668 | 0.748 | 0.707 | 0.941 |
| H27 | 0.726 | **0.843** | 0.741 | 0.677 | 0.712 |

### 4.3.2 RSEP, TS and TSKM Methods

RSEP and three–step methods (TS and TSKM) attained similar results. As expected, we did not find any significant differences in their AUCs (Fig. 4.4, Table 4.3) since both TS and TSKM define a threshold extent based on the asymptotic AUC value $Vm$ (Fig. 4.3), close to the expected value of the maximum distance threshold used by the RSEP method. With this regard, TS and TSKM methods are preferable than RSEP, since using the theoretical AUC value given by $Vm$ ensures the selection of a good model, while RSEP method may result in a sub-optimal model if the last point in the X-axis lies significantly below the $Vm$ value by chance (Fig. 4.3).

The suitability plots (Fig. 4.6) show a similar behaviour, clearly different from RS and TG. Thus, we conclude that the relevant step that affects SDM results is the environmental profiling of the background, which constitutes the common characteristic of the RSEP and three–step methods. As a result, RSEP

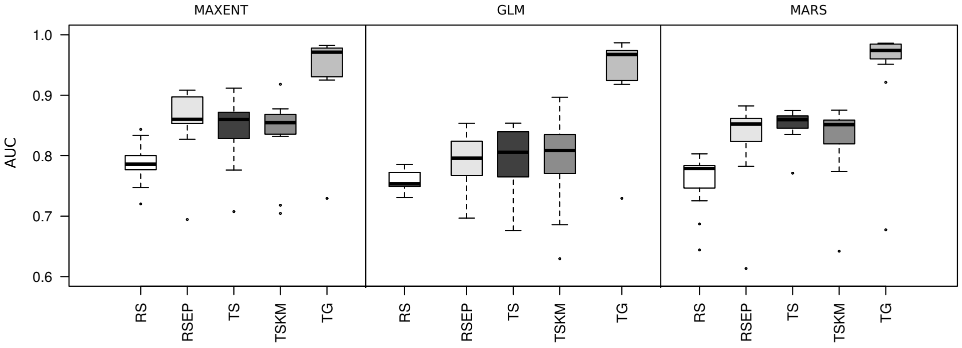

**Figure 4.4:** AUC box–plots of the 11 oak groups modeled with the five pseudo–absence generation methods for each modeling technique. Oak groups were modeled with higher accuracy by MAXENT and MARS. The average AUC values improved for all modeling techniques when using a different method from RS.

was equally effective while entailing a more straightforward implementation. Analogously, since the background extent restriction does not impair final results, three–step methods are also recommendable as the effect of non informative pseudo–absences from far regions could be significant in other case studies, especially when a wider study area is considered. In this sense, several authors argue that pseudo–absences from far regions should be avoided (Van der Wal & Shoo, 2009; Anderson & Raza, 2010). Moreover, Jiménez-Valverde *et al.* (2008) and Lobo *et al.* (2010) suggested that pseudo–absences should be located near the external boundary of the suitable environment to adequately represent the potential distribution of a species. At this respect, we consider that the three–step method proposed in this study satisfies this requirement while avoids misleading models with reduced AUCs. Finally, since the TSKM method does not improve SDM results in relation to TS, the introduction of the k-means clustering in Step 2 of TSKM can be skipped in favour of a simple random selection within the background extent.

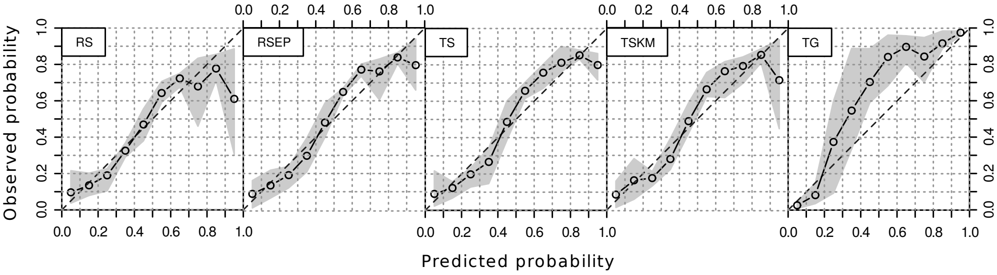

**Figure 4.5:** Calibration plots of the multimodel predictions. Points connected by lines are the mean obtained from the different Oak groups and the grey area correspond to the range between maximum and minimum values. Values below the diagonal indicate over-estimated probabilities and values above it under-estimated predictions. The smallest Oak groups H4(n=53), H6(n=41), H15(n=36) and H27(n=31), are excluded in the calibration plots, because their low sample size systematically yields poorly calibrated models that mask observable differences between methods.

### *4.3.3 RS Method vs. RSEP, TS and TSKM Methods*

The RS method produced well calibrated SDMs, excepting in the zones of higher environmental suitability, where the latter was over-estimated for all Oak groups (Fig. 4.5). This is due to the fact that many pseudo–absences are distributed around presences inside the potentially suitable environment, resulting in a lower rate of observed presences against absences in the zones predicted as most suitable, and is arguably one major disadvantage of the RS method with regard to methods applying environmental profiling as a previous step (RSEP, TS and TSKM). Furthermore, RS yielded the worst discrimination results, with the lowest AUC values for all algorithms tested (Fig. 4.4) and for most Oak groups (Table 4.3).

The use of a profiling technique as an intermediate step, characteristic of the three-step methods (TS and TSKM), has been criticized by some authors for producing artificially high probabilities of occurrence (Wisz & Guisan, 2009; Stokland *et al.*, 2011) and wider predicted suitability areas. In ecological terms, the variability in the predicted probabilities is related to the ability of the SDMs to represent realized *vs.* potential species distributions, lying spatially

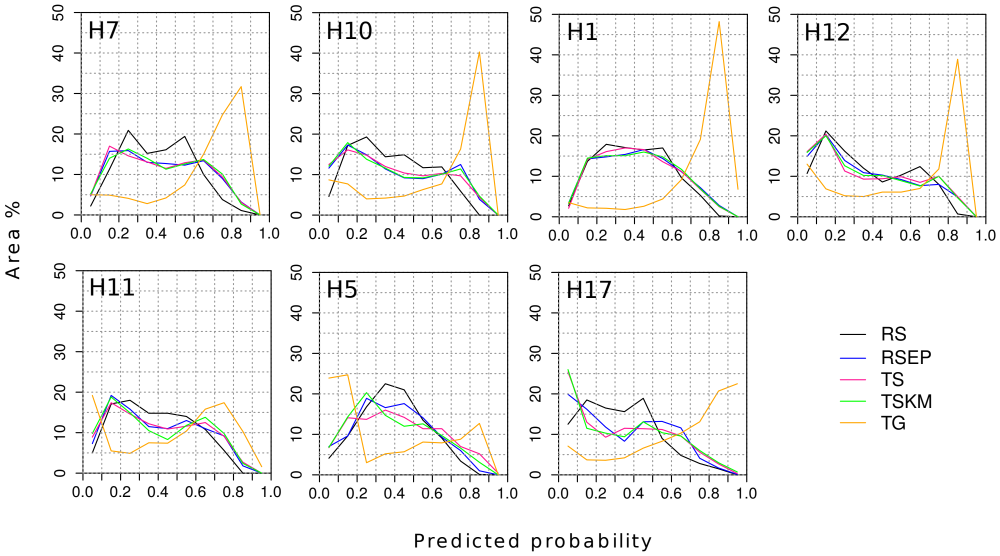

**Figure 4.6:** Suitability plots. Percentage of area predicted into each interval of probability of occurrence for the Oak groups producing well calibrated models (see Figure 4.5). These graphics give quantitative information on the suitability maps for a better interpretation of the results obtained. The first plot (H7) correspond to the suitability maps shown in Figure 4.7. Compared to RS, the RSEP, TS and TSKM methods produce incremented areas of high and low suitability and reduced mid suitable areas. The TG method predicts large areas of high suitability.

wider predicted distributions closer to the fundamental niche of the target species (Chefaoui & Lobo, 2008). However, since the potential distribution of the species is uncertain, we see no reason to penalize the model based on the extent of the area predicted as suitable (see e.g. Jiménez-Valverde, 2012). Furthermore, our results indicate that the predicted potential areas are not significantly shrink/widened with the use of either profiling/RS techniques (they are though in case of TG method, Fig. 4.7). In fact, the most remarkable difference between both is a higher resolution of the profiling-based models as compared to RS for most Oak groups, as depicted by the suitability plots (Fig. 4.6). This means that ambiguous probabilities (around 0.5) are less likely to occur when RSEP or three–step methods are introduced, in favor of more informative predicted probabilities closer either to 1 or to 0, as opposed to the traditional RS approach. (see e.g. Bedia *et al.*, 2011, for a more detailed explanation of model resolution in the context of SDMs). This is particularly important in order to reduce uncertainties when binary presence/absence maps are required for decision making and/or management plans.

Furthermore, the lack of records from suitable regions may simply derive from an inadequate sampling (Anderson, 2003; Hanspach *et al.*, 2011). In fact, presence data is quite often environmentally biased (Bierman *et al.*, 2010) resulting in presence data that does not represent the whole environmental range of the realized niche. In these cases, the RS method introduces false absences (within both the realized and fundamental niches) introducing a major source of uncertainty (Lobo *et al.*, 2010) and resulting in over-constrained areas of high suitability (Fig. 4.6). In this sense, as long as RSEP, TS and TSKM methods sample pseudo–absences within a previously profiled unsuitable area, the risk of introducing false pseudo–absences is minimized, even in the case of relatively biased species collections. On the other hand, in case of error in the initial presence data (e.g. false positives), then profiling techniques may bear the risk of further reinforcing this bias rather than correcting it, although this particular situation should be further investigated.

### 4.3.4 Sensitivity of Model Performance to the Pseudo–absence Generation Method

Our results show that the method of pseudo–absence generation strongly conditions output SDMs. Whilst the choice of the SDM algorithm is generally recognized as the principal factor of uncertainty (see e.g. Buisson *et al.*, 2010; Fronzek *et al.*, 2011), in this case study we demonstrate that pseudo–absence sampling design is even more important, leading to a larger variation of model AUC (Fig. 4.4, Table 4.3) than the modeling algorithms tested or the initial presence dataset choice, even though MAXENT and MARS performed better than GLMs (Fig. 4.4), indicating that algorithm selection is also an important factor (Phillips *et al.*, 2009; Bedia *et al.*, 2011; Senay *et al.*, 2013). Our results also suggest that MARS performance was more sensitive to the pseudo–absence

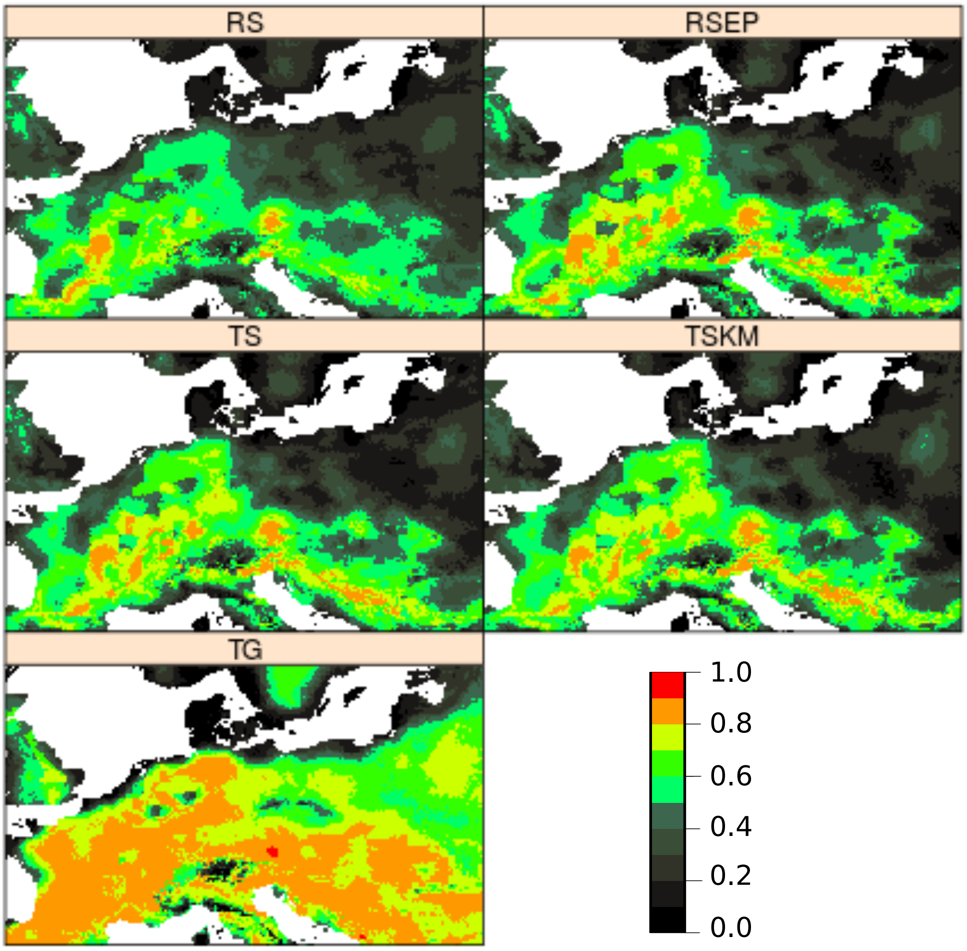

**Figure 4.7:** Multimodel suitability maps according to the five pseudo–absence generation methods tested for Oak group H7. Maps for the rest Oak groups show the same pattern on the prediction change between methods as is shown in Figure 4.6. Suitability is here expressed as a probability of occurrence given the environmental conditions, in the range [0,1].

configuration than MAXENT (Fig. 4.4), although a more intensive testing beyond the scope of this study is required to ascertain the sensitivity of different algorithms to the pseudo–absence generation scheme (Chapter 5).

### *4.3.5 Sample Size Effect on Results*

As sample sizes are heterogeneous across Oak groups, this allowed us to indirectly evaluate the influence of the sample size in the performance. Caution has to be given to interpreting inflated AUC values due to small number of records (Wisz *et al.*, 2008). For instance, Hanspach *et al.* (2011) excluded species with less than 50 records to allow reliable modeling. In this study, the calibration analysis shows that group H4 (53 presence records) and smaller groups (Table 4.1), did not produce reliable models for any of the pseudo–absence generation methods compared (not shown), even though AUC values were generally high (Table 4.3). In addition, the poor performance of the models for the smallest Oak groups (H15 and H27) is also reflected in the relationship of AUC and background extent, resulting in poor model fits in the TS and TSKM methods (equation 4.1) and yielding small threshold extents and lower AUCs (Tables 4.2 and 4.3).

# CHAPTER 5

# On the Impact of Pseudo–absences in Future Climate-Driven Projections

## 5.1 Introduction

Climate change projection ensembles from SDMs are strongly conditioned by different sources of uncertainty that decrease their potential informative value. In addition to the variability derived from alternative climate change scenarios, methodological aspects involved in SDM applications have the potential to affect model transferability and increase the variability of the projected future distributions, contributing significantly to the overall uncertainty. An important source of uncertainty often neglected in climate change studies comes from the use of background data (a.k.a. pseudo–absences) for model calibration. In this Chapter, we study the sensitivity to the pseudo–absence sample as a determinant factor for SDM stability and transferability.

The goal of this work is to assess the impact of pseudo–absences in SDM applications for climate change studies. For this purpose, we explore the range of uncertainty in SDM future projections derived from ten realizations of pseudo–absence data, using the distribution of a *Quercus robur* L. phylogeny in Europe as case study, and considering several pseudo–absence generation methods, SDM techniques and regional future climate projections (RCMs).

## 5.2 Methods and Materials

### 5.2.1 Presence/pseudo–absence Data

Here we use the distribution of a *Quercus robur* phylogeny ($GD^2$ database, Ehrenmann *et al.*, 2016), consisting in oak occurrence data that corresponds to chloroplast haplotype H7 (n = 359) and belongs to genetic linage A (Fig. 5.1). The main reason for the choice of this particular haplotype was its wide distribution and the greater number of samples available, thus improving model robustness. More details on the oak genetic lineages can be found in Petit *et al.* (2002c,b,a).

For analysis purposes, we divided the study area according to the climatic regions defined in the EU-funded PRUDENCE project (Christensen & Christensen, 2007). With respect to the distribution of phylogeny H7, in this study we defined as "peripheral" regions MD, IP, BI and SC (Fig. 5.1).

From the pseudo–absence generation methods evaluated in Chapter 4, here were considered the simplest (RS method) and the most elaborated (TS method), in order to encompass the full range of complexity at this respect and to analyze a possible influence on the results.

Based on the recommendations provided by Barbet-Massin *et al.* (2012), we considered the cases of using the same number of pseudo–absences as presences ($n = 359$) and three times more pseudo–absences than presences ($n = 1077$). Additionally, in order to further analyze the effect of prevalence (proportion of

The main results of these Chapter were submitted to *Global and Planetary Change* in May 2017 and were under review when the Thesis was printed.

presences *vs.* absences) on the results we also considered $n = 718$ and $n = 1795$ (two and five times the number of presences respectively). In order to minimize the false absence ratio, pseudo–absences were generated setting an exclusion buffer of 25 Km (i.e. one grid cell) around the occurrence points (Chefaoui & Lobo, 2008).

Although Barbet-Massin *et al.* (2012) recommended a minimum of ten realizations of pseudo–absences, this has rarely been performed in previous studies. In this work, we computed ten realizations for each of the two generation methods and each prevalence setting, and used them independently to train each of the three different SDMs.

### 5.2.2 Climate Data

Observational data for the reference period 1971-2000 was obtained from the E-OBS gridded observational dataset (Haylock *et al.*, 2008, v14), providing historical information of daily temperature and precipitation for Europe over a regular 0.22 grid. Using E-OBS data, we calculated a set of 19 standard bioclimatic variables (see e.g. Hijmans & Graham, 2006).

After a pairwise cross-correlation analysis of the resulting bioclimatic variables (following Bedia *et al.*, 2013), we discarded variables highly cross-correlated ($r > 0.9$).Then, we performed a stepwise (forward and backward) variable selection procedure using GLM, and retained a subset of variables that are relevant for all pseudo–absence realizations (see Chapter 6 for a more detailed description), these are: BIO1, BIO4, BIO5, BIO9, BIO15, BIO18 and BIO19 (Table 1.1).

Climate projections were obtained from the Regional Climate Model (RCM) simulations of the project ENSEMBLES (van der Linden & Mitchell, 2009, `http://www.ensembles-eu.org`) over the same 0.22 grid, under the historical emissions scenario (20C3M, period $1971 - 2000$) and the A1B transient emissions scenario (period $2001 - 2100$). We considered seven future climate scenarios generated by a subset of RCM-GCM couplings (Table 5.1), discarding

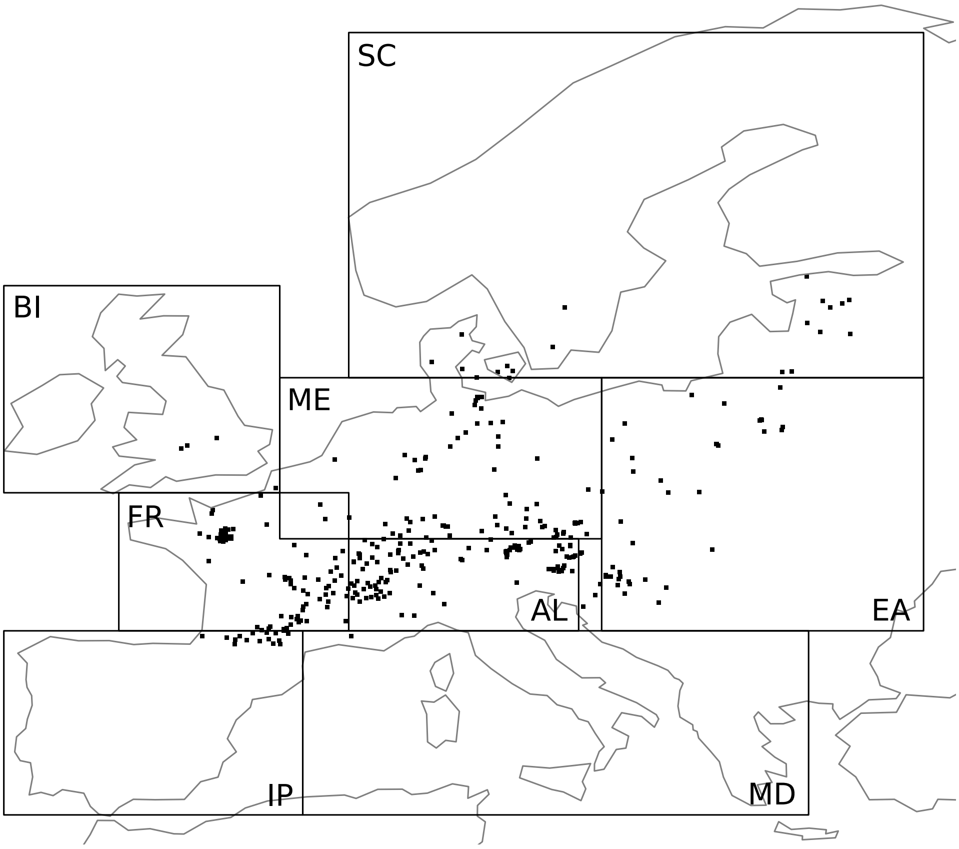

**Figure 5.1:** Distribution of phylogeny H7 (n=359) (*Quercus robur*) in Europe, and climatic regions defined in PRUDENCE: (MD) Mediterranean; (IP) Iberian Peninsula; (BI) British Isles; (SC) Scandinavia; (EA) Eastern Europe; (ME) Mid-Europe; (AL) Alps; (FR) France. Taking as reference the distribution of phylogeny H7, in this paper we consider as peripheral regions MD, IP, BI and SC.

those that have been shown to have large biases for particular GCM couplings (Turco *et al.*, 2013).

We calculated the future projected bioclimatic variables applying the "delta" method to the climatologies of max/min temperatures and precipitation (see, e.g., Räisänen, 2007; Zahn & von Storch, 2010, for a description and application of delta method). According to this, the historical simulation $(1971-2000)$ was subtracted from the future period climatology $(2071-2100)$ for each member to obtain the change signals ("deltas", see Section 2.2.1). The "deltas" were then added to the baseline (E-OBS) climatology at a grid-box level, obtained as the difference/ratio of the temperature/precipitation values in the future period. We then calculated the future bioclimatic variables from the resulting future temperature/precipitation climatologies.

**Table 5.1:** Regional climate models from the ENSEMBLES project used in this study. See Tables 2.2 and 2.3.

| **Acronym** | **RCM** | **Driving GCM** | **Reference** |
|---|---|---|---|
| CNRM | ALADIN | ARPEGE | Radu *et al.* (2008) |
| DMI | HIRHAM | ARPEGE | Christensen *et al.* (2008b) |
| ETHZ | CLM | HadCM3Q0 | Jaeger *et al.* (2008) |
| HC | HadRM3Q0 | HadCM3Q0 | Haugen & Haakensatd (2005) |
| ICTP | RegCM3 | ECHAM5-r3 | Pal *et al.* (2007) |
| MPI | M-REMO | ECHAM5-r3 | Jacob (2001) |
| SMHI-BCM | RCA | BCM | Samuelsson *et al.* (2011) |

### 5.2.3 SDM Development, Evaluation and Projection

SDMs were built using generalized linear models (GLMs, Guisan *et al.*, 2002), multivariate adaptive regression splines (MARS, Friedman, 1991) and random forest (RF, Breiman, 2001). For all prevalence settings, model fitting was done with equal weighting of presences vs pseudo–absences (i.e. the total weight of all presences is the same as the total weight of all pseudo–absences, see section 5.2.5).

Constrained by data availability, we resorted to a 10-fold cross-validation approach (Steyerberg *et al.*, 2010) in order to assess model performance. We calculated four metrics used in previous studies as suitable criteria for addressing the best formula of pseudo–absence data generation (Barbet-Massin *et al.*, 2012) and model transferability (Petitpierre *et al.*, 2016). These are 1) AUC (area under the receiver operating characteristic curve), 2) TSS (true skill statistic), 3) Sensitivity and 4) the Boyce Index (Fig. 5.2). The latter two, rely solely on predicted *vs.* observed presences (see Section 1.6 for details about the different accuracy measures).

Finally, models fitted with each pseudo–absence realization (10 levels) were projected into reference (1971-2000) and future (2071-2100) conditions to obtain probability maps of the potential distribution (i.e. suitability maps ranging from 0 to 1) for each particular SDM (3 levels) and regional climate projection (RCM, 7 levels). This was done for each method (TS and RS) and prevalence considered, resulting in a total of $10 \times 2 \times 3 \times 7 = 420$ members, representing probability maps of the future potential distribution for each prevalence setting (4 levels) and pseudo–absence generation method (2 levels).

### 5.2.4 Uncertainty Derived from Pseudo–absence Data

The uncertainty was analyzed by computing the range among projected suitability probabilities in every grid cell (location), and calculating the variance explained by the pseudo–absence realization in front of the SDM and the RCM. On the one hand, the range was obtained as the maximum–minimum difference of the ten pseudo–absence realizations (hereafter referred to as *sensitivity range*), for each SDM and climate projection combination (Figs. 5.3 and 5.4).

The relative contribution of each component to the total ensemble spread (variability) was assessed using a simple analysis of variance approach, where the total variance ($V$) can be decomposed as the summation of the variance explained by the realization ($P$), the RCM ($R$) and the combination of the previous two ($PR$):

$$V = P + R + PR. \tag{5.1}$$

Following the notation in Déqué *et al.* (2012) and San-Martín *et al.* (2016), let $i$ be the index of the pseudo–absence realization ($i = 1, ..., 10$), $j$ the index of the RCM ($j = 1, ..., 7$), and $X_{ij}$ is the response (e.g., predicted distribution for the particular realization and climate projection). Then,

$$P = \frac{1}{10}\sum_{i=1}^{10}(X_i - \bar{X})^2 \quad \text{and} \quad R = \frac{1}{7}\sum_{i=1}^{7}(X_j - \bar{X})^2 \tag{5.2}$$

are the terms resulting from the realization alone ($P$), and RCM alone ($R$), and

$$PR = \frac{1}{10}\sum_{i=1}^{10}\frac{1}{7}\sum_{i=1}^{7}(X_{ij} - X_i - X_j + \bar{X})^2 \tag{5.3}$$

is the interaction term of the realization with the RCM ($PR$).

We also computed the variance resulting from the pseudo–absence realization relative to the variability explained by the SDMs ($j = 1, ..., 3$). In order to illustrate thoroughgoing information on the spread in the projected potential distributions, variance percentage maps are shown together with the maps of the mean ($\bar{X}$ in Equations 5.2 and 5.3) and the standard deviation (square root of V in Equation 5.1)(Figs. 5.5 and 5.6).

Finally, in order to summarize the results, the spatial mean of the variance percentage was computed for each PRUDENCE region (Fig. 5.7).

### 5.2.5 Implementation and Tools

All the analysis performed in this study were undertaken using the open source R software for statistical computing (R Core Team, 2015). Climate data was loaded and handled using the package **loadeR** (v0.1-0, `https://github.com/SantanderMetGroup/loadeR/wiki`). Bioclimatic variables were calculated using the R package **dismo** (v1.0-15, Hijmans *et al.*, 2017).

In connection to pseudo–absence sample size, Barbet-Massin *et al.* (2012) recommended using 1000 pseudo–absences with equal weight to presences when 10 realizations are computed for GLM fitting. In the case of RF and MARS, less pseudo–absences are recommended, since by the time of the correspondent analysis, the weighting option for these two algorithms was not available in the particular R implementations used. In this case, we used the MARS algorithm implementation of a newer version of the R package **earth** (v4.4.4, Milborrow, 2015) and the RF algorithm implementation of the R package **ranger** (v0.6.0, Wright, 2016), both including a suitable weighting option. This allowed to perform a fair model fitting with all tested SDMs for the different prevalence settings considered, without penalizing the resulting probability distributions.

## 5.3 Results

### 5.3.1 Model Performance

RF achieved the best performance scores, followed by MARS, being GLM the technique showing lowest performance (Fig. 5.2). Regarding the method for pseudo–absence generation, in agreement with previous studies (e.g. Senay *et al.*, 2013; Iturbide *et al.*, 2015) and the results obtained in Chapter 4, TS achieved higher scores of model performance, except for some SDMs for sensitivity and the Boyce index (e.g. sensitivity by RF or Boyce index by GLM). Although RS shows lower performance, it is the most widely used method due to its simplicity (Iturbide *et al.*, 2015), and provides more easily interpretable results, avoiding possible effects derived from intermediate steps in the generation of pseudo absences. Therefore, hereinafter, we will mainly describe and illustrate results corresponding to the RS method, although results obtained for the TS method are also commented.

Figure 5.2 shows that different prevalence settings yield a similar performance. However, the sensitivity ranges of the resulting projections were higher when less pseudo–absences were used ($n = 359$, not shown), as the non-sampled background is wider and thus, the variability among realizations is larger. This results in projections with higher uncertainty (i.e. higher sensitivity range and standard deviation). Therefore, in the following we mainly illustrate the results obtained when using 1077 pseudo–absences with equal weight of presences *vs.* pseudo–absences for all tested SDMs. Note that if models are not fitted with equal weighting, increasing the number of pseudo–absences decreases the uncertainty at the expense of obtaining lower probability values in the projections (see Chapter 6).

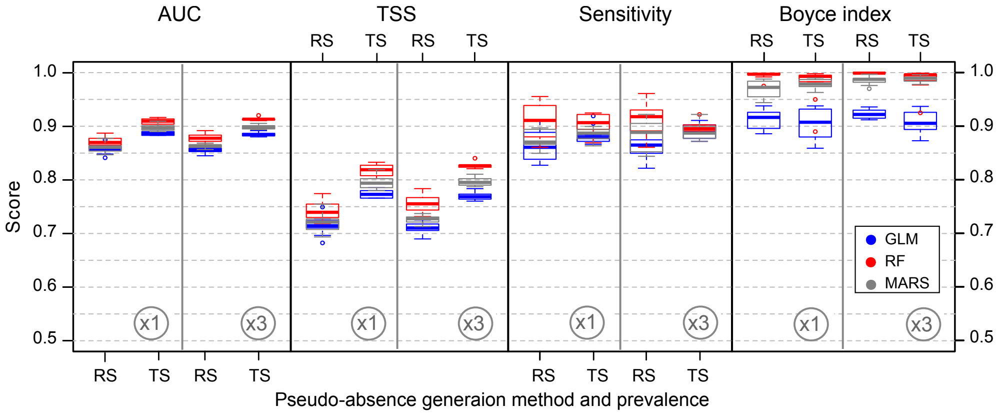

**Figure 5.2:** Model performance scores obtained for each SDM (GLM, RF and MARS) for different prevalence settings: Same number of pseudo–absences as presences (x1) and three times more pseudo–absences than presences (x3). Each chart correspond to a different accuracy measure (AUC, TSS, Sensitivity and the Boyce index) and shows the results for the two different pseudo–absence generation method (RS and TS).

### 5.3.2 *Sensitivity Range*

Figure 5.3 shows maps of the mean suitability and the sensitivity range resulting from the 10 pseudo–absence realizations, for the reference period and future climate projection given by an illustrative regional climate projection, the MPI model (similar results were obtained for the rest of RCMs). These maps show a small sensitivity range for GLM, in both reference and future climates, while the sensitivity is large for RF, but decreasing in the future. On the contrary, MARS exhibits a remarkable increase of uncertainty from reference to future period affecting a large part of the study area, specially Iberia, with range values over 0.5 indicating that predictions switch from absence to presence, or the other way round. Therefore, MARS yielded contradictory predictions regarding the future presence/absence at regional scales, due solely to the pseudo–absence sampling randomness in a certain background. Thus, the uncertainty analysis performed in the historical period cannot be extrapolated into the future.

In order to analyze in detail results obtained in the Iberian Peninsula (IP PRUDENCE region, Fig. 5.1), Figure 5.4 shows the future projected individual

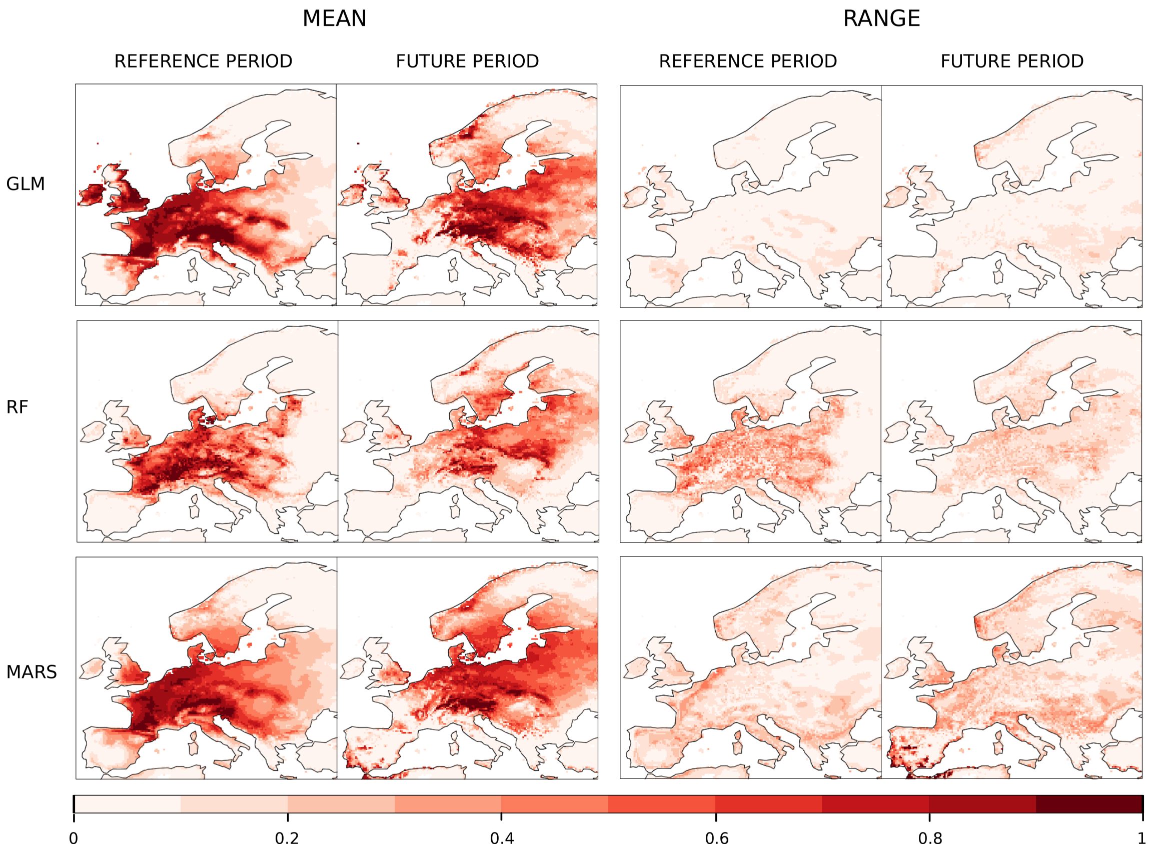

**Figure 5.3:** Mean suitability (MEAN) and Sensitivity range (RANGE) obtained from the set of 10 pseudo–absence realizations, for each SDM (rows) and period (columns). The mean and range of the forth row (MULTISDM) considers the joint of 30 realizations from all SDMs. These maps correspond to method RS and climate projection given by MPI.

suitability for each realization. There are not significant departures from the overall mean in GLM and RF (low sensitivity range), which project a reduced area of potential distribution in the region, according to the habitat shift towards the North-East predicted at European scale (mean maps in Fig. 5.3). Contrarily, the majority prediction of MARS points towards a suitability increment in the southern half of the IP region, with the exception of two realizations (number 2 and 3 in Fig. 5.4), which could be considered more similar to the projections obtained by RF and GLM than to the rest of realizations of MARS. This suggests that the more plausible predictions of MARS among 10 realizations are also the less likely ones. This poses some concerns about the commonplace procedure of combining members and models to construct ensembles, either with an equal probability approach or applying model-weighting according to their performance in reference climate (Buisson *et al.*, 2010; Zhang *et al.*, 2015).

### *5.3.3 Future Projections Uncertainty Due to Pseudo-absences*

Figure 5.5 illustrates the analysis of variance applied to the set of projections that correspond to each SDM and pseudo–absence realization (3 SDMs × 10 realizations) for an example RCM (MPI). The mean suitability map and the standard deviation are shown in the top two panels, while the ones in the bottom are the variance percentage maps showing the contribution of each component to the total variance (realization, SDM and realization & SDM) of the observed deviation. Here we see that the contribution due to the pseudo–absences is considerable —specially in the peripheral areas— since the pseudo–absence realization alone explains up to a 30 % of the variability in wide areas and even a 50 % in some locations (Fig. 5.5). The percentage of the variance is higher for the combination of the two components (realization & SDM) meaning that the contribution of the pseudo–absence realization varies depending on the SDM. Therefore, the variance explained by the SDMs alone is under the 30 % in many areas. This indicates that a significant fraction of the uncertainty attributed to the SDM in different climate change studies may be due to the

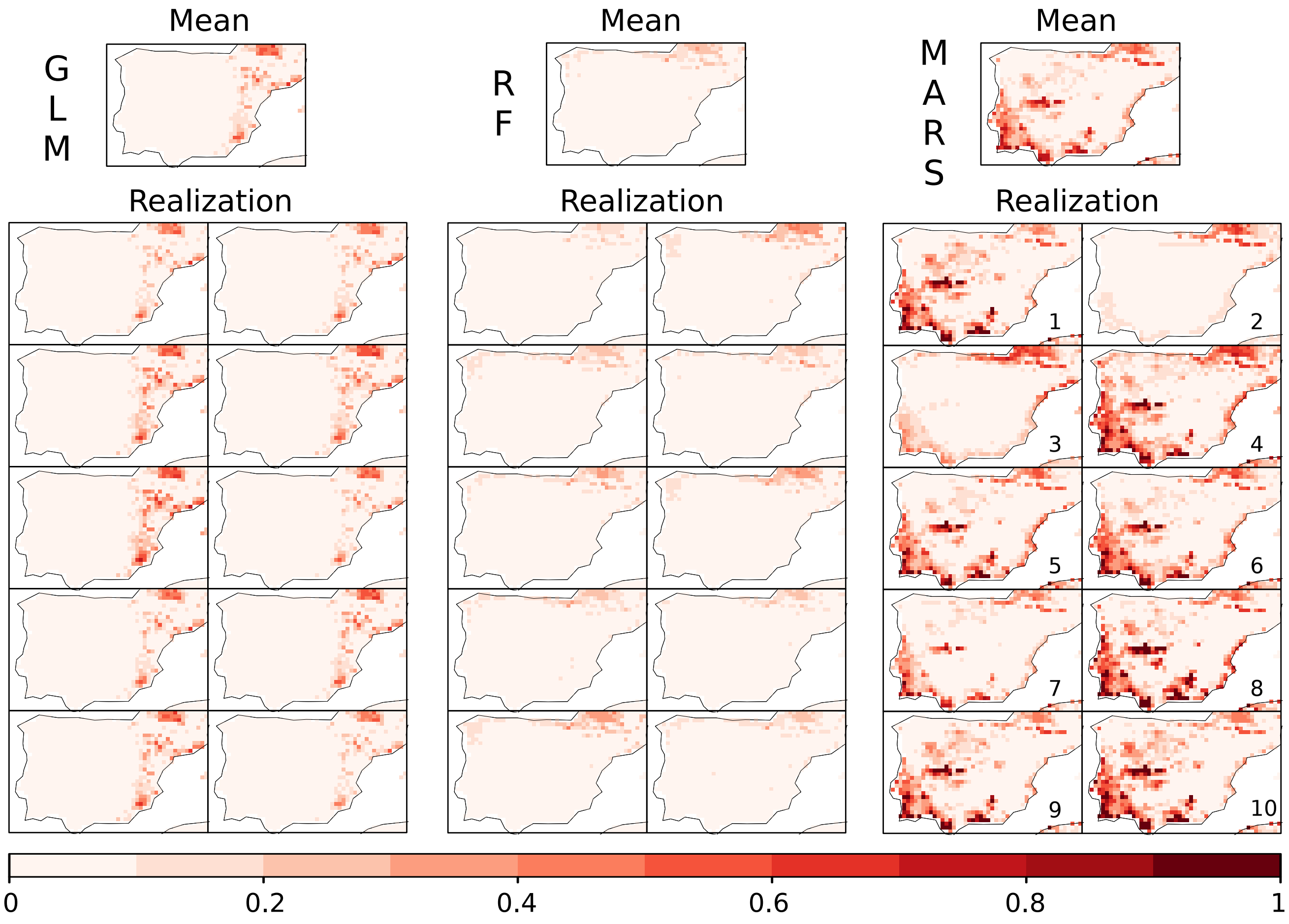

**Figure 5.4:** Future suitability maps of PRUDENCE region IP (Iberian Peninsula, Fig. 5.1) for different pseudo–absence realizations and SDMs. These maps correspond to method RS and future climate projection MPI.

pseudo–absence sample. For instance, a three-member ensemble based on the first realization (see Fig. 5.4) would yield much larger uncertainty than based on the second one. Studies based on a single realization of pseudo–absences, or in the mean of a number of realizations, have the potential to mask results from bad performing SDMs, thus diluting the useful information.

Regarding the variability of the realization with respect to the climate projection (7 RCMs × 10 realizations), Figure 5.6 shows the results obtained for each SDM. The contribution of the RCM clearly differs among SDMs (in connection to what we see in Figure 5.5), being dominant for GLM projections and subordinated to the realization contribution at the peripheral regions for MARS projections (results for RF at this respect are intermediate between GLM and MARS). The areas most influenced by the pseudo–absence realization in GLM projections are those with minimum spread ($s.d. \in [0 - 0.1]$), while this is not a general rule for MARS (e.g. regions IP and MD). Moreover, the contribution of the RCM alone is around the 80 % in wide areas that are not peripheral and have a considerable spread (e.g. region FR). Therefore, to a greater or lesser degree the realization contributes considerably to the MARS projections spread in the major part of the study domain, particularly in the peripheral areas of the current *Quercus* haplotype distribution.

The same overall conclusions hold when applying the TS method for pseudo–absence data generation, even being the spread coming from the realization bigger in some cases. This is depicted in Figure 5.7, that shows the spatial mean of the variance fraction by regions, for both pseudo–absence generation methods and all prevalence settings. This Figure summarizes the information by only showing the contribution of the RCM alone, as the percentage of variance that is explained by the realization is the complementary of the percentage observed therein. Here we can see that the previously described differences among SDMs are maintained across all PRUDENCE regions, prevalence settings and pseudo–absence generation methods, and that even considering the best case scenario, MARS still shows a considerable uncertainty as compared to GLM. In addition,

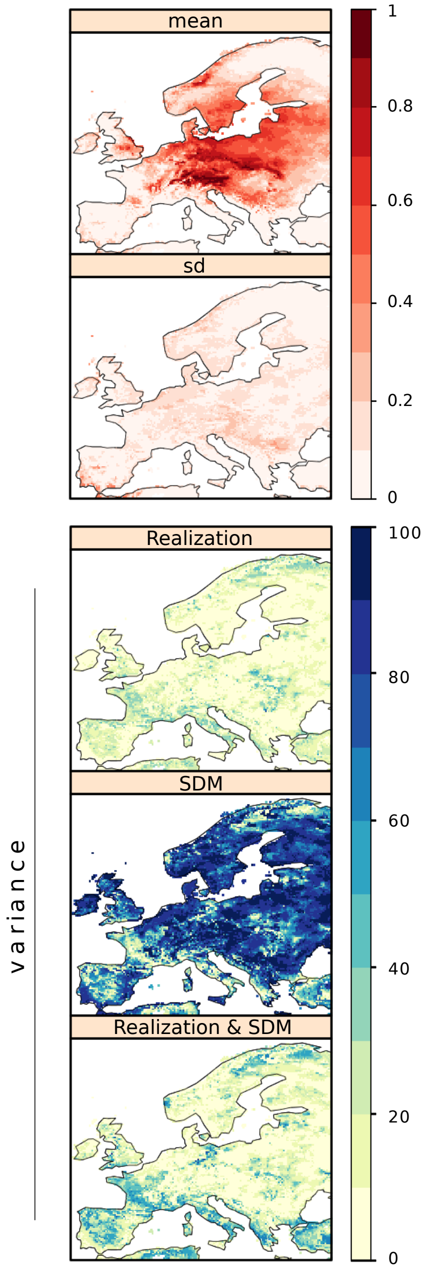

**Figure 5.5:** Mean and standard deviation of the suitability maps corresponding to 3 SDMs x 10 realizations (red maps), and variance percentage explained by each component (realization, SDM and realization & SDM)(yellow-blue maps). These maps correspond to method RS and climate projection given by MPI.

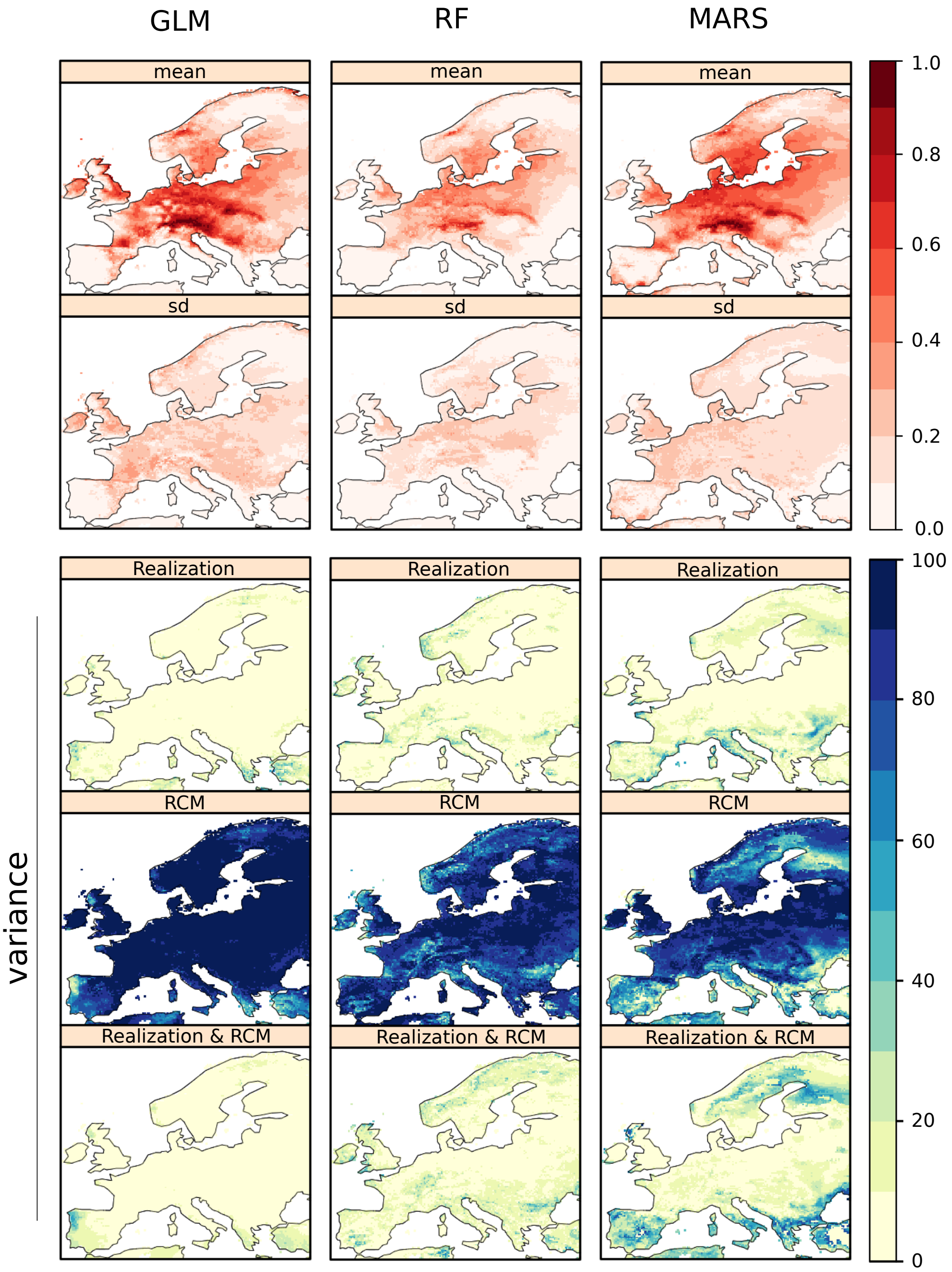

**Figure 5.6:** For each tested SDM (columns), mean and standard deviation of the suitability maps corresponding to 7 RCMs x 10 realizations (red maps), and variance percentage explained by the RCM alone (yellow-blue maps). These maps correspond to method RS.

it is also confirmed that results for RF are in between the other two (except regions BI and ME) and they are less affected by the prevalence setting in most of the cases.

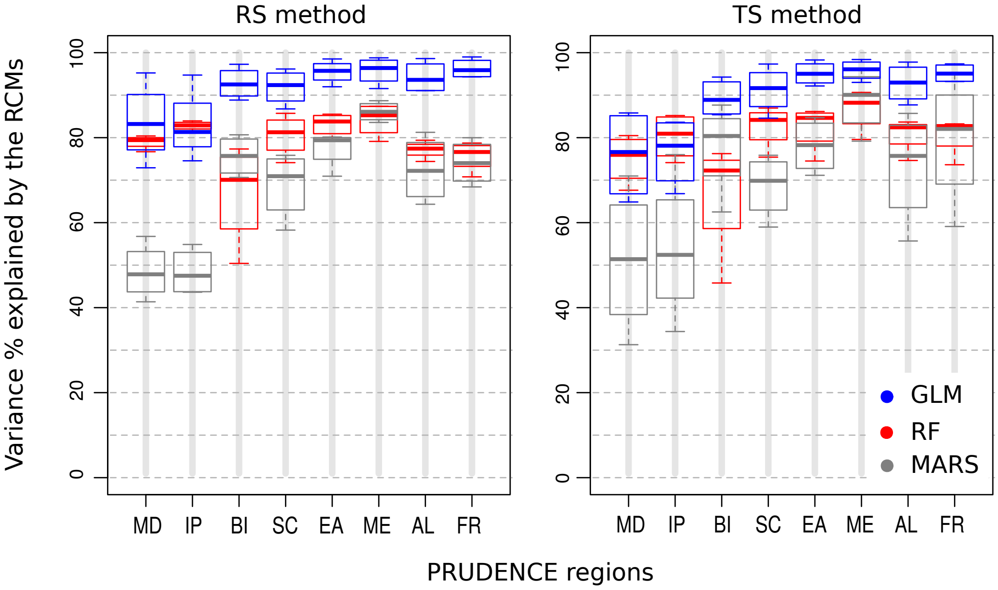

**Figure 5.7:** Box plot of the variance percentage explained by the climate projections (y axis) relative to the pseudo–absence realizations, for each PRUDENCE region (x axis, ordered from peripheral to central), each pseudo absence generation method (RS and TS) and each SDM (GLM, RF and MARS). The spread of the boxes correspond to four different prevalences (same number of pseudo–absences as presences and 2, 3 and 5 times the number of presences).

## 5.4 Discussion

The results obtained in this study reveal a varying sensitivity to the pseudo–absence sample in future projections obtained with different SDMs, being MARS the most sensitive among the tested ones, and GLM the most stable, with the lowest uncertainty derived from different pseudo–absence realizations. Moreover, MARS projections showed unrealistic probability distributions at a regional level (an example has been shown for the Iberian Peninsula), depending on the particular pseudo–absence realization. The contribution of the pseudo–absence realization to the uncertainty was high also in the rest of peripheral areas (excepting the British Isles), specially for MARS, indicating poor transferability (predictive ability) and pointing to an overfitting problem. This is consistent with previous studies in which the stability and reliability of MARS projections have been reported to be dramatically affected by presence sample size (Mateo *et al.*, 2010b).

Note that these results cannot be explained according to the performance of each particular SDM in reference climate conditions, since MARS outperformed GLM, in agreement with previous analysis on multiple-model comparison which indicate that more complex models tend to be more accurate (Elith & et al, 2006). This gives further evidence on the previous finding that model performance gives no indication about the transferability to a non-sampled environment (Fronzek *et al.*, 2011), in this case to future climate conditions. In particular, AUC has been criticized as a measure for evaluating models based on pseudo–absence data, arguing that it can not be meaningfully interpreted and that leads naturally to the selection of complex models (Golicher *et al.*, 2012). The present paper contributes to this discussion and warns about the blind use of ensembles combining models of different complexities, where the members could be differently affected by the particular realization of the pseudo–absence sample.

MARS used around twice the number of parameters used by GLM in most of the cases. In essence, SDMs combine response curves across multiple predictor variables to model the environmental space. A more complex model can fit more complex niche shapes. However, if the model is overly complex (overparametrized), it is likely to make predictions that fit too closely to known occurrences (overfitting) leading to a poor predictive ability for unsampled cells (Peterson *et al.*, 2011). Therefore, in the framework of future niche modeling, we defend that parsimonious models (i.e., with less parameters) are better than complex ones, specially when pseudo–absence data is used (Wisz & Guisan, 2009), given that pseudo–absences are an approximation of real absences and so are occurrences with respect to a non-biased distribution of presences. Thus, if model fitting is also approximated, the inherent bias and false absence rate in the training data is relieved. However, there are still situations where even parsimonious methods yield uncertain results; for example, when a low number of pseudo–absences is used. Therefore, this aspect constitutes a relevant source of uncertainty that should be accounted in SDM applications to climate change studies. In addition, even in the case that non-biased presences and enough reliable absence information were available for modeling, the extrapolation capability of SDMs that are prone to overfitting would be still limited, given that part of the projection environment is out of the sampled range in the calibration phase (Varela *et al.*, 2009; Peterson *et al.*, 2011)

In the same vein, Petitpierre *et al.* (2016) used an independent dataset to evaluate model transferability by measuring the Sensitivity and the Boyce index in the invaded ranges of multiple species, and found that parsimonious models built with less predictors (less parameters) are more transferable to other geographic areas, and that excellent performance in the native range does not necessarily imply good transferability.

A proper validation of SDM future projections is unfeasible by definition. However, here we exposed the sensitivity to the pseudo–absence realization as a model stability and transferability dependent characteristic. In this sense, part of the uncertainty in ensemble forecasts that include non-stable SDMs could be the result of diluting insightful SDM signals with noise from inadequate (e.g. over-parameterized) SDMs (Thuiller *et al.*, 2004; Peterson *et al.*, 2011).

Applying the TS method for pseudo–absence data generation reduces the environmental range available for sampling and, thus, limits the environmental variability among each set of randomly generated pseudo–absences. In this sense, less variability among projections could be expected. On the other hand, sampling pseudo–absences in a narrower environmental range widens the non-sampled range, leading to a low predictive ability in case of overfitting (Wisz & Guisan, 2009), specially for complex SDMs. This explains the higher contribution of the pseudo–absence realization to the uncertainty in the case of the TS method. Nevertheless, the method for pseudo–absence data generation is considered as a study aim dependent choice (Lobo *et al.*, 2010) that conditions model predictions in the gradient between potential and realized distributions of biological entities (Chefaoui & Lobo, 2008).

# CHAPTER 6

# On the Impact of Predictors in Future Climate-Driven Projections

## *6.1 Introduction*

The possibility of building predictive models that are able to extrapolate across space or time are contingent on the choice of appropriate predictors (Peterson, 2011; Rödder *et al.*, 2009), as depicted by previous studies addressing the implications of using different sets of predictors on SDM transferability. For instance, Petitpierre *et al.* (2016) tested how the strategy used to choose predictor variables impacts the extrapolation capability of SDMs, by using an independent set of distribution data in the extrapolation range, allowing the measurement of model accuracy in the native and also the invaded range. Similarly, Peterson & Nakazawa (2008) studied the implications of different environmental datasets in developing general, predictive and extrapolative ecological niche models, suggesting that some environmental datasets may be less useful, in agreement with the results obtained in the illustrative example provided in the introduction of this Thesis (Sections 1.7 and 2.3).

On the other hand, Pliscoff *et al.* (2014) pointed that, despite the generally high predictive performance achieved under all different sets of predictor, they can have statistically significant effects on the spatial patterns of the predictions, that are transferred to the projections of climate change on species distributions and estimates of habitat shifts (Fordham *et al.*, 2011; Braunisch *et al.*, 2013; Wenger *et al.*, 2013). In this sense, Baker *et al.* (2015) considered future projections derived from different baseline climates to account for the uncertainty explained by the baseline climate data in contrast to other well known sources of uncertainty (e.g. future climate projections) and concluded that constitutes an important source of uncertainty in future ensemble forecasts. However, as far as we know, the interrelationship between predictors and pseudo–absences has been neglected in SDM applications to climate change studies.

This chapter extends the work presented in Chapter 5 in order to further analyze the contribution of pseudo–absences to the uncertainty in future SDM projections, with a focus on the interrelationship of the pseudo–absence sample and the set of explanatory variables used to build the models, which constitute two important methodological aspects affecting uncertainty and SDM transferability.

## 6.2 Methods and Materials

The methods and data used are the same as those in the foregoing chapter (see Section 5.2), Additionally, we analyze the dissimilarity among the obtained future suitability maps for different sets of predictors and SDMs, by means of niche distance matrices derived from computing the *niche overlap* between each pair combination of the projected probabilities.

Niche overlap measures the similarity of the environmental ranges occupied by each constructed model via operating the difference between two vectors of probability distributions ($p$), where $p_{x,i}$ and $p_{y,i}$ are the normalized suitability scores for biological entity or model $X$ and $Y$ in grid cell $i$. We considered the Schoener's statistic $D$ for niche overlap (Warren *et al.*, 2008; Broennimann *et al.*, 2012), defined as

$$D(p_x, p_y) = 1 - \frac{1}{2}\sum_{i=1}^{n} |p_{x,i} - p_{y,i}|, \tag{6.1}$$

$D$ ranges from 0 to 1, thus $1 - D$ gives the niche dissimilarity between two models (Fig. 6.4).

### 6.2.1 Strategies for Variable Selection

After a pairwise cross-correlation analysis of the 19 standard bioclimatic variables (see Section 1.5) following Bedia *et al.* (2013), we discarded variables highly cross-correlated ($>$ 0.9, Fig. 6.1). Then, different strategies for variable selection were applied over the resulting set of predictors. These strategies are described in Table 6.1.

The set of predictors corresponding to the stepwise selection procedure is the same as in Chapter 5. The relevant variables for each pseudo–absence realization are indicated in Table 6.2.

## 6.3 Results and Discussion

Given that similar results were obtained for different pseudo–absence generation methods regarding the contribution of pseudo–absences to the uncertainty (Chapter 5), and that the RS method provides more interpretable results, hereafter, we only show the results obtained for the RS method. In addition we also show results for the non-weighted modeling scheme of presences *vs* pseudo–absences, in order to show how does the weighting scheme affect the resulting suitability maps.

### 6.3.1 Model Performance and Niche Dissimilarities

Accuracy measures were similar among weighting schemes, but differed among predictor sets (Fig. 6.2) —more than among prevalence settings—, being $V2pcs$

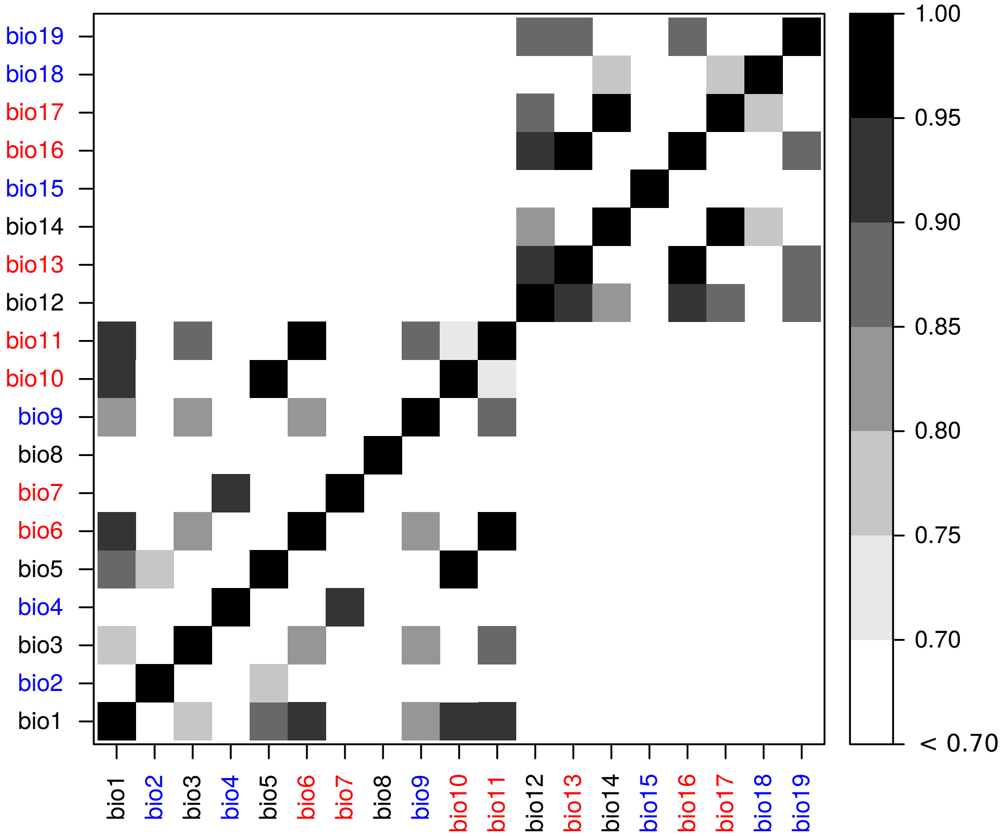

**Figure 6.1:** Pearson correlation of the 19 standard set of bioclimatic variables. The first eleven variables derive from temperature data and the rest from precipitation data. Highly correlated ($> 0.9$) variables are written in red and the subset of uncorrelated variables ($< 0.7$) corresponding to strategy $Vuncor$ are written in blue (see Table 6.1).

**Table 6.1:** Abbreviation and description of each strategy used to select predictors from the standard set of the 19 bioclimatic variables. These strategies are aimed at reducing dimensionality and collinearity of the set of predictors used to build the models, except for $Vall$, which actually is equivalent to not applying any strategy.

| | |
|---|---|
| $Vall$ | The entire set of bioclimatic variables (BIO1 to BIO19). |
| $Vuncor$ | After a pairwise cross-correlation analysis of the resulting bioclimatic variables (following Bedia *et al.*, 2013), we retained a subset of 6 uncorrelated predictors ($< 0.7$), these are BIO2, BIO4, BIO9, BIO15, BIO18 and BIO19. |
| $Vsw$ | After discarding highly correlated variables ($> 0.9$: BIO6, BIO7, BIO10, BIO11, BIO13, BIO16, BIO17) we performed a stepwise glm and retained a subset of variables that are relevant for all the realizations of pseudo–absences. These are BIO1, BIO4, BIO5, BIO9, BIO15, BIO18, BIO19 (see Table 6.2). |
| $V6pcs$ | After discarding highly correlated variables (r $> 0.9$), the first 6 components of a principal component analysis (PCA) —relatively calculated to the climate projections— were retained. |
| $V2pcs$ | Same as $V6pcs$ but only retaining the first two components. |

the one producing the worst performing models in terms of AUC and TSS. On the other hand, strategy $V2pcs$ attained better Sensitivity and Boyce index scores, excepting for GLM, specially regarding the Boyce index, that showed even negative scores (out of graph in Fig. 6.2) when weighted pseudo–absences were modeled, proving not to be an appropriate set of predictors for GLM, the most parsimonious modeling technique among tested SDMs. In order to see how this particular result is reflected by the projected distribution probabilities (suitability maps), Figure 6.3 show the suitability maps for GLM (built with three times more pseudo–absences than presences) and each predictor set. Here we see that $V2pcs$ produced highly suitable areas in regions where the other sets of predictors produced low or even none suitablity (e.g. Scandinavia, Iberian Peninsula). This points to an over-prediction problem of model transferability due to an excessive simplicity (under-parametereization) of the built model (see

**Table 6.2:** Significant variables (indicated with symbol *) resulting from the stepwise analysis performed for each pseudo–absence realization. The first column (r) enumerates each realization. The rest of columns correspond to the bioclimatic variables with correlation < 0.9 (BIO = B).

| **r** | **B1** | **B2** | **B3** | **B4** | **B5** | **B8** | **B9** | **B12** | **B14** | **B15** | **B18** | **B19** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | * |  | * | * | * | * | * | * | * | * | * | * |
| **2** | * |  | * | * | * |  | * |  | * | * | * | * |
| **3** | * |  | * | * | * |  | * |  | * | * | * | * |
| **4** | * |  | * | * | * | * | * |  |  | * | * | * |
| **5** | * | * |  | * | * | * | * |  | * | * | * | * |
| **6** | * |  | * | * | * |  | * |  |  | * | * | * |
| **7** | * |  | * | * | * |  | * | * |  | * | * | * |
| **8** | * |  | * | * | * |  | * | * |  | * | * | * |
| **9** | * | * |  | * | * | * | * |  | * | * | * | * |
| **10** | * |  | * | * | * |  | * | * | * | * | * | * |

Section 2.2.2).

Niche dissimilarities among predicted probabilities corroborate that the suitability maps produced by GLM when using the set of predictors $V2pcs$, clearly differ from the rest of predictions, since present up to a 50% of non-overlapping niche with respect to the rest projections for the non-weighted scheme, and more than a 30% regarding other GLM projections for the weighted scheme. This is depicted in Figure 6.4, which shows the hierarchical clustering of the $1 - D$ metric.

Regarding the rest of strategies for variable selection, these are grouped according to the particular SDM, meaning that predictions differ more among SDMs ($0.25 < 1 - D < 0.5$ in Fig. 6.4) than among predictor sets. Still, there is a considerable percentage (between 10% and 20%) of non-overlapping niches among different sets of predictors for a particular SDM ($0.1 < 1 - D < 0.2$ in Fig. 6.4). Additionally, it should be noted that, when using strategy $Vall$, GLM projected probabilities more similar to those projected by RF under the weighed modeling scheme, while attaining the highest accuracy values as

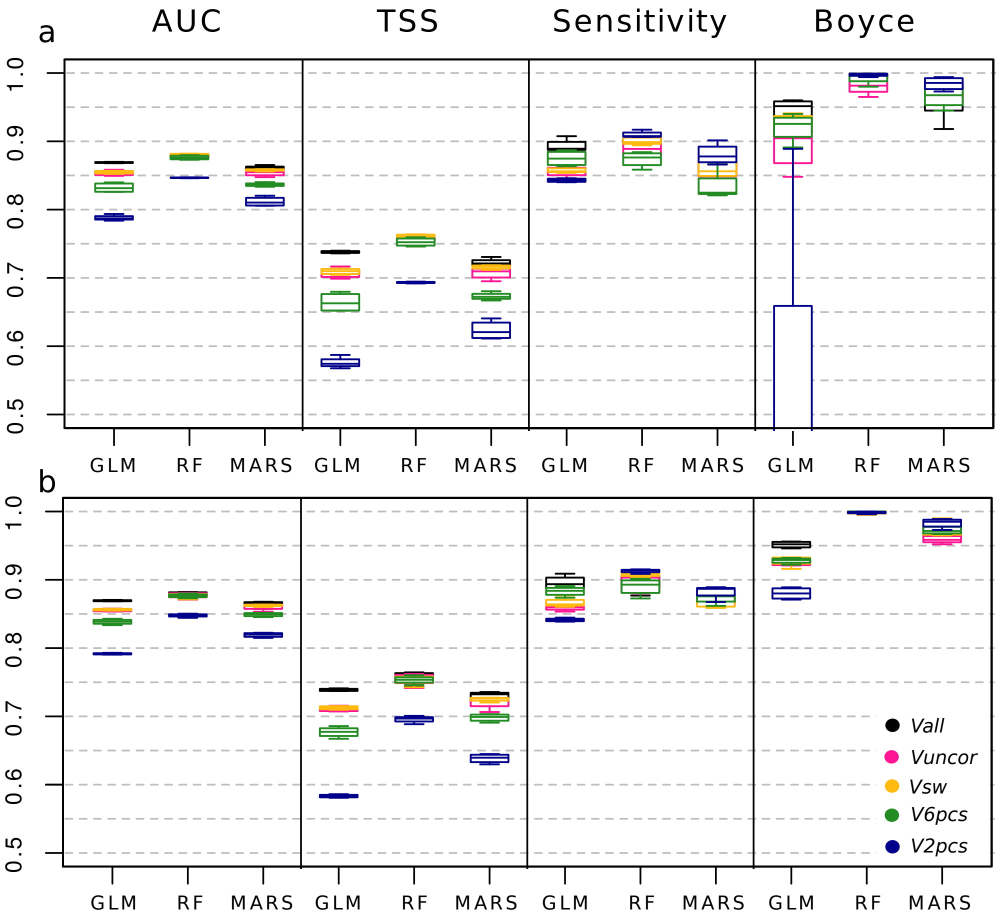

**Figure 6.2:** Model performance scores obtained for each SDM (GLM, RF and MARS) and prevalence setting (same number of pseudo–absences as presences and two, three and five times more pseudo–absences than presences). Each chart correspond to a different accuracy measure (AUC, TSS, Sensitivity and the Boyce index) and shows the results for weighted (a) and unweighted (b) modeling of pseudo–absences.

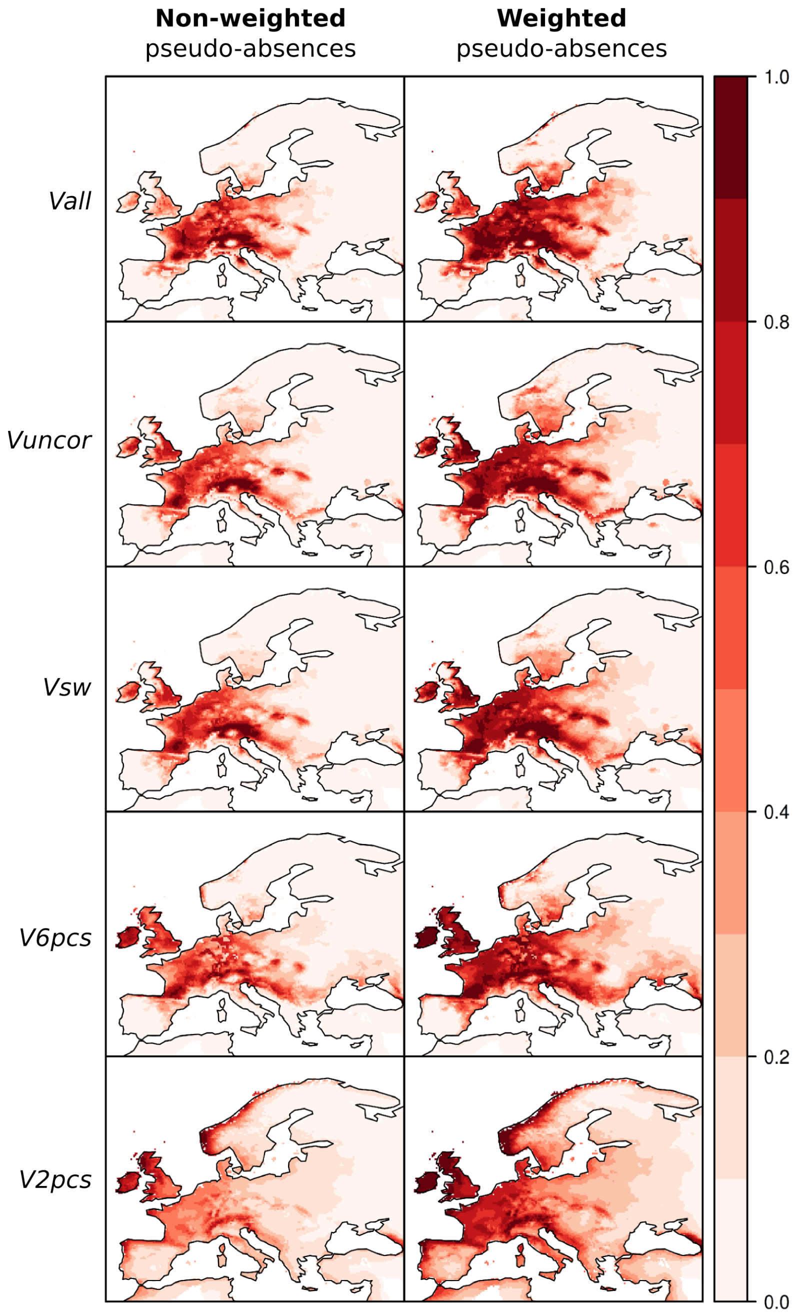

**Figure 6.3:** Future suitability maps projected by GLM when using three times more pseudo–absences than presences, for different sets of predictors (rows) and weighting schemes (columns). These maps show the mean of the 10 pseudo–absence realizations.

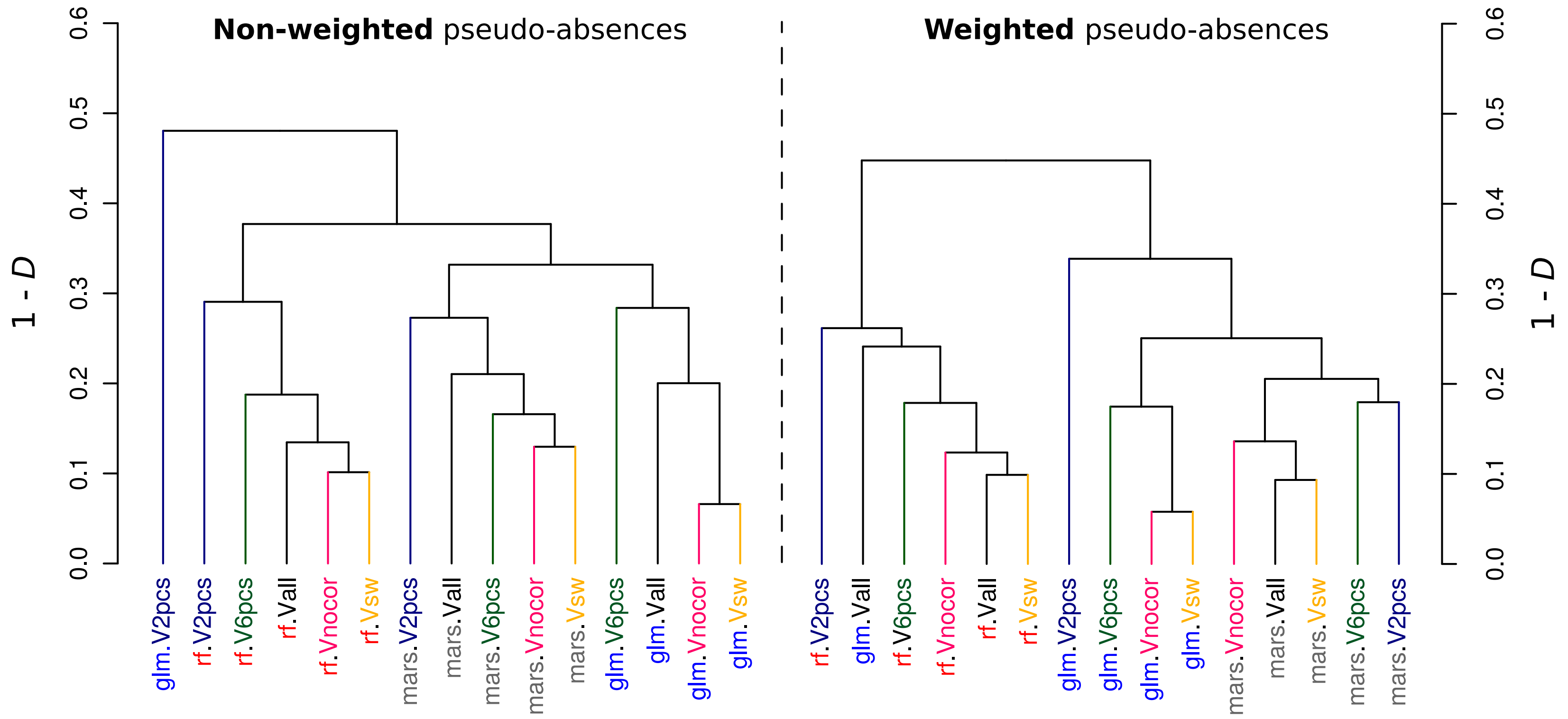

**Figure 6.4:** Dendrograms of the hierarchical cluster analysis of niche dissimilarity $(1 - D)$. Colored lines make reference to the different sets of predictors (black = $Vall$; pink = $Vuncor$; yellow = $Vsw$, green = $V6pcs$; blue = $V2pcs$). SDMs are differenced with colored writing (blue = GLM; red = RF; gray = MARS). Results corresponding to the pseudo–absence realization mean, climate projections given by MPI, and the use of three time more pseudo–absences than presences are shown, for the weighted (right) and unweighted (left) modeling schemes.

compared to the rest of predictor sets (Fig. 6.2) for both weighting schemes. In fact, $Vall$ showed high model performance for all cases, together with the rest of predictor sets, excepting $V2pcs$, whose outputs could be discarded at the model calibration and evaluation phase due to model performance problems.

These results confirm that, despite the general high model performance, different sets of predictors produce different suitability maps, thus constituting an important source of uncertainty in future ensemble forecasts (Fordham *et al.*, 2011; Braunisch *et al.*, 2013; Wenger *et al.*, 2013).

As expected, predictions corresponding to the non-weighted scheme show lower probabilities of suitability (e.g. Fig. 6.3) as the proportion of pseudo–absences *vs* presences increases. On the other hand, model performance is not affected by the weighting scheme (Fig. 6.2). Therefore, modeling should be performed by equal weighting of presences and (pseudo–)absences when possible.

### *6.3.2 Uncertainty of Future Projections Due to Pseudo–absences*

Figure 6.5 is analogous to Figure 5.7 in Chapter 5, and shows the results of the variance analysis (described in Section 5.2) performed for each set of predictors —including the set analyzed in the previous chapter ($Vsw$)—. Here, the pattern of the variance proportion explained by each component (pseudo–absence realization and future climate projection) across all PRUDENCE regions and SDMs is very similar for all sets of predictors. Thus, the results of Chapter 5 are here reinforced, given that the contribution of the pseudo–absence realization to the total variability of the future SDM projections is higher for the most complex SDMs —specially in peripheral regions—, regardless of the strategy considered to build the set of predictors for modeling.

Nevertheless, the variability in future SDM projections due to different pseudo–absence sampling realizations increased for all tested SDMs when no strategy for reducing collinearity and dimensionality was applied ($Vall$) to the initial set of 19 bioclimatic variables (Fig. 6.5). Therefore, the use of inappropriate predictors could potentiate the uncertainty derived from the use of pseudo–absences in future ensemble forecasts of species distributions. Contrarily, using strategy $V6pcs$ reduced the variability in MARS and RF projections as compared to the rest of strategies, even though model performance shown by $V6pcs$ was generally the lowest (excluding $V2pcs$, Fig. 6.2). Therefore, model performance is not determinant for selecting appropriate predictors regarding SDM transferability. On the other hand, they are still valuable, since in this case study, reflected an over-prediction problem of GLM for strategy $V2pcs$ due to underfitting (Thuiller *et al.*, 2004).

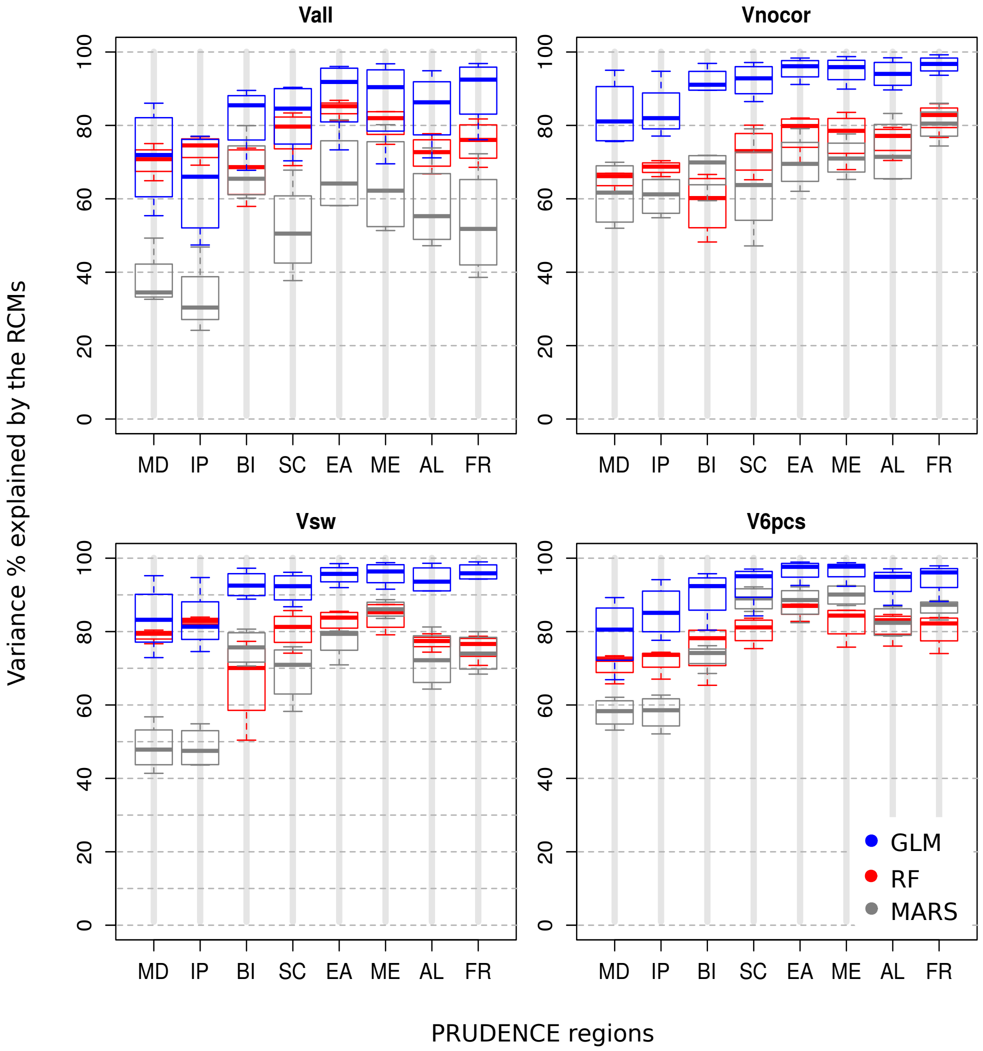

**Figure 6.5:** Box plots of the variance percentage explained by the climate projections (y axis) relative to the pseudo–absence realizations, for each PRUDENCE region (x axis, ordered from peripheral to central) and each SDM (GLM, RF and MARS). The spread of the boxes correspond to four different prevalences (same number of pseudo–absences as presences and 2, 3 and 5 times the number of presences). Each chart corresponds to a different set of predictors ($Vall$, $Vuncor$, $Vsw$ and $V6pcs$). The box plot corresponding to $Vsw$ is the same as the one shown in Chapter 5 (Fig. 5.7).

# Part III

# Developed Tools

# CHAPTER 7

# The R Package MOPA for Species Distribution MOdelling with Pseudo–Absences

## 7.1 Introduction

SDMs have become a valuable tool as a means of estimating distribution shifts due to climate variations, a problem of current interest in environmental conservation studies (see e.g.: Araújo *et al.*, 2004; Hamann & Wang, 2006; Jeschke & Strayer, 2008). As a result, there is an increasing demand of climate products, requiring historical climate databases (see Section 1.5) and future climate projections (see Section 2.1). Despite the increased use of future SDM projections as a support tool for decision-making in biological conservation, the communication of the inherent uncertainties of these products remains as an ongoing challenge (see, e.g. Araújo *et al.*, 2005; Beaumont *et al.*, 2008; Fronzek *et al.*, 2011). There are important sources of uncertainty that are rarely quantified, yet crucial, in order to assess the credibility of the future distributions, such as the varying errors of the different modeling algorithms used to characterize the ecological niche (Bedia *et al.*, 2011), the SDM extrapolation ability outside the training period/spatial extent (transferability in time/space; e.g. Fronzek *et al.*, 2011), uncertainties regarding the training data (Mateo *et al.*, 2010b; Bedia *et al.*, 2013), the assumptions underlying the different emission scenarios (Nakićenović, 2000), the global/regional climate model (GCM/RCM) biases (Turco *et al.*, 2013) and others (see e.g.: Falloon *et al.*, 2014, for an overview). Among them, two have been highlighted in this Thesis, namely the SDM choice (see e.g. Buisson *et al.*, 2010; Fronzek *et al.*, 2011; Garcia *et al.*, 2012, who find that different statistical methods can differ wildly in their projected distributions, being not all of them equally plausible), and the approach used for pseudo–absence data generation.

In this context, the R package **mopa** (MOdeling with Pseudo-Absences) has been built and developed as part of the work of this Thesis, providing tools —based in the open-source R language (R Core Team, 2015)—, for pseudo–absence data generation and species distribution modeling, with a focus on the above aspects related to SDM transferability and uncertainty. All methods and techniques described in previous chapters are implemented in **mopa**, which packs specific functions that allow to flexibly explore and combine different ensemble configurations of the projected probability distributions and perform a variance partitioning approach that allows to quantitatively assess the contribution of different factors to the overall spread of the SDM projections.

In this Chapter, we show the package functionality through a case-study that reproduces part of the analysis performed in Chapter 5, but considering a distinct *Quercus robur* phylogeny (H11), that in turn has a differentiated geographical distribution pattern (see Fig. 4.1). We use publicly available data that are included in the package to guarantee the reproducibility of the R code shown in the following sections.

### *7.1.1* MOPA *Within the "SDM ecosystem" in R*

**mopa** is oriented towards the analysis of components that add variability to the projected distributions in non-sampled environmental spaces (e.g. under climate change conditions or new geographical areas), thus directly addressing the problem of SDM transferability, that can not be properly evaluated during model calibration. Besides, unlike previously existing packages (see Section 2.2.4), **mopa** allows pseudo–absence data generation as an independent step prior to model fitting, thus providing a finer control to the user for the analysis of several alternative methods and specific tuning options. In addition, the novel Three-Step method for pseudo–absence data generation is implemented (TS hereafter, Senay *et al.*, 2013), that has been shown to improve model transferability (Iturbide *et al.*, 2015), providing a convenient interface that allows a fine tuning of the technique with simple arguments. Furthermore, **mopa** is also seamlessly integrated with standard R packages for spatial data manipulation like **raster** (Hijmans, 2015) and **sp** (Pebesma & Bivand, 2005), allowing their usage at any stage of the modeling process (e.g. for data visualization and post-processing), and also a direct extensibility to other SDM tools available in **sdm**, **biomod2**, etc., also handling the same spatial data classes.

### *7.1.2 Integration of* MOPA *with Climate Services*

An important barrier for SDM development is climate data retrieval and preparation. With this regard, the **climate4R** bundle has been recently developed, a set of R packages specifically designed to ease climate data access, analysis and processing in a straightforward manner, tailored to the needs of the impacts and vulnerability assessment community. Further details and references

The main content of these Chapter was submitted to *R Journal* in May 2017 and was under review when the Thesis was printed.

to worked examples and tutorials can be found for instance in Cofino *et al.* (2017) and Bedia *et al.* (2017). **mopa** is fully integrated within the **climate4R** bundle, directly handling the climate data structures of the **climate4R** and providing conversion features to other types of R data classes, in order to obtain appropriate climate variables for modeling.

### 7.1.3 Package Installation

**mopa** is available trough a public GitHub Repository (`https://github.com`). The recommended installation for most users is a direct install from the master branch with the latest stable release. To this aim, the function `install_github` from the **devtools** R package (Wickham & Chang, 2016) is recommended.

```
> devtools::install_github("SantanderMetGroup/mopa")
```

## 7.2 Input Data Pre-processing

### 7.2.1 Climate Data

Predictor variables (in this case-study a number of bioclimatic variables, but not necessarily so) are introduced in the analysis as collections of **raster** objects of the classes `rasterBrick` or `rasterStack`, similarly as other SDM-oriented packages. For instance, function `biovars` from package **dismo** uses precipitation and temperature climatologies in the form of `rasterBrick/Stack` to calculate a standard set of bioclimatic variables widely used in SDM applications (Busby, 1991). For instance, the built-in **mopa** dataset `biostack` contains a set of bioclimatic variables (present and future) constructed with the function `biovars` of package **dismo**. The precipitation and temperature climatologies have been calculated from the E-OBS gridded observational dataset (Haylock *et al.*, 2008), and from 7 Regional Climate Model (RCM) simulations of the project ENSEMBLES (van der Linden & Mitchell, 2009, `http://www.ensembles-eu.org`). Further details about the data sources are included in the help of the dataset:

```
> library(mopa)
> data(biostack)
> help(biostack)
```

### 7.2.2 *Species Distribution Data*

Several impact studies indicate that species should be modeled by treating sub-specific groups of organisms independently (e.g. distinct genetic linages, see Section 1.3) due to their differing adaptive responses to changes in their environment (Hernández *et al.*, 2006; Beierkuhnlein *et al.*, 2011; Serra-Varela *et al.*, 2015). Although this is not always possible, due to the rare availability of information on the distribution of sub-specific groups for most of species, **mopa** has been conceived with this idea in mind, being able to deal with several sets of presences simultaneously. This adds flexibility to the modeling process in order to carry out experiments considering different sub-collections of presences, not only for sub-specific analyses (Iturbide *et al.*, 2015), but also to address the sensitivity of the modeled distributions to different characteristics of the training sample (e.g. the sample size, Hernández *et al.*, 2006; Mateo *et al.*, 2010b). Thus, the `Oak_phylo2` **mopa** dataset contains a named list of length two, containing the geographical coordinates of presence localities for two different Oak phylogenies (H01 and H11, Petit *et al.*, 2002b). More details about the source data are provided in the help file of the dataset.

```
> data(Oak_phylo2)
> help(Oak_phylo2)
> presences <- Oak_phylo2$H11
```

### 7.2.3 *Geographic Background*

The geographic background is often defined as the spatial extent of the area considered in the SDM calibration stage. Here, we refer to the *background* as a regular, geo-referenced grid with a specific size and resolution, in which both the environmental variables and the presence localities are located, so its grid-points are the sampling units. Function `backgroundGrid` provides a simple way of generating a backgroud using a raster-class object as reference. It also includes an additional argument (spatial.subset) for spatial subsetting, set by a raster::extent object or by one or several sets of bounding-box coordinates, providing great flexibility and ease of use for the analysis of SDM spatial aspects. For instance, it allows straightforward exploration of SDM geographical transferability or performing cross-validation experiments based on spatial folds (e.g.: Randin *et al.*, 2006). As a result, when the object Oak_phylo2 is passed to backgroundGrid, two different backgrounds are created by default, each one spatially restricted by its phylogeny distribution (H11 and H01).

```
> bg <- backgroundGrid(raster = biostack$baseline$bio1)
```

A smaller domain than the previous one can be arbitrarily indicated by the user by providing a specific spatial extent:

```
> bg.subdomain <- backgroundGrid(
        raster = biostack$baseline$bio1,
        spatial.subset = extent(c(-10, 35, 45, 65)))
```

Similarly, the user might be interested in a background strictly constrained by the bounding box of the actual species localities, by just passing to spatial.subset their coordinates:

```
> bg.species <- backgroundGrid(
        raster = biostack$baseline$bio1,
        spatial.subset = presences)
```

Thus, the user has flexibility to perform further modifications of the background, so it would be also possible to discard specific areas based on expert knowledge (e.g. Serra-Varela *et al.*, 2015). In this case study, we will retain the full background (bg) for further analyses.

## 7.3 Pseudo-absence Generation

Pseudo-absence sampling in **mopa** is performed by the pseudoAbsences function. It implements a wide range of methodologies described in the literature (see Iturbide *et al.*, 2015, for an overview and comparison of methods) for maximum user flexibility, but at the same time its arguments have been kept as simple as possible to ease its application (Table 7.1). Here, three methods are described: random sampling, random sampling with environmental profiling and the three-step method. Their main characteristics are next briefly described. A more extended explanation can be found in (Iturbide *et al.*, 2015) and reference therein.

**Table 7.1:** Arguments of function `pseudoAbsences` controlling the parameter values involved in pseudo–absence generation.

| **Argument** | **Description** |
|---|---|
| `realizations` | Number of realizations of pseudo–absence generation |
| `exclusion.buffer` | Minimum distance to be kept between presence data and pseudo–absence data |
| `prevalence` | Proportion of presences against absences |
| `kmeans` | Performs a k-means clustering of the background to extract the pseudo–absences instead of sampling at random |
| `varstack` | `RasterStack` of variables for computing the k-means clustering |

*Random Sampling (RS).* The RS method is the simplest and most frequent way of generating pseudo–absences. In the next example three times more pseudo–absences than presences are generated at random, keeping a 0.249° ($\simeq$ 30 km) exclusion buffer around known presence localities. Ten pseudo–absence realizations are considered:

```
> pa_RS <- pseudoAbsences(xy = presences,
        background = bg$xy,
        realizations = 10, exclusion.buffer = 0.249,
        prevalence = -0.5)
```

As an alternative to strict RS, a stratified random sampling approach can be performed, based on homogeneous environmental conditions. To this aim, a clustering of the environmental space is often applied (Senay *et al.*, 2013):

```
> pa_RS_kmeans <- pseudoAbsences(xy = presences,
        background = bg$xy,
        exclusion.buffer = 0.249,
        prevalence = -0.5,
        kmeans = TRUE, varstack = biostack$baseline)
```

*Random Sampling with Environmental Profiling (RSEP).* The RSEP method imposes restrictions on the environmental range of the background to be sampled for pseudo–absences. In **mopa** this is done by performing an environmental profiling of the background (function `OCSVMprofiling`) that, following Senay *et al.* (2013), applies a one-class support vector machine algorithm (OCSVM, implemented in package **e1071**, Meyer *et al.*, 2017) returning a binary (presence/absence) classification of the background gridboxes based solely on the presence information (`bg.profiled$presence` and `bg.profiled$absence` in the example below). Only the predicted absence background is then retained for pseudo–absence generation.

```
> bg.profiled <- OCSVMprofiling(xy = presences,
        varstack = biostack$baseline,
        background = bg$xy)
> pa_RSEP <- pseudoAbsences(xy = presences,
        background = bg.profiled$absence,
        realizations = 10, exclusion.buffer = 0.249,
        prevalence = -0.5)
```

*Three-step method (TS).* TS is based on imposing restrictions to both the environmental range and the spatial extent of the background from which pseudo–absences are sampled. This method has been shown to outperform other common approaches in terms of resulting SDM robustness (Iturbide *et al.*, 2015). The TS method adds an additional step to the RSEP method, consisting on the partition of the background space (as yielded by RSEP) in multiple bands using different radius from presence localities. In the example below, multiple distance bands with an increasing radius of 30 km between each other are created (argument `by = 0.249`, in degrees). The first one (with the shortest radius from presence localities) is at 30 km from the closest presence point (`start = 0.249`), and the largest one (the longest radius from presences) is set by default to half the length of the diagonal of the background bounding-box (see Iturbide *et al.*, 2015, for more details).

```
>  bg.radius <- backgroundRadius(xy = presences,
        background = bg.profiled$absence,
        start = 0.249, by = 0.249, unit = "decimal degrees")
>  pa_TS <- pseudoAbsences(xy = presences,
        background = bg.radius,
        realizations = 10, exclusion.buffer = 0.249,
        prevalence = -0.5)
```

A spatial representation of the results yielded by the pseudo–absence methods described is next generated (Fig. 7.1):

```
> # Generates Fig. 8.1
> par(mfrow = c(2, 2), mar = c(2, 2, 2, 1.2))
> # Panel 1a (Presence data)
> plot(bg$xy, pch = 18, cex = 0.4, col = "gray", asp = 1)
        > points(presences, pch = 18, cex = 0.6, col = "red")
> # Panel 1b (RS method)
> plot(bg$xy, pch = 18, cex = 0.4, col = "gray", asp = 1)
        > points(pa_RS$species1$PA01[[1]], pch = 18,
                col = "darkviolet", cex = .6)
        > points(pa_RS_kmeans$species1$PA01[[1]], pch = 18,
                col = "yellow", cex = .6)
        > points(presences, pch = 18, cex = 0.6, col = "red")
> # Panel 1c (RSEP method)
> plot(bg.profiled$absence, pch = 18, cex = 0.4,
        col = "gray", asp = 1)
        > points(bg.profiled$presence, pch = 18, cex = 0.4,
                col = "aquamarine")
        > points(pa_RSEP$species1$PA01[[1]], pch = 18,
                cex = 0.6, col = "darkviolet")
        > points(presences, pch = 18, cex = 0.6, col = "red")
> # Panel 1d (TS method)
> plot(bg.radius[[1]]$km3120, col = "gray", asp = 1,
        pch = 18, cex = 0.4)
        > points(bg.profiled$presence, pch = 18, cex = 0.4,
                col = "aquamarine")
        > for (i in 1:10) {
        l <- (11 - i) * 10
        points(bg.radius[[1]][[l]],
        col = gray.colors(10, start = .9,end = 0.1)[i],
        pch = 18, cex = 0.4)
        }
> points(pa_TS$species1$PA01[[50]], pch = 18, cex = 0.6,
        col = "darkviolet")
> points(presences, pch = 18, cex = 0.6, col = "red")
```

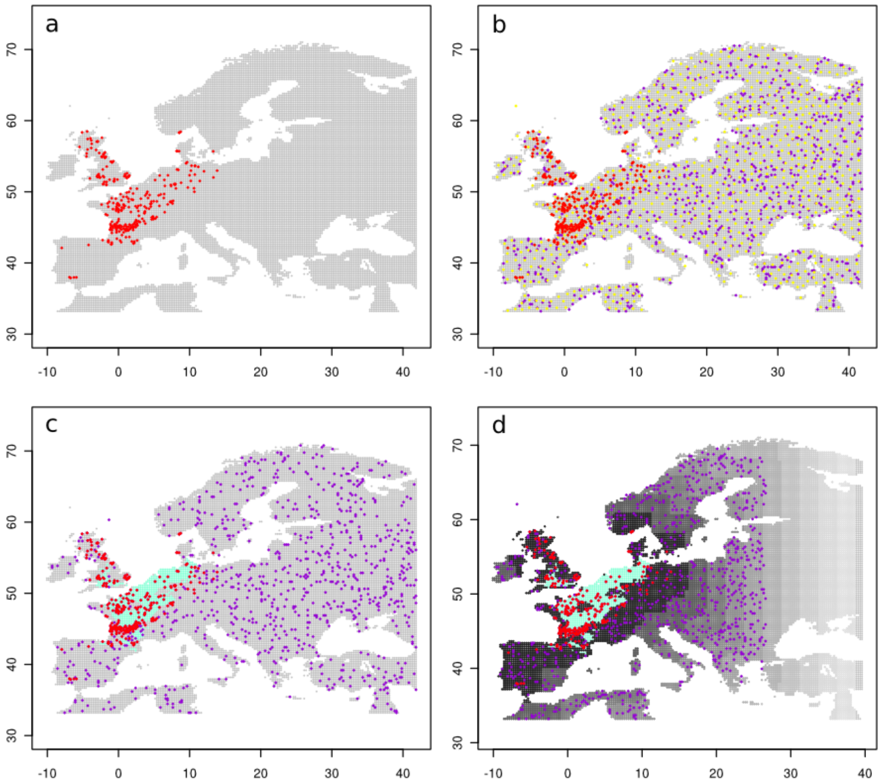

**Figure 7.1:** Pseudo-absence dataset maps, as generated by function `pseudoAbsences`. **(a)** Known presence locations of the Oak phylogeny H11 (red points) and initial background for pseudo–absence sampling (grey grid points). **(b)** pseudo–absences generated using the RS method randomly (purple points) and with k-means clustering (yellow points). **(c)** Pseudo-absences generated with the RSEP method (purple), where the turquoise area corresponds to the discarded suitable background space as identified by the OCSVM profiling approach. **(d)** TS approach. Environmentally stratified as RSEP (c), but also spatially stratified background, the different strata (spatial extents) identified by the different gray-scale colors. Pseudo-absences for one of the background extents (3120 km) are depicted as example (purple points).

Thus, **mopa** allows for the generation of a wide range of combinations of environmental restriction criteria (using `OCSVMprofiling`) and spatial extent constraints (using `backgroundRadius`, see Table 7.2), providing unrivalled functionality for the development and inter-comparison of multiple pseudo–absence setups for SDM refinement and ensemble prediction generation.

**Table 7.2:** Combinations of functions `OCSVMprofiling` and `backgroundRadius` for background definition. These are used prior to pseudo–absence data generation with function `pseudoAbsences`, that controls the different sampling methods.

| `OCSVMprofiling` | `backgroundRadius` | **Method** |
|---|---|---|
| × | × | No restriction (RS method) |
| ✓ | × | Environmental restriction (RSEP method) |
| ✓ | ✓ | Environmental and spatial restriction (TS method) |
| × | ✓ | Spatial restriction (Particular case of RS) |

## 7.4 SDM Fitting and Prediction

### 7.4.1 Model Fitting

Once the pseudo–absence dataset(s) chosen by the user is(are) built, the `mopaTrain` function performs SDM fitting. The function is a wrapper for different statistical method implementations commonly used in SDM applications (see summary in Table 7.3). Moreover, `mopaTrain` adds extended functionality for cross-validation for each set of presence/absence data and for each different species contained in the presence dataset, as routinely done in SDM applications (see e.g.: Verbyla & Litvaitis, 1989). In the next line of code, the Oak H1 phylogeny is fitted using a generalized linear model (GLM, Guisan *et al.*, 2002) and multivariate adaptive regression splines (MARS, Friedman, 1991), applying a 10-fold cross validation approach. Moreover, equal weighting of presences and pseudo–absences is indicated with the argument `weighting = TRUE` (see e.g.: Barbet-Massin *et al.*, 2012).

```
> trainRS <- mopaTrain(y = pa_RS, x = biostack$baseline,
         weighting = TRUE,
         k = 10, algorithm = c("glm", "mars"))
```

**Table 7.3:** SDM techniques available in **mopa** through the function `mopaTrain`. The corresponding `algorithm` argument values are also indicated.

| **SDM technique** | `algorithm` value | **pkg::function** | **Reference** |
|---|---|---|---|
| Generalized Linear Model | "glm" | **stats**::glm | Part of R |
| Random Forest | "rf" | **ranger**::ranger | Wright (2016) |
| Multivariate Adaptive Regression Splines | "mars" | **earth**::earth | Milborrow (2015) |
| Maximum Entropy | "maxent" | **dismo**::maxent | Hijmans *et al.* (2017) |
| Support Vector Machine | "svm" | **e1071**::best.svm | Meyer *et al.* (2017) |
| Classification and regression tree (tree) | "cart.tree" | **tree**::tree | Ripley (2016) |
| Classification and regression tree (rpart) | "cart.rpart" | **rpart**::rpart | Therneau *et al.* (2017) |

### 7.4.2 The Special Case of Model Fitting with TS Pseudo–absences

After the generation of TS pseudo–absences, multiple background extents exist as a result of the different distances defined by `backgroundRadius`. It has been noted that the background extent from which pseudo–absences are sampled is an important factor affecting not only model performance, but also biological meaning (Van der Wal & Shoo, 2009). With this regard, in Chapter 4 we propose a selection criterion based on the response of model performance as a function of distance radius, that is generalizable to different SDM characteristics and spatial scales. The performance criterion chosen is the Area Under the ROC Curve (AUC), one of the most widely used accuracy measures of binary classification systems (Swets, 1988). Essentially, the method performs a non-linear regression of the AUC obtained by each SDM extent against their background radius, considering three possible asymptotic models implemented in **mopa** (also described in Chapter 4):

1. Michaelis-Menten model: $v(x) = \frac{ax}{Km+x}$

2. 2-parameter exponential model: $v(x) = a(1 - e^{-bx})$

3. 3-parameter exponential model: $v(x) = a - be^{-cx}$

, where $v$ and $x$ represent the AUC and the background extent respectively. $a$ is the asymptotic AUC value achieved by the system and $a - b$ is the intercept. $Km$ is the *Michaelis constant* (i.e. the extent at which the AUC is half of $a$, and $c$ is the coefficient of the point where the curve is most pronounced. The asymptotic model that better fits the AUC response to the different background extents is automatically selected to extract the AUC asymptotical value. The minimum extent at which the AUC lies above the asymptote is retained as the optimal threshold radius, being the corresponding fitted SDM returned. The asymptotic models are fitted internally by `mopaTrain` via the `nls` function from package **stats** always the TS method is used (this is automatically detected by the function). Optionally, a diagram displaying the results is also returned by setting the argument `diagrams=TRUE` (Fig. 7.2).

```
> # Train TS model and generate Fig. 8.2
> trainTS <- mopaTrain(y = pa_TS, x = biostack$baseline,
        weighting = TRUE,
        k = 10, algorithm = c("glm", "mars"),
        diagrams = TRUE)
```

### 7.4.3 Model Assessment

The object returned by `mopaTrain` is a list of several components generated in the model calibration and evaluation process. Several performance measures are included apart from the AUC, like the True Skill Statistic (TSS) and Cohen's Kappa obtained in the cross-validation, frequently used for the assessment in SDMs (Allouche *et al.*, 2006). These and other ocmponents of the SDM fitted object can be accessed using `extractFromModel`. For, instance, to extract the TSS:

```
> tss.RS <- extractFromModel(models = trainRS,
        value = "tss")
```

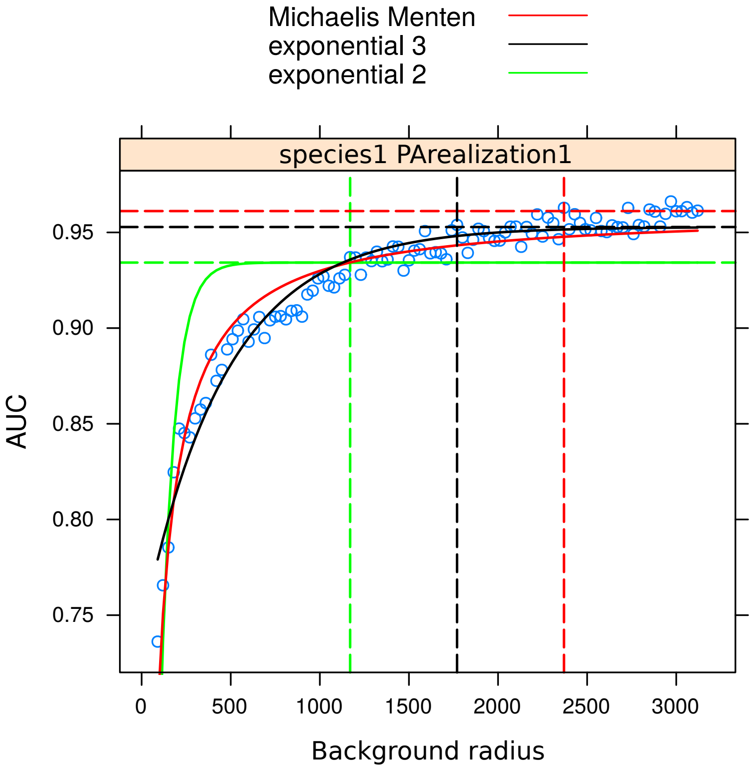

**Figure 7.2:** Asymptotic model fitting in SDMs using the TS approach for pseudo–absence generation. The blue points are the AUC values (y-axis) obtained by the SDMs for different background radius extents (x axis). Non-linear fits to the three asymptotic models considered (Michaelis Menten, 2 and 3-parameter exponential). The vertical and horizontal lines indicate the optimal radius and resulting AUC value of the final `mopaTrain` SDM output.

However, and for maximum user flexibility, a matrix containing the observed and predicted probability values for each calibration point is returned, allowing other types of user-tailored model performance assessments.

```
> ObsPred.RS <- extractFromModel(models = trainRS,
        value = "ObsPred")
```

The fitted models are stored in the `"model"` (or `"fold.models"`) component, required for subsequent model prediction.

```
> models.RS <- extractFromModel(models = trainRS,
        value = "model")
```

Additionally, variable importance may be also estimated. One straightforward possibility is to pass the fitted models (e.g. `models.RS`) to function `varImp` from package **caret** (Kuhn, 2011).

### *7.4.4 Model Predictions*

SDM predictions are obtained by passing a new set of predictors (e.g.: future bioclimatic variables) to the generated models. The `model` component corresponds to the models fitted using all available data for model training, while the SDM predictions for the k-cross-validation setup are generated from the component `fold.models` –instead of `model`–. Thus, **mopa** allows handling both the cross-fitted models for flexible model performance assessment and the global model –fitted with all presences and pseudo–absences– for predicting distributions, accomplished through the use of the function `mopaPredict`. In the following example, models corresponding to the RS method are projected to reference climate conditions (`biostack$baseline`) and to 7 future climate projections (`biostack$future`):

```
> ensemble.present <- mopaPredict(models = models.RS,
        newClim = biostack$baseline)
> ensemble.future <- mopaPredict(models = models.RS,
        newClim = biostack$future)
```

## 7.5 Exploring the Uncertainty in SDM Projections

Projections returned by `mopaPredict` are structured in a nested `list`. Each depth or level in the `list` corresponds to a different component. These are: presence data sets (`SP`), pseudo–absence realizations (`PA`), modeling algorithms (`SDM`), baseline climate (baseClim), and the new climate (`newClim`) used to project models (e.g. future climate projections). The function used to extract components is `extractFromPrediction`. In the next example, projections corresponding to the first pseudo–absence realization (object `rcms_run1`) and to the future climate projection from the MPI RCM (object `runs_rcm1`) are extracted:

```
> rcms_run1 <- extractFromPrediction(ensemble.future, "PA01")
> runs_rcm1 <- extractFromPrediction(ensemble.future, "MPI")
```

Then, the function is again applied to object `runs_rcm1` to extract the SDM results for MPI and GLM. The resulting object is of S4-class `raster*`, thus being straightforward to apply any of the plotting/analysis methods for spatial objects. Here, we use `spplot` from **sp** for output visualization (Fig. 7.3).

```
> glm_runs_rcm1 <- extractFromPrediction(runs_rcm1, "glm")
> # Generates Fig. 8.3
> data(wrld)
> spplot(glm_runs_rcm1, layout = c(5, 2),
         at = seq(0, 1, 0.1),
         col.regions = colorRampPalette(c("white",  "red3")),
         sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

Thus, it is easy to explore the results by inspecting the different components of the `mopaPredict` outputs. For instance, the **raster** package can be particularly useful this aim allowing for a wide variety of map algebra operations through the function `stackApply` over user-defined subsets of SDM projections.

### 7.5.1 Partition of the Uncertainty into Components Using Variance Analysis

The relative contribution of each component to the total ensemble spread (i.e. variability) is implemented in **mopa** using a simple variance approach described

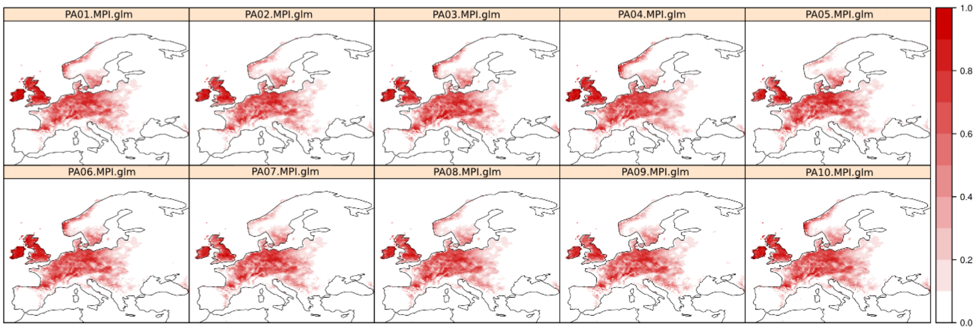

**Figure 7.3:** Future species distribution projections (2071-2100) according to the MPI RCM projections, considering 10 different pseudo–absence realizations of the RS method, as stored in the object `glm_runs_rcm1`.

in Chapter 5, through the function `varianceAnalysis`, following the method in Déqué *et al.* (2012) and San-Martín *et al.* (2016). For instance, in this example, the total variance $V$ can be decomposed as the summation of the variance explained by the pseudo–absence realization $P$ (`component1 = "PA"`), the RCM $R$ (`component2 = "newClim"`) and the combination of both $PR$, so $V = P + R + PR$ (see Section 5.2).

The following example shows the analysis performed for the pseudo–absence realizations and the climate projections in GLM projections (`fixed = "glm"`). In order to illustrate thoroughgoing information on the spread in the projected potential distributions, variance percentage maps are returned together with the maps of the mean and standard deviation. Again, the results can be conveniently visualized with function `spplot` (Figs. 7.4 and 7.5).

```
> var.glm <- varianceAnalysis(predictions = ensemble.future,
        component1 = "PA", component2 = "newClim",
        fixed = c("glm"))
> # Generates Fig. 8.4
> spplot(var.glm$mean,
        at = seq(0,1,0.1),
        col.regions = colorRampPalette(c("white", "red3")),
        sp.layout= list(wrld, first = FALSE, lwd = 0.5))
> # Generates Fig. 8.5
> spplot(var.glm$variance,
        col.regions = rev(gray.colors(10, end = 1)),
        at = seq(0, 100, 10),
```

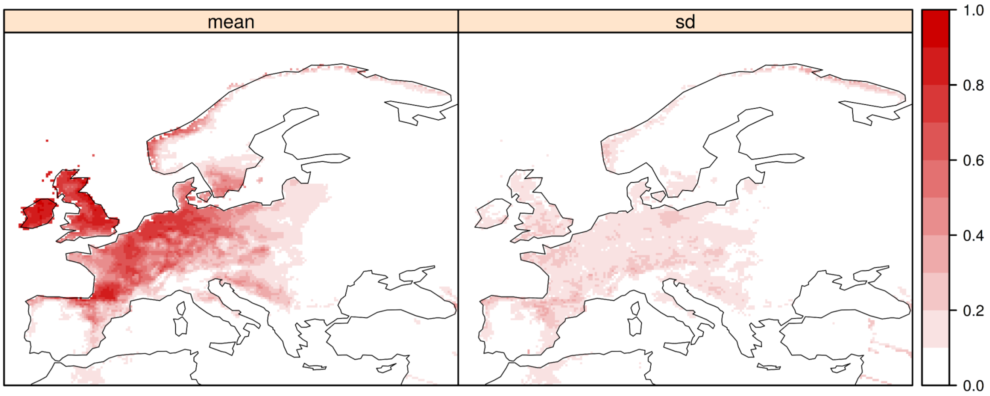

**Figure 7.4:** Mean and standard deviation of the SDM ensemble projections (GLM), formed by 7 RCMs × 10 pseudo–absence realizations (RS method, object `var.glm$mean`).

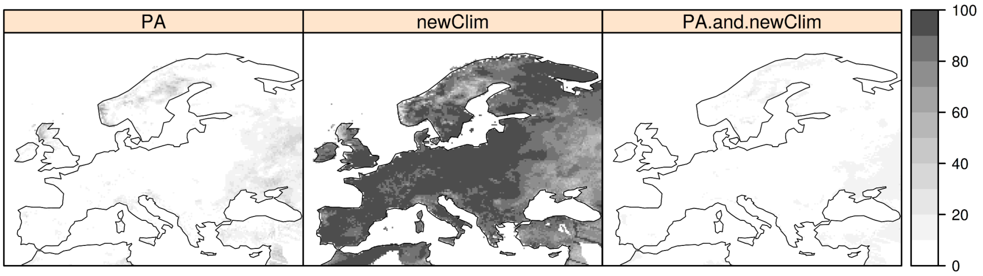

**Figure 7.5:** Variance percentage explained by each component: pseudo–absence realization (*PA*), RCM future climate projections (*newClim*) and their joint contribution (*PA.and.newClim*), considering GLM projections (object `var.glm$var`).

```
        sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

Figures 7.4 and 7.5 depict the ensemble SDM projections and the variance analysis results, applied to the set of projections that correspond to the 10 pseudo–absence realization and 7 climate projections (10 realizations x 7 RCMs). The mean suitability map and the standard deviation are shown in Figure 7.4, while Figure 7.5 are the variance fraction maps (%), depicting the contribution of each component (realization, RCM and realization & RCM) to the overall variance. For instance, the results displayed in Fig. 7.5 unveil that the RCM choice is by far the most important factor contributing to the ensemble spread, while pseudo–absence realization has some impact in areas that are outside the current domain of the Oak phylogeny H1 (e.g. Scandinavia).

Similarly, the next lines perform the same analysis, but considering considering MARS instead of GLM as the statistical modelling technique (Figs. 7.6 and 7.7):

```
> var.mars <- varianceAnalysis(predictions = ensemble.future,
        component1 = "PA", component2 = "newClim",
        fixed = c("mars"))
> # Generates Fig. 8.6
>  spplot(var.mars$mean,
        at = seq(0,1,0.1),
        col.regions = colorRampPalette(c("white", "red3")),
        sp.layout= list(wrld, first = FALSE, lwd = 0.5))
> # Generates Fig. 8.7
> spplot(var.mars$variance,
        at = seq(0, 100, 10),
        col.regions = rev(gray.colors(10, end = 1)),
        sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

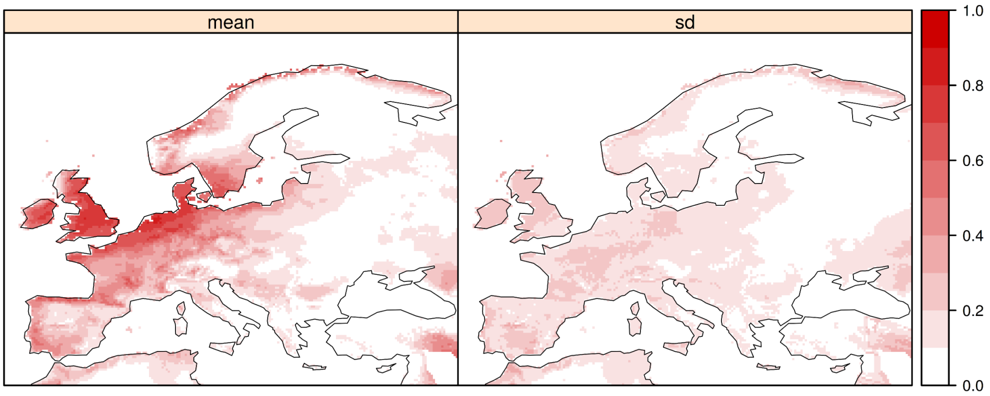

**Figure 7.6:** Same as Fig. 7.4, but considering MARS instead of GLM as statistical modeling technique for SDM production (object `var.mars$mean`).

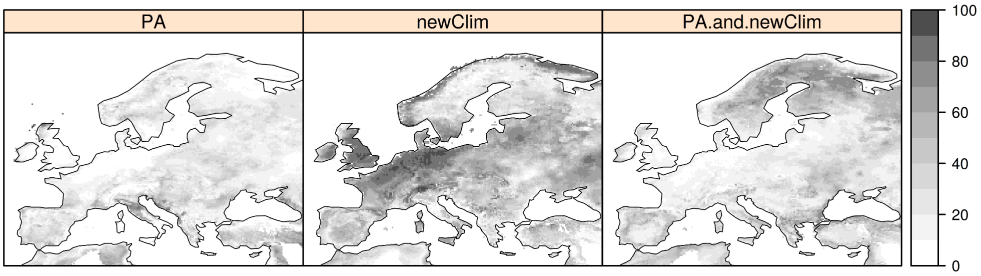

**Figure 7.7:** Same as Fig. 7.5, but considering MARS instead of GLM as the statistical modeling technique for SDM production (object `var.mars$var`).

Unlike GLM, in the case of MARS the ensemble spread (Fig. 7.6) is greatly affected by the pseudo–absence realization in a wide area of the study domain (Fig. 7.7). The much higher sensitivity of MARS to the pseudo–absence sample warns about its instability, while GLM reveals much better properties in terms of model stability and transferability. These findings are possible after variance analysis thanks to the utilities included in **mopa**, enabling a flexible experimental setup with a simple user interface. Model transferability is thus not apparent during the SDM calibration stage and is not coupled to model performance (even with the application of the 10-fold cross validation approach), so for instance TSS among realizations was 0.82 for GLM and 0.85 for MARS, and the mean AUC, 0.91 and 0.92 respectively. The uncertainty analysis results are extremely valuable for the construction of an ensemble of SDM projections that minimizes the risk of including unuseful realizations, thus yielding more plausible results.

In the same vein, the analysis of the SDM to the overall spread is achieved by adding a new component argument to `varianceAnalysis`, while the RCM projection (MPI model in this case) is kept as a fixed factor:

```
> MPI.var <- varianceAnalysis(ensemble.future,
          component1 = "PA",
          component2 = "SDM",
          fixed = c("MPI"))
```

## 7.6 SDM Ensemble Building

Finally, the ensemble forecast is built. In this particular example, we could discard those MARS projections that we consider are the result of bad transferability, e.g. corresponding to the pseudo–absence realizations that resulted in unrealistic predictions. Let us consider the simplified case where, after a more detailed analysis of the results, we conclude that MARS projections corresponding to pseudo–absence realization 8 along with GLM projections, are valid forecasts, then, as shown in the next example, the definitive ensemble is easily built with function `extractFromPrediction` and the utilities of the

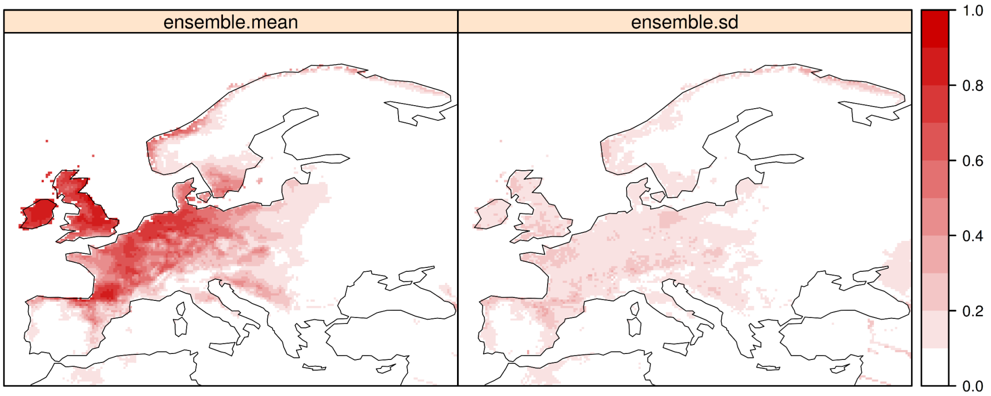

**Figure 7.8:** Future ensemble forecast (mean and standard deviation) of the suitability of the oak phylogeny H11 under climate conditions given by 7 different RCMs.

**raster** package. Here we calculate and plot the ensemble mean and standard deviation of the final SDM ensemble projections (Fig. 7.8):

```
> marsEns <- extractFromPrediction(ensemble.future,
        value = "mars")
> marsEnsPA08 <- extractFromPrediction(marsEns,
        value = "PA08")
> glmEns <- extractFromPrediction(ensemble.future,
        value = "glm")

> ensemble.future.def <-  stack(list(glmEns, marsEnsPA08))
> mean.ensemble <- stackApply(ensemble.future.def,
        fun = mean,
        indices = rep(1, nlayers(ensemble.future.def)))
> sd.ensemble <- stackApply(ensemble.future.def, fun = sd,
        indices = rep(1, nlayers(ensemble.future.def)))
> forecast.future <- stack(mean.ensemble, sd.ensemble)
> names(forecast.future) <- c("ensemble mean",
        "ensemble sd")
> # Generates Fig. 8.8
> spplot(forecast.future, at = seq(0,1,0.1),
        col.regions = colorRampPalette(c("white", "red3")),
        sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

Basically, this is a weighting exercise that favors GLM predictions in front of those of MARS, beyond the performance shown in the calibration phase.

The current ensemble forecast is obtained the same way but considering predictions made in reference climate, (Fig. 7.9). We suggest the **raster** package for further analysis on, for instance, habitat shifts among reference and future

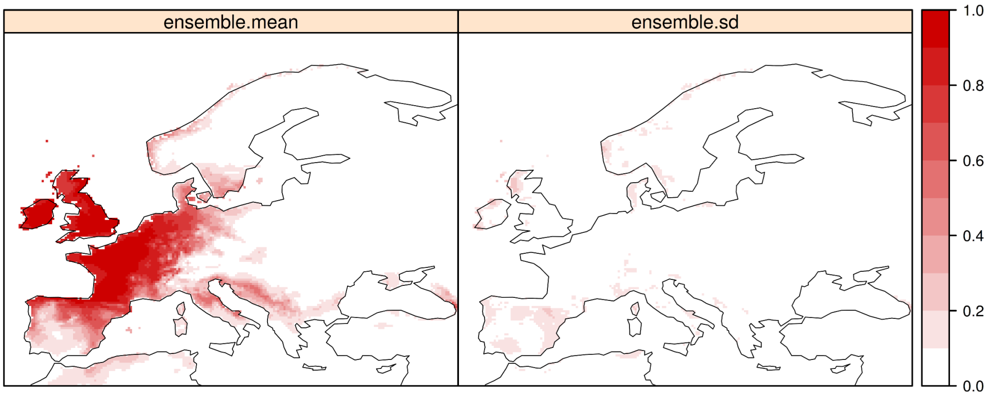

**Figure 7.9:** Ensemble forecast (mean and standard deviation) of the suitability of oak phylogeny H11 under reference climate conditions.

projections. Similarly, further typical manipulations can be done using other packages. For instance, binary (deterministic) presence/absence maps can be directly calculated with function `cut` from **raster**. We also suggest the **SDMTools** package, providing a set of analytical tools for SDM outputs.

In this work, we generated a set of SDM projections considering multiple combinations of climate change projections from a set of state-of-the-art RCMs, two popular statistical modeling methods (GLM and MARS) and different pseudo–absence realizations. The analyses undertaken with **mopa** enabled the identification of stable and plausible future projections for building the final ensemble. Moreover, through the illustrative case study used in this chapter, we show that the results of Chapter 5 are consistent for other groups of presences.

# Part IV

# Concluding Remarks

# CHAPTER 8

# Conclusions, Achievements and Future Work

## 8.1 Main Conclusions

This section aims to summarize the work done in order to achieve the three main objectives of the Thesis (included below in italics; see Chapter 3) as well as to briefly expose the most important achievements and conclusions which have been obtained in relation to them.

- *Objective 1: To compare and assess the limitations of standard methods for pseudo–absence data generation in terms of model performance, considering a representative set of SDMs. Research will be also conducted for the development of new methods, focusing on new alternatives for the implementation of the background extent restriction.*

  Regarding the first objective, in Chapter 4 we evaluated the influence of different pseudo–absence generation methods on model performance (area under the ROC curve), calibration (reliability diagrams) and the resulting suitability maps in reference climate. Five methods were compared, ranging from the classical random sampling of the whole region (RS), to the more elaborated three–step technique (TS), introducing a novel methodology for background extent restriction that does not penalize model performance.

As a result of this analysis we demonstrate that pseudo–absence sampling design can lead to a larger variation of model AUC (Fig. 4.4) than the choice of alternative SDMs, since the method for pseudo–absence generation strongly affected output SDM performance regardless of the modeling algorithm chosen and for all the Oak groups tested. The classical random sampling method (RS) yielded the lowest overall performance, while the target group (TG) approach attained high AUC values at the cost of poorly calibrated models, resulting in unreliable suitability maps. Methods that include environmental profiling in a previous step (RSEP, TS and TSKM), clearly outperformed both RS and TG, yielding high AUC values and better calibrated predictions, resulting in suitability maps with a higher resolution of the predicted probabilities. This stresses the importance of the pseudo–absence generation methods for the development of accurate and reliable SDMs.

The modeling algorithm is also an important factor affecting performance (Phillips *et al.*, 2009; Bedia *et al.*, 2011; Senay *et al.*, 2013). In this case, MAXENT and MARS performed better than GLM (Fig. 4.4). This agrees with previous studies pointing that more complex models tend to be more accurate (Elith & et al, 2006). Our results also suggest that MARS performance was more sensitive to the pseudo–absence configuration (Fig. 4.4).

- *Ovjective 2: To analyze pseudo–absence sampling as a determinant factor to characterize model stability and transferability in climate change conditions. This will be done by assessing the uncertainty in future ensembles of SDM projections (suitability maps) due to this factor. The interrelationship between predictors and pseudo–absences in this context will be also analyzed.*

With respect to the second objective, in Chapters 5 and 6 we explore the uncertainty in SDM future projections due to the sampling randomness in the background, for which different strategies for variable selection were considered (for building different sets of predictors). For this purpose, we performed 10 realizations of randomly generated pseudo–absences for each considered method (RS and TS) and sample size (prevalence). We tested the sensitivity to the pseudo–absence sample of three SDMs (GLM, RF and MARS) when projecting to future climate change conditions given by seven regional climate models (RCMs) from the ENSEMBLES project.

MARS proved to be the most sensitive algorithm to the pseudo–absence sample, whereas GLM was the most stable, being the uncertainty derived from different pseudo–absence realizations the lowest. These results are not related to the accuracy shown by each SDM in the calibration phase and, thus, future SDM projections can not be evaluated relying solely in the assessment of SDM performance. The contribution of the pseudo–absence realization to the uncertainty was higher in peripheral regions, specially for MARS, as a results of a limited extrapolation capability (see Figure 6.5). Although these results are consistent among different variable selection strategies, an increasing collinearity and dimensionality of the predictors potentiates the uncertainty derived from the use of pseudo–absences in future ensemble forecasts. On the other hand, using a small number of predictors could lead to over-prediction, specially for parsimonious SDMs (e.g. GLM).

Therefore, the sampling of pseudo–absence data constitutes a relevant source of uncertainty in SDM applications for climate change studies. Modeling algorithms are not equally affected, being parsimonious methods preferable in this context, since complex methods (such as MARS) are prone to yield wildly different future projections as a result of the pseudo–absence realization, indicating poor model transferability due to overfitting. Accounting for the pseudo–absence generation component of uncertainty is crucial to avoid the introduction of unreliable SDM signals confounding the final ensemble projections.

- *Objective 3: To develop an open-source modeling framework implementing the state-of-the-art SDM techniques, incorporating tools for pseudo–absence data generation and uncertainty analysis, envisaged to yield optimal future estimates of habitat suitability. Special attention will be paid to the transparent connection with standard climate data repositories, thus helping to bridge the gap between the niche and the climate modeling communities. This package will be develop in R language.*

Finally, with regard to the third objective, Chapter 7 introduces **mopa**, the R package developed as part of the work of this Thesis. We illustrate the functionalities of **mopa** by means of a case study that reproduces part of the analysis performed in Chapter 5, but considering an Oak phylogeny with other geographical distribution pattern (see Fig. 4.1).

The ability to quantitatively assess the individual contribution of each component in the modeling and prediction chain to the overall spread of the SDM outputs, as implemented in function `varianceAnalysis`, proved to be crucial in the evaluation of uncertainty and SDM transferability. While previously existing R packages already provide functionalities for SDM building and their assessment during the calibration stage, this is not related with their transferabilty into future climate conditions, as it has been shown through this Thesis, being therefore this feature specific of **mopa**. Other characteristic aspects introduced by the package consist of the novel methods for pseudo–absence generation, and the ability to perform a fine-tuning of these methods prior to model fitting.

Therefore, the new package **mopa** provides tools for species distribution modeling and for the straightforward design of relatively complex experiments with multiple factors or components affecting SDM uncertainty (pseudo–absence generation, climate projections, statistical technique, etc.), allowing users to quantify the contribution of different factors to the final uncertainty of the results, for optimal ensemble generation of future projections from SDMs. Furthermore, **mopa** is seamlessly integrated with other SDM-oriented packages as well as already standard geospatial data classes in R, thus providing maximum flexibility and inter-operability with a wide range of SDM-related tools. It is also integrated in the **climate4R** bundle for an easy retrieval and post-processing of climate data, helping to overcome complex, time-consuming data downloads and error-prone processing steps prior to SDM development. Hence, **mopa** takes a step forward in connecting the climate and niche modeling communities, which is of paramount importance for SDM applications to climate change studies.

Overall, as shown throughout the Thesis, the generation of pseudo–absences constitutes an extra source of uncertainty which can have a considerable impact in the projected results. This is highly relevant given that, as reported in this Thesis (Fig. 1.2 in Section 1.4), the most popular approach in species distribution modeling consists in using pseudo–absence data —generated by sampling the background areas from which presence records have not been collected—. There are different methods to this aim whose choice has an important effect on model performance and results (Chapter 4). However, there is not a consensus on the way in which pseudo–absences should be generated (e.g. Hengl *et al.*, 2009; Wisz & Guisan, 2009; Stokland *et al.*, 2011; Senay *et al.*, 2013). Nevertheless, in addition to the uncertainty that involves the use of alternative methods, we demonstrate that, independently from the method used, the variability of future SDM projections (uncertainty) derived from different realizations of pseudo–absences is significant, indicating transferability problems in some cases (Chapters 5 and 6), specially when a complex SDM is used (in this case MARS and to a lesser extent also RF). Therefore, we conclude that parsimonious models (e.g. GLMs), are preferable in the context of species distribution modeling under climate change conditions, although they generally obtain lower performance scores in the model training/calibration phase. In fact, as indicated in previous chapters, if true-absences are missing, the accuracy measures can only indicate how well models discriminate data considered in the model training process, but provides limited information about their real predictive capability (Václavík & Meentemeyer, 2009). Therefore, exploring different sources of uncertainty in future SDM projections is very important in order to avoid diluting insightful SDM signals with noise from inadequate (e.g. over-parameterized) SDMs (Thuiller *et al.*, 2004; Peterson *et al.*, 2011). To this aim, we implemented specific tools in the R package **mopa** (Chapter 7), which is of public domain and facilitates climate data preparation for the niche modeling community. Thus, the utilities in package **mopa** can help in the SDM production chain since the early stage (climate data retrieval and post-processing) to the ultimate phase in which a final set of SDM outputs is retained for ensemble generation and map production. This constitutes an important contribution, since SDMs have become a key tool for the vulnerability and impact assessment community to assess the impacts of climate change on the biological systems, an issue of current concern worldwide.

## 8.2 Publications and Contributions

This Thesis builds from the following research papers:

- Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M. & Gutiérrez, J.M. (2015) A framework for species distribution modelling with improved pseudo–absence generation. *Ecological Modelling* **312**, 166–174.
- Iturbide, M., Bedia, J. & Gutiérrez, J.M. (2015) Background sampling and transferability of species distribution models for climate change projections: Implications for the multimodel ensemble approach. Submitted to *Global and Planetary Change.*
- Iturbide, M., Bedia, J. & Gutiérrez, J.M. (2017) Tackling uncertainties to address the transferability of future species distribution models with package mopa. Submitted to *R journal.*

As well as from the contributions to the following events and initiatives:

- Poster presentation at BES Annual Symposium, Forest and Global Change, 2011, Cambridge (UK)
- Poster presentation at Klimagune Workshop "De Euskadi a Río +20", 2012 Bilbao (Spain)
- Poster presentation at Conference Adapting to Global Change in the Mediterranean Hotspots. 2013, Seville (Spain).
- Oral presentation at 5th international EcoSummit 2016, 29 Aug - 1 Sep. Montpellier (France).

- Work developed in the framework of WG1 of the EPS COST Action FP1202 (MaP-FGR, "Strengthening conservation: a key issue for adaptation of marginal/peripheral populations of forest trees to climate change in Europe").

Additionally, in parallel to the development of this Thesis, I collaborated in a number of initiatives dealing with climate data access and post-processing (bias adjustment) and its applicability and contribution to seasonal prediction and climate change projections. As a result I co-authored the following publications:

- Cofino, A., Bedia, J., Iturbide, M., Vega, M., Herrera, S., Fernández, J., Frías, M., Manzanas, R. & Gutiérrez, J.M. (2017) The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of Climate Services. *Climate Services* **in press**.
- Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C. & Gutiérrez, J.M. (2017) Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Services*, DOI:10.1016/j.cliser.2017.04.001.

## 8.3 Future Work

In connection to climate data post-processing and the preparation of appropriate variables for species distribution modeling, some of the results obtained during the realization of this Thesis have opened the door for the development of new works, by further integrating package **mopa** in the **climate4R** R bundle, in order to take full advantage of its functionalities in the preparation of predictor variables for species distribution modeling. In particular, in this Thesis we used the basic "delta" method (change factor) to produce the future climate projections (see Chapter 2), however, there are alternative methods for adjusting the bias of the GCM/RCM outputs. In this sense, and related to the publications shown above, I have acquired some experience in this field and have contributed to the development of related tools. Therefore, the work that follows this Thesis will consist in exploring alternative methods of bias adjustment in the production of predictor and projection variables that represent climate variations in a reduced time scale (e.g. weekly values). As well as analyzing the utilization of the resulting predictors in SDMs.

# Part V

# Summary in Spanish

# CHAPTER 9

# Resumen

De acuerdo con la normativa que regula los estudios de doctorado de la Universidad del País Vasco (UPV/EHU), se incluye a continuación un resumen de los principales resultados y conclusiones de la Tesis Doctoral.

## 9.1 Introducción

Los Modelos de Distribución de Especies (SDMs según sus siglas en inglés), son herramientas estadísticas utilizadas para la generación de predicciones probabilísticas de la presencia de poblaciones de especies en el espacio geográfico (Guisan & Zimmermann, 2000; Elith & et al, 2006). Los SDMs funcionan mediante el establecimiento de una relación empírica entre las localizaciones de presencia/ausencia conocidas (predictando) y las características físicas de su entorno (predictores). Dada la amenaza que supone el cambio climático, una aplicación popular de estos modelos es la proyección futura de las distribuciones potenciales de las especies —a partir de proyecciones climáticas futuras, véase el Capítulo 2— con el fin de evaluar temas claves en la conservación del medio ambiente, como el seguimiento de las respuestas biológicas al Cambio climático (Hamann & Wang, 2006), invasiones de especies (Jeschke & Strayer, 2008) o transmisión de enfermedades (Drake & Beier, 2014) entre otros. Por lo tanto, los SDMs se han convertido en una valiosa herramienta para la comunidad de evaluación de vulnerabilidad e impactos. Sin embargo, hay fuentes importantes de incertidumbre que afectan la credibilidad de las predicciones, como la capacidad predictiva de los SDMs fuera del dominio espacial y/o temporal de entrenamiento (conocida como transferibilidad o capacidad de extrapolación; Fronzek *et al.*, 2011), la incertidumbre asociada a los datos de entrenamiento (Mateo *et al.*, 2010b; Bedia *et al.*, 2013), las suposiciones subyacentes a los diferentes escenarios de emisiones (Nakićenović, 2000), los sesgos en los modelos climáticos globales/regionales (GCM/RCM) (Turco *et al.*, 2013) y otros (e.g. Falloon *et al.*, 2014, para una visión general). Entre ellos, en esta Tesis se destacan dos, la elección del SDM (ver e.g. Buisson *et al.*, 2010; Fronzek *et al.*, 2011; Garcia *et al.*, 2012), y la utilización de datos de pseudo–ausencia.

## 9.2 Generación de Pseudo–Ausencias

Además de los datos de presencia de una especie, la mayoría de los SDMs requieren también datos de ausencia para modelizar la respuesta binaria de presencia/ausencia (predictando) en función de las diferentes variables ambientales (predictores). En la mayoría de los casos no hay información explícita sobre la ausencia de las especies, de forma que la práctica más popular en la modelización de distribución de especies consiste en el uso de datos de pseudo–ausencia —generados mediante el muestreo de localidades donde no se han recogido registros de presencia—. Existen diferentes métodos de muestreo cuya elección tiene un efecto importante en el rendimiento y los resultados de los modelos (Capítulo 4, Iturbide *et al.*, 2015). Sin embargo, no hay un consenso sobre la manera en que se deben generar las pseudo–ausencias (e.g. Hengl *et al.*, 2009; Wisz & Guisan, 2009; Stokland *et al.*, 2011; Senay *et al.*, 2013).

El método más utilizado para generar pseudo–ausencias es la selección aleatoria considerando todo el área de estudio (método RS), sin embargo, esto aumenta el riesgo de introducir falsas ausencias en el modelo en lugares que, aún no habiendo registro de presencia, son en realidad adecuados para la especie, lo cual lleva a subestimaciones del nicho potencial (Anderson & Raza, 2010). Esto ocurre naturalmente debido a las interacciones bióticas y a las limitaciones de dispersión que no permiten que la especie habite en ciertos lugares, y también muy a menudo como resultado de sesgos en el muestreo de presencias. Frente a este problema, una práctica común es establecer una área de exclusión desde las localidades de presencia conocidas con el fin de minimizar la tasa de falsos negativos (e.g. Mateo *et al.*, 2010a; Bedia *et al.*, 2013).

Otros enfoques más elaborados aplican una exclusión geográfica ponderada, que mantiene las pseudo–ausencias fuera de las presencias usando mapas de distancia, o emplean un algoritmo para hacer una clasificación previa del área de estudio, de manera que las áreas clasidicadas como adecuadas se excluyen del muestreo, de esta manera, las pseudo–ausencias se alejan en el espacio ambiental (método RSEP, e.g. Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Barbet-Massin *et al.*, 2012; Liu *et al.*, 2013). Estas estrategias pretenden reducir el dominio de muestreo a aquellas áreas en las que es menos probable que ocurran falsas ausencias, mientras que el método del “grupo objetivo” (método TG) se ha postulado como una solución para eliminar parte del sesgo en el conjunto de datos de presencia, mediante el uso de localidades de presencia de otras especies como datos sesgados de pseudo–ausencia (Phillips *et al.*, 2009).

Otra cuestión crítica con respecto a los datos de pseudo–ausencia es la extensión del área en la cual se muestrean. Una distribución restringida de pseudo–ausencias alrededor de las localizaciones de presencia puede conducir a modelos engañosos, mientras que el muestreo sin restricción puede inflar artificialmente las estadísticas de evaluación del rendimiento del modelo (véase la Sección 1.6), así como el peso de variables predictivas menos informativas (Van der Wal & Shoo, 2009).

## *9.3 Transferabilidad de los modelos y cambio climático*

La capacidad predictiva por parte de los SDMs de la probabilidad de la distribución potencial de una población de especies en regiones o periodos diferentes a los utilizados para entrenar/calibrar el modelo, se conoce como transferibilidad o capacidad de extrapolación de los SDMs (dado un conjunto de presencias, (pseudo-) ausencias y predictores). Las distribuciones futuras se proyectan bajo el supuesto de que el rango ambiental actual será retenido bajo el Cambio Climático (Thuiller *et al.*, 2005). Por lo tanto, independientemente del escenario y el modelo climático considerado (GCM/RCM), los SDMs deben ser capaces de reproducir correctamente el rango ocupado en el futuro. A este respeto, las sobre-predicciones y sobre parametrizaciones de los modelos podrían explicar por qué dos SDMs calibrados con los mismos datos pueden producir diferentes proyecciones futuras (Thuiller *et al.*, 2004). Además de la capacidad de extrapolación del propio algoritmo de modelización, la transferibilidad de un SDM podría verse afectada de manera significativa debido a otras limitaciones metodológicas, como la disponibilidad y elección de variables predictoras apropiadas (Dormann *et al.*, 2008; Petitpierre *et al.*, 2016).

Una técnica común para abordar la incertidumbre en las proyecciones futuras de los SDMs se basa en la producción de conjuntos de proyecciones que derivan de múltiples SDMs, GCMs/RCMs, climatologías de referencia, etc., con el fin de abarcar un rango amplio de variabilidad de las proyecciones futuras (Araújo & New, 2007; Buisson *et al.*, 2010; Bagchi *et al.*, 2013; Baker *et al.*, 2015).

En particular, la contribución relativa de los SDMs a la variabilidad total de los conjuntos de proyecciones, ha demostrado ser la mayor (Buisson *et al.*, 2010; Fronzek *et al.*, 2011; Garcia *et al.*, 2012), ya que los resultados varían significativamente dependiendo de la técnica (GLMs, RF, MARS, etc.) y la configuración del modelo (veáse e.g. Araújo *et al.*, 2005; Beaumont *et al.*, 2008; Fronzek *et al.*, 2011). En este sentido, el enfoque de “conjuntos de proyecciones” tiene limitaciones, ya que asume que todos los SDMs son igualmente transferibles en condiciones del cambio climático, lo cual implica el riesgo de diluir las predicciones informativas con el ruido y error producido por SDMs menos útiles o defectuosos (Thuiller *et al.*, 2004; Peterson *et al.*, 2011). Sin embargo, no existe un criterio objetivo para realizar una selección de SDMs apropiados, ya que una validación adecuada de las proyecciones futuras de los SDMs es inherentemente imposible. Por lo tanto, la provisión de nuevas metodologías que evalúen la transferibilidad de los SDMs y que ayuden a reducir el rango de incertidumbre en los conjuntos de predicciones futuras es de suma importancia.

A este respecto, la falta de información sobre los lugares de ausencia de una población de especies plantea varios problemas metodológicos para los SDMs (Varela *et al.*, 2009). Las diferentes metodologías propuestas para la generación de pseudo–ausencias (Sección 1.4) se han evaluado atendiendo al rendimiento de los modelos resultantes en condiciones ambientales o climáticas de referencia (Sección 1.6). Sin embargo, se pueden obtener valores similares de rendimiento para predicciones de distribución potencial no similares (Lobo *et al.*, 2010). En este contexto, si faltan ausencias reales, las medidas de rendimiento de los modelos sólo pueden indicar el éxito de discriminación de los datos considerados en el proceso de entrenamiento o calibración, pero revelan poco acerca de su capacidad predictiva real (Václavík & Meentemeyer, 2009). De hecho, los SDMs que muestran un alto rendimiento en la fase de calibración pueden tener una capacidad de extrapolación limitada, no pudiendo predecir correctamente distribuciones futuras de las especies (Fronzek *et al.*, 2011). Sin embargo, la sensibilidad de diferentes SDMs a la muestra de pseudo–ausencias cuando se utilizan para proyectar distribuciones potenciales en un ambiente no muestreado (por ejemplo, bajo condiciones de Cambio Climático) ha sido ignorada hasta ahora.

Los objetivos principales de esta Tesis se enmarcan en este contexto, y se detallan en la siguiente sección.

## 9.4 Objetivos

A continuación se enumeran los objetivos de esta Tesis tal y como se presentan en el Capítulo 3:

1. Comparar y evaluar las limitaciones de los métodos estándar para la generación de datos de pseudo–ausencia en términos de rendimiento del modelo, considerando un conjunto representativo de SDMs. También se llevarán a cabo investigaciones para el desarrollo de nuevos métodos, centrándose en nuevas alternativas para restringir la extensión del área de muestreo de pseudo–ausencias.

2. Analizar el muestreo de pseudo–ausencias como un factor determinante para caracterizar la estabilidad y transferibilidad de los modelos en condiciones de cambio climático. Esto se llevará a cabo mediante la evaluación de la incertidumbre en conjuntos de proyecciones futuras (mapas de idoneidad) debido a este factor. También se analizará la interrelación entre predictores y pseudo–ausencias en este contexto.

3. Desarrollar un paquete de código abierto que implemente las técnicas de SDM de vanguardia, incorporando herramientas para la generación de datos de pseudo–ausencia y análisis de incertidumbre, dirigidos a producir estimaciones óptimas de la idoneidad de hábitats futuros. Se prestará especial atención a la conexión transparente con los repositorios de datos climáticos estándar, ayudando así a superar la brecha entre el las comunidades de modelización de nichos y del clima. Este paquete se desarrollará en el lenguaje de programación R.

Para responder a estas cuestiones se han desarrollado los estudios que se describen en los Capítulos 4, 5, 6 y 7.

A continuación, los principales resultados y conclusiones se resumen brevemente en español.

## 9.5 Resultados y Conclusiones

### 9.5.1 Objetivo 1

Con respecto al primer objetivo, en el capítulo 4 se evaluó la influencia de diferentes métodos de generación de pseudo–ausencia en el rendimiento de los modelos (AUC: área bajo la curva ROC), la calibración (diagramas de fiabilidad) y los mapas de idoneidad resultantes en condiciones climáticas de referencia. Se compararon cinco métodos, desde el clásico muestreo aleatorio de todo el área de estudio (RS), hasta la técnica más elaborada de tres pasos (TS), introduciendo una metodología novedosa para la restricción de la extensión del área de muestreo que no penaliza el rendimiento del modelo.

Como resultado de este análisis demostramos que el diseño de muestreo de pseudo–ausencias puede conducir a una mayor variación del AUC (Fig. 4.4) que la elección de SDMs alternativos, ya que el método de generación de pseudo–ausencias afectó fuertemente al rendimiento de los SDMs independientemente del algoritmo de modelización elegido y para todos los grupos de roble considerados. El método clásico de muestreo aleatorio (RS) produjo un rendimiento general menor, mientras que el grupo objetivo (TG) alcanzó altos valores de AUC pero produjo modelos mal calibrados, lo que resultó en mapas de idoneidad poco fiables. Los métodos que incluyen la clasificación previa del área de muestreo (RSEP, TS y TSKM), claramente superaron a RS y TG, produciendo valores de AUC altos y predicciones mejor calibradas, resultando en mapas de idoneidad con una mayor resolución de las probabilidades predichas. Esto subraya la importancia de los métodos de generación de pseudo–ausencias para el desarrollo de SDMs precisos y fiables.

El algoritmo de modelización es también un factor importante que afecta al rendimiento (Phillips *et al.*, 2009; Bedia *et al.*, 2011; Senay *et al.*, 2013). En este caso, MAXENT y MARS mostraron un mejor ajuste que GLM (Fig. 4.4). Esto concuerda con estudios previos que señalan que los modelos más complejos tienden a ser más precisos (Elith & et al, 2006). Nuestros resultados también sugieren que el rendimiento de MARS fue más sensible a la configuración de las pseudo–ausencias (Fig. 4.4).

### 9.5.2 Objetivo 2

Con respecto al segundo objetivo, en los Capítulos 5 y 6 exploramos la incertidumbre en las proyecciones futuras de los SDMs debido a la aleatoriedad del muestreo de pseudo–ausencias en el área de estudio, para lo cual se consideraron diferentes estrategias de selección de variables (para la construcción de diferentes conjuntos de predictores). Para ello, se generaron 10 realizaciones aleatorias de pseudo–ausencias para cada método (RS y TS) y tamaño de muestra (prevalencia) considerados. Hemos analizado la sensibilidad a la realización de pseudo–ausencias de tres SDMs (GLM, RF y MARS) cuando se proyectaron a condiciones futuras de cambio climático, dadas por siete modelos climáticos regionales (RCMs) del proyecto ENSEMBLES.

MARS demostró ser el algoritmo más sensible a la muestra de pseudo–ausencias, mientras que GLM fue el más estable, siendo la incertidumbre derivada de diferentes realizaciones de pseudo–ausencias la más baja. Estos resultados no están relacionados con el rendimiento mostrado por cada SDM en la fase de calibración. Por lo tanto, las proyecciones futuras de los SDMs no pueden ser evaluadas confiando únicamente en la evaluación del rendimiento de los SDMs. La contribución de la realización de la muestra de pseudo–ausencias a la incertidumbre fue mayor en las regiones periféricas, especialmente para MARS, como resultado de una capacidad de extrapolación limitada (ver Fig. 6.5)). Aunque estos resultados son consistentes entre las diferentes estrategias de selección de variables, una creciente colinealidad y dimensionalidad de los predictores potencian la incertidumbre derivada del uso de pseudo–ausencias en los conjuntos de proyecciones futuras. Por otra parte, el uso de un pequeño número de predictores podría conducir a la sobre-predicción, especialmente para SDMs parsimoniosos (e.g. GLM).

Por lo tanto, el muestreo de los datos de pseudo–ausencia constituye una fuente relevante de incertidumbre en las aplicaciones de los SDMs para estudios de cambio climático. Los algoritmos de modelización no se ven igualmente afectados, siendo los métodos parsimoniosos preferibles en este contexto, ya que los métodos complejos (como MARS) son propensos a producir proyecciones futuras muy diferentes como resultado de la realización de la muestra de pseudo–ausencias, lo que indica una pobre transferibilidad del modelo debido a problemas de sobre-ajuste. Es crucial tener en cuenta el factor o componente de generación de pseudo–ausencias en la incertidumbre para evitar la introducción de señales de SDMs no fiables que confundan los conjuntos de proyecciones finales.

### *9.5.3 Objetivo 3*

Finalmente, con respecto al tercer objetivo, el Capítulo 7 presenta el paquete de R **mopa**, desarrollado como parte del trabajo de esta Tesis. Se ilustran las funcionalidades de **mopa** mediante un estudio de caso que reproduce parte del análisis realizado en el Capítulo 5, pero considerando una filogenia de roble con otro patrón de distribución geográfica (ver Fig. 4.1).

La capacidad de evaluar cuantitativamente la contribución individual de cada componente en la cadena de predicción y modelización a la variabilidad general de los resultados de los SDMs, tal y como se implementa en la función `VarianceAnalysis`, resultó ser crucial en la evaluación de la incertidumbre y la transferibilidad del SDM. Mientras que los paquetes de R existentes ya proveen funcionalidades para la construcción de SDMs y su evaluación durante la etapa de calibración, esto no está relacionado con su transferibilidad en condiciones climáticas futuras, tal y como se demuestra a través de esta Tesis, siendo esta característica específica de **mopa**. Otros aspectos característicos que diferencian a **mopa** del resto de paquetes existentes, consisten en los nuevos métodos para la generación de pseudo–ausencias, y la capacidad de diseñar y afinar estos métodos antes de calibrar los SDMs.

Por lo tanto, el nuevo paquete **mopa** proporciona herramientas para la modelización de la distribución de especies y para el diseño directo de experimentos relativamente complejos con múltiples factores o componentes que afectan la incertidumbre del SDM (pseudo–ausencias, proyecciones climáticas, etc.), permitiendo a los usuarios cuantificar la contribución de diferentes factores a la incertidumbre final de los resultados, para la generación óptima de conjuntos de proyecciones futuras a partir de SDMs. Además, **mopa** se integra perfectamente con otros paquetes de R, proporcionando así máxima flexibilidad e interoperabilidad con una amplia gama de herramientas relacionadas con los SDMs. También está integrado en el conjunto de paquetes **climate4R** para una fácil obtención y post-procesado de datos climáticos, ayudando a superar descargas de datos complejas y pasos de procesamiento propensos a errores antes del desarrollo de los SDMs. Por lo tanto, **mopa** da un paso adelante en la conexión de las comunidades de modelización del clima y de nichos, lo cual es de suma importancia para las aplicaciones de los SDMs a estudios sobre Cambio Climático.

### *9.5.4 Conclusiones Generales*

Como se muestra a lo largo de la Tesis, la generación de pseudo–ausencias constituye una fuente adicional de incertidumbre que puede tener un impacto considerable en los resultados proyectados por parte de los SDMs. Esto es muy relevante dado que, como se indica en esta Tesis, el enfoque más popular en el modelización de distribución de especies consiste en el uso de datos de pseudo–ausencia (Sección 1.4). Existen diferentes métodos para la generación de pseudo–ausencias cuya elección tiene un efecto importante en el rendimiento y los resultados de los modelos (Capítulo 4), pero no hay un consenso sobre la manera en que se deben generar (por ejemplo Hengl *et al.*, 2009; Wisz & Guisan, 2009; Stokland *et al.*, 2011; Senay *et al.*, 2013). Sin embargo, además de la incertidumbre que conlleva el uso de métodos alternativos, demostramos que, independientemente del método utilizado, la variabilidad (incertidumbre) de las proyecciones futuras de los SDMs que derivan de diferentes realizaciones de pseudo–ausencias es significativa, indicando problemas de transferibilidad en algunos casos (Capítulos 5 y 6), especialmente cuando se utiliza un SDM complejo (en este caso MARS y en menor medida también RF). Por lo tanto, se concluye que los modelos parsimoniosos (como los GLMs), son preferibles en el contexto de la modelización de distribución de especies en condiciones de cambio climático, aunque generalmente obtengan valores de rendimiento más bajos en la fase de calibración de los modelos. De hecho, como se indica en capítulos anteriores, si no hay información sobre ausencias reales, las medidas de rendimiento sólo pueden indicar la capacidad de los modelos para discriminar los datos considerados en el proceso de construcción del modelo, pero proporcionan información limitada sobre su capacidad predictiva real (Václavík & Meentemeyer, 2009). Por lo tanto, la exploración de diferentes fuentes de incertidumbre en proyecciones futuras de SDMs es muy importante para evitar la introducción de señales erróneas de SDMs no transferibles (Thuiller *et al.*, 2004; Peterson *et al.*, 2011). Para ello, hemos implementado herramientas específicas en el paquete de R **mopa** (Capítulo 7), el cual es de dominio público y facilita la preparación de datos climáticos para la comunidad de modelos de nicho. Por lo tanto, las utilidades en el paquete **mopa** pueden ayudar en la cadena de producción y poryección de SDMs, desde la fase inicial (preparación de datos climáticos) hasta la fase final en la que se retiene un conjunto final de resultados óptimos. Esto constituye una contribución importante, ya que los SDMs se han convertido en una herramienta clave para que la comunidad de evaluación de vulnerabilidad e impactos con respecto a los riesgos que supone el Cambio Climático para los sistemas biológicos, un asunto de actualidad en todo el mundo.

## 9.6 Publicaciones y Contribuciones

Esta Tesis se basa en los siguientes artículos de investigación:

- Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M. & Gutiérrez, J.M. (2015) A framework for species distribution modelling with improved pseudo–absence generation. *Ecological Modelling* **312**, 166–174.

- Iturbide, M., Bedia, J. & Gutiérrez, J.M. (2015) Background sampling and transferability of species distribution models for climate change projections: Implications for the multimodel ensemble approach. Submitted to *Global and Planetary Change.*

- Iturbide, M., Bedia, J. & Gutiérrez, J.M. (2017) Tackling uncertainties to address the transferability of future species distribution models with package mopa. Submitted to *R journal.*

Así como en las contribuciones a los siguientes eventos e iniciativas:

- Poster presentation at BES Annual Symposium, Forest and Global Change, 2011, Cambridge (UK)

- Poster presentation at Klimagune Workshop ”De Euskadi a Río +20”, 2012 Bilbao (Spain)

- Poster presentation at Conference Adapting to Global Change in the Mediterranean Hotspots. 2013, Seville (Spain).

- Oral presentation at 5th international EcoSummit 2016, 29 Aug - 1 Sep. Montpellier (France).

- Work developed in the framework of WG1 of the EPS COST Action FP1202 (MaP-FGR, “Strengthening conservation: a key issue for adaptation of marginal/peripheral populations of forest trees to climate change in Europe”).

Adicionalmente, paralelamente al desarrollo de esta Tesis, colaboré en varias iniciativas relacionadas con el acceso y post-procesado (corrección de sesgo) de datos climáticos y su aplicabilidad y contribución a la predicción estacional y las proyecciones del cambio climático. Como resultado, soy coautora de las siguientes publicaciones:

- Cofino, A., Bedia, J., Iturbide, M., Vega, M., Herrera, S., Fernández, J., Frías, M., Manzanas, R. & Gutiérrez, J.M. (2017) The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of Climate Services. *Climate Services* **in press**.

- Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C. & Gutiérrez, J.M. (2017) Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Services*, DOI:10.1016/j.cliser.2017.04.001.

## 9.7 Líneas Futuras de Trabajo

En relación con el procesamiento de datos climáticos y la preparación de variables apropiadas para la modelización de distribución de especies, algunos de los resultados obtenidos durante la realización de esta Tesis han abierto la puerta para el desarrollo de nuevos trabajos, aprovechando al máximo las funcionalidades del conjunto de paquetes **climate4R** con el fin de construir variables predictoras para su uso en SDMs. En particular, en esta Tesis utilizamos el "método Delta" básico (factor de cambio) para producir las proyecciones climáticas futuras (ver Capítulo 2), sin embargo, existen métodos alternativos para ajustar el sesgo de las salidas de los GCM/RCM. En este sentido, y en relación con las publicaciones mostradas anteriormente, he adquirido cierta experiencia en este campo y he contribuido al desarrollo de herramientas relacionadas. Por lo tanto, el trabajo que sigue a esta Tesis consistirá en explorar métodos alternativos de ajuste de sesgo en la producción de variables que representen variaciones climáticas en una escala de tiempo reducida (por ejemplo, valores semanales). Así como analizar la utilización de los predictores resultantes en los SDM.

# Bibliography

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of Applied Ecology* **43**, 1223–1232.

Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: tests for oryzomys albigularis (rodentia: Muridae) in venezuela. *Journal of Biogeography* **30**, 591–605.

Anderson, R.P. & Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography* **37**, 1378–1393.

Araújo, M.B., Cabeza, M., Thuiller, W., Hannah, L. & Williams, P.H. (2004) Would climate change drive species out of reserves? an assessment of existing reserve-selection methods. *Global Change Biology* **10**, 1618–1626.

Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation* **96**, 331–345.

Araújo, M., F., G., Neto D. R., Pozo, I. & R, C. (2011) *Impactos, vulnerabilidad y adaptación al cambio climático de la biodiversidad española*, vol. 2. Fauna de vertebrados. Dirección general de medio Natural y Política Forestal. Ministerio de Medio Ambiente, y Medio Rural y Marino. Madrid, 640 páginas.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* **22**, 42–47.

Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography* **14**, 529–538.

Atkinson, P.M. & Lloyd, C.D. (1998) Mapping precipitation in Switzerland with ordinary and indicator kriging. Special issue: Spatial Interpolation Comparison 97. *Journal of Geographic Information and Decision Analysis* **2**, 72–86.

Bagchi, R., Crosby, M., Huntley, B., Hole, D.G., Butchart, S.H.M., Collingham, Y., Kalra, M., Rajkumar, J., Rahmani, A., Pandey, M., Gurung, H., Trai, L.T., Van Quang, N. & Willis, S.G. (2013) Evaluating the effectiveness of conservation site networks under climate change: accounting for uncertainty. *Global Change Biology* **19**, 1236–1248.

Baker, D.J., Hartley, A.J., Burgess, N.D., Butchart, S.H.M., Carr, J.A., Smith, R.J., Belle, E. & Willis, S.G. (2015) Assessing climate change impacts for vertebrate fauna across the West African protected area network using regionally appropriate climate projections. *Diversity and Distributions* **21**, 991–1003.

Baker, D.J., Hartley, A.J., Butchart, S.H.M. & Willis, S.G. (2016) Choice of baseline climate data impacts projected species' responses to climate change. *Global Change Biology* pp. n/a–n/a.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**, 327–338.

Barredo, J., Strona, G., de, R., Caudullo, G., Stancanelli, G. & San-Miguel-Ayanz, J. (2015) Assessing the potential distribution of insect pests: Case studies on large pine weevil (Hylobius abietis L) and horse-chestnut leaf miner (Cameraria ohridella) under present and future climate conditions in European forests. *EPPO Bulletin* **45**, 273–281.

Beaumont, L.J., Hughes, L. & Pitman, A.J. (2008) Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters* **11**, 1135–1146.

Bedia, J., Busqué, J. & Gutiérrez, J.M. (2011) Predicting plant species distribution across an alpine rangeland in northern spain: a comparison of probabilistic methods. *Applied Vegetation Science* **14**, 415–432.

Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C. & Gutiérrez, J.M. (2017) Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Services* .

Bedia, J., Herrera, S. & Gutiérrez, J.M. (2013) Dangers of using global bioclimatic datasets for ecological niche modeling. limitations for future climate projections. *Global and Planetary Change* **107**.

Beierkuhnlein, C., Thiel, D., Jentsch, A., Willner, E. & Kreyling, J. (2011) Ecotypes of european grass species respond differently to warming and extreme drought. *Journal of Ecology* **99**, 703–713.

Berger, L., DellaPietra, S.A. & Dellapietra, V.J. (1996) A maximum entropy approach to natural lenguaje processing. *Computational Linguistics* **22**, 39–71.

Biau, G., Zorita, E., von Storch, H. & Wackernagel, H. (1999) Estimation of Precipitation by Kriging in the EOF Space of theSea Level Pressure Field. *Journal of Climate* **12**, 1070–1085.

Bierman, S.M., Butler, A., Marion, G. & Kuehn, I. (2010) Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. *ECOGRAPHY* **33**, 451–460.

Brands, S., Herrera, S., San-Martín, D. & Gutiérrez, J. (2011) Validation of the ENSEMBLES Global Climate Models over southwestern Europe using probability density functions: A downscaler's perspective. *Climate Research* **48**, 145–161.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H. & Bollmann, K. (2013) Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography* **36**, 971–983.

Breiman, L. (1996) Bagging Predictors. *Mach. Learn.* **24**, 123–140.

Breiman, L. (2001) Random forests. *Machine Learning* **45**, 5–32.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H. & Guisan, A. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography* **21**, 481–497.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* **16**, 1145–1157.

Busby, J. (1991) *Nature Conservation: cost effective biological surveys and data analysis*, chap. BIOCLIM - a bioclimatic analysis and prediction system. CSIRO.

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

Christensen, J., Boberg, F., Christensen, O. & Lucas-Picher, P. (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters* **35**, L20709.

Christensen, J.H. & Christensen, O.B. (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change* **81**, 7–30.

Christensen, J.H., Kjellstrom, E., Giorgi, F., Lenderink, G. & Rummukainen, M. (2010) Weight assignment in regional climate models. *Climate Research* **44**, 179–194.

Christensen, O., Drews, M., Christensen, J., Dethloff, K., Ketelsen, K., Hebestadt, I. & Rinke, A. (2006) The HIRHAM regional climate model version 5. Tech. rep., Danish Meteorological Institute, Copenhagen, Denmark, `http://www.dmi.dk/dmi/en/print/tr06_17.pdf`.

Christensen, O.B., Drews, M., Christensen, J.H., Dethloff, K., Ketelsen, K., Hebestadt, I. & Rinke, A. (2008b) The HIRHAM Regional Climate Model. Version 5 (beta). Tech. Rep. 06-17, Danish Meteorological Institute (DMI).

Chu, D.P. & Dowsett, M.G. (1997) Dopant spectral distributions: Sample–independent response function and maximum entropy reconstruction. *Physical Review* **56**, 15167–15170.

Cofino, A., Bedia, J., Iturbide, M., Vega, M., Herrera, S., Fernández, J., Frías, M., Manzanas, R. & Gutiérrez, J.M. (2017) The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of Climate Services. *Climate Services* **in press**.

Collins, M., Booth, B., Harris, G., Murphy, J., Sexton, D. & M., W. (2006) Towards quantifying uncertainty in transient climate change. *Climate Dynamics* **27**, 127–147.

Curtis, C. & Bradley, B. (2016) Plant distribution data show broader climatic limits than expert-based climatic tolerance estimates. *PLoS ONE* **11**.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology* **88**, 2783–2792.

D'Amen, M., Zimmermann, N. & Pearman, P. (2013) Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography* **22**, 93–104, cited By 17.

DellaPietra, S.A., DellaPietra, V.J. & Lafferty, J. (1997) Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 1–13.

Dixon, P.M. (2006) Ripley's k function. *Encyclopedia of Environmetrics*, John Wiley & Sons, Ltd.

Domisch, S., Kuemmerlen, M., Jähnig, S. & Haase, P. (2013) Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. *Ecological Modelling* **257**, 1–10.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology* **89**, 3371–3386.

Drake, J. & Beier, J. (2014) Ecological niche and potential distribution of anopheles arabiensis in africa in 2050. *Malaria Journal* **13**, cited By 2.

Drake, J.M. & Bossenbroek, J.M. (2009) Profiling ecosystem vulnerability to invasion by zebra mussels with support vector machines. *Theoretical Ecology* **2**, 189–198.

Drake, J.M., Randin, C. & Guisan, A. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology* **43**, 424–432.

Drange, H. (2006) ENSEMBLES BCCR-BCM2.0 20C3M run1, daily values. CERA Database, World Data Center for Climate, Hamburg, Germany, http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=BCCR_BCM2.0_20C3M_1.

Dufresne, J. (2007) Ensembles ipsl-cm4 20c3m run1, daily values. CERA Database, World Data Center for Climate, http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_IPCM4_20C3M_1_D.

Dufresne, J. (2009) Ensembles stream2 ipslcm4-v2 20c3m run1, daily values. CERA Database, World Data Center for Climate, http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_IPCM4v2_20C3M_1_D.

Déqué, M., Rowell, D.P., Lüthi, D., Giorgi, F., Christensen, J.H., Rockel, B., Jacob, D., Kjellström, E., Castro, M.d. & Hurk, B.v.d. (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. *Climatic Change* **81**, 53–70.

Déqué, M., Somot, S., Sanchez-Gomez, E., Goodess, C.M., Jacob, D., Lenderink, G. & Christensen, O.B. (2012) The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability. *Climate Dynamics* **38**, 951–964.

Ehrenmann, F., Glaubitzer, S., Kopecky, D., Schmidt, J., Fluch, S., Maria Sehr, E. & Kremer, A. (2016) *Evolution of Trees and Forest Communities. Ten years of the EVOLTREE network*, chap. Evoltree E-Lab - An information system for forest genetics. ISBN: 978-2-9519296-3-9.

Elith, J. & et al (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* **41**, 263–274.

Engler, R., Randin, C.F., Vittoz, P., Czáka, T., Beniston, M., Zimmermann, N.E. & Guisan, A. (2009) Predicting future distributions of mountain plants under climate change: does dispersal capacity matter? *Ecography* **32**, 34–45.

Evans, J., Murphy, M.A., Holden, Z.A. & Cushman, S.A. (2011) *Predictive Species and Habitat Modeling in Landscape Ecology*, chap. Modeling species distribution and change using random forest [Chapter 8], pp. 139–159. Springer.

Falkowski, M.J., Evans, J.S., Martinuzzi, S., Gessler, P.E. & Hudak, A.T. (2009) Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. *Remote Sensing of Environment* **113**, 946–956.

Falloon, P., Challinor, A., Dessai, S., Hoang, L., Johnson, J. & Koehler, A.K. (2014) Ensembles and uncertainty in climate change impacts. *Frontiers in Environmental Science* **2**, 33.

Felicísimo, A.M., Muñoz, J., Villalba, C.J. & Mateo, R.G. (2011) *Impactos, vulnerabilidad y adaptación al cambio climático de la biodiversidad española*, vol. 1. Flora y vegetación. Oficina Española de Cambio Climático, Ministerio de Medio Ambiente y Medio Rural y Marino. Madrid, 552 pág.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**, 38–49.

Fordham, D.A., Wigley, T.M.L. & Brook, B.W. (2011) Multi-model climate projections for biodiversity risk assessments. *Ecological Applications* **21**, 3317–3331.

Franklin, J., Davis, F., Ikegami, M., Syphard, A., Flint, L., Flint, A. & Hannah, L. (2013) Modeling plant species distributions under future climates: how fine scale do climate projections need to be? *Global Change Biology* **19**, 473–483.

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling* **217**, 48–58.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics* **19**.

Fronzek, S., Carter, T. & Luoto, M. (2011) Evaluating sources of uncertainty in modelling the impact of probabilistic climate change on sub-arctic palsa mires. *Natural Hazards and Earth System Sciences* **11**, 2981–2995.

Garcia, R.A., Burgess, N.D., Cabeza, M., Rahbek, C. & Araújo, M.B. (2012) Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology* **18**, 1253–1269.

Gastón, A. & García-Viñas, J. (2011) Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. *Ecological Modelling* **222**, 2037–2041.

Giorgi, F. & Mearns, L.O. (1999) Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research* **104**, 6335–6352.

Golicher, D., Ford, A., Cayuela, L. & Newton, A. (2012) Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve. *International Journal of Geographical Information Science* **26**, 2049–2063.

González, S., Soto-Centeno, J. & Reed, D. (2011) Population distribution models: Species distributions are better modeled using biologically relevant data partitions. *BMC Ecology* **11**.

Gould, S.F., Beeton, N.J., Harris, R.M.B., Hutchinson, M.F., Lechner, A.M., Porfirio, L.L. & Mackey, B.G. (2014) A tool for simulating and communicating uncertainty when modelling species distributions under future climates. *Ecology and Evolution* **4**, 4798–4811.

Grantham, N. (2012) *Analyzing Multiple Independent Spatial Point Processes.* Ph.D. thesis, California Polytechnic State University.

Gude, J.A., Mitchell, M.S., Ausband, D.E., Sime, C.A. & Bangs, E.E. (2009) Internal validation of predictive logistic regression models for decision-making in wildlife management. *Wildlife Biology* **15**, 352–369.

Guisan, A., Edwards, T.C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157**, 89–100.

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

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological modelling* **135**, 147–186.

Gutiérrez, J.M., Herrera, S., San Martín, D., Sordo, C., Rodríguez, J., Frochoso, M., Ancell, R., Fernández, R., Cofiño, A., Pons, M. & Rodríguez, M. (2010) Escenarios Regionales Probabilísticos de Cambio Climático en Cantabria: Termopluviometría. Tech. rep., Gobierno de Cantabria-Consejería de Medio Ambiente y Universidad de Cantabria, Santander, Spain (In Spanish).

Hamann, A. & Wang, T. (2006) Potential effects of climate change on ecosystem and tree species distribution in british columbia. *Ecology* **87**, 2773–2786.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hanspach, J., Kühn, I., Schweiger, O., Pompe, S. & Klotz, S. (2011) Geographical patterns in prediction errors of species distribution models. *Global Ecology and Biogeography* **20**, 779–788.

Hastie, T., Tibshirani, R. & Friedman, J. (2010) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics, Springer, 2nd ed. 2009. corr. 3rd printing edn.

Haugen, J.E. & Haakensatd, H. (2005) Validation of hirham version 2 with 50 km and 25 km resolution. Tech. Rep. 9, Regional Climate Development Under Global Warming (RegClim).

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D. & New, M. (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research* **113**.

Hengl, T., Heuvelink, G.B.M. & Rossiter, D.G. (2007) About regression-kriging: From equations to case studies. *Computers & Geosciences* **33**, 1301–1315.

Hengl, T., Sierdsema, H., Radović, A. & Dilo, A. (2009) Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling* **220**, 3499–3511.

Hernández, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773–785.

Herrera, S., Fita, L., Fernández, J. & Gutiérrez, J. (2010) Evaluation of the mean and extreme precipitation regimes from the ENSEMBLES regional climate multimodel simulations over Spain. *Journal of Geophysical Research* **115**, 21117.

Hijmans, R.J. (2015) *raster: Geographic Data Analysis and Modeling.* R package version 2.4-20.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965–1978.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology* **12**, 2272–2281.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2017) *dismo: Species Distribution Modeling.* R package version 1.1-4.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological modelling* **145**, 111–121.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* **199**, 142–152.

Huebener, H. & Koerper, J. (2008) Ensembles stream2 egmam2 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=ENSEMBLES2_FUBEMA2_20C3M_1_D`.

Hutchinson, G.E. (1957) Population studies - animal ecology and demography - concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* **22**, 415–427.

IPCC (2000) IPCC Special Report Emissions Scenarios. Summary for Policy Makers. Special Report of IPCC Working Group III, World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP).

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M. & Gutiérrez, J.M. (2015) A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling* **312**, 166–174.

Jacob, D. (2001) A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. *Meteorology and Atmospheric Physics* **77**, 61–73.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O.B., Bouwer, L.M., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., Meijgaard, E.v., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.F., Teichmann, C., Valentini, R., Vautard, R., Weber, B. & Yiou, P. (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional Environmental Change* **14**, 563–578.

Jacob, D., Van den Hurk, B., Andrae, U., Elgered, G., Fortelius, C., Graham, L., Jackson, S., Karstens, U., Kopken, C., Lindau, R., Podzun, R., Rockel, B., Rubel, F., Sass, B., Smith, R. & Yang, X. (2001) A comprehensive model inter-comparison study investigating the water budget during the BALTEX–PIDCAP period. *Meteorology and Atmospheric Physics* **77**, 19–43.

Jaeger, E.B., Anders, I., Lüthi, D., Rockel, B., Schär, C. & Seneviratne, S.I. (2008) Analysis of ERA40-driven CLM simulations for Europe. *Meteorologische Zeitschrift* pp. 349–367.

Jeschke, J.M. & Strayer, D.L. (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *YEAR IN ECOLOGY AND CONSERVATION BIOLOGY 2008*, vol. 1134 of *ANNALS OF THE NEW YORK ACADEMY OF SCIENCES*, pp. 1–24, BLACKWELL PUBLISHING, 9600 GARSINGTON RD, OXFORD OX4 2DQ, OXEN, ENGLAND.

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.

Johns, T. (2008) Ensembles meto-hc-hadgem1 20c3m run1, 6h and 12h instantaneous values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_HADGEM_20C3M_1_6H12H`.

Johns, T. (2009a) Ensembles stream2 meto-hc-hadcm3c 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_HADCM3C_20C3M_1_D`.

Johns, T. (2009b) Ensembles stream2 meto-hc-hadgem2ao 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_HADGEM2_20C3M_1_D`.

Jolliffe, I. & Stephenson, D. (2003) *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons.

Juang, K.W. & Lee, D.Y. (1998) Simple Indicator Kriging for Estimating the Probability of Incorrectly Delineating Hazardous Areas in a Contaminated Site. *Environmental Science & Technology* **32**, 2487–2493.

Kjellström, E., Bärring, L., Gollvik, S., Hansson, U., Jones, C., Samuelsson, P., Rummukainen, M., Ullerstig, A., Willén, U. & Wyser, K. (2005) A 140–year simulation of European climate with the new version of the Rossby Centre regional atmospheric climate model (RCA3). *Rep. Meteorol. Climatol.* **108**, 681–687.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1137–1143.

Kriticos, D.b., Webber, B.d., Leriche, A.g., Ota, N., Macadam, I.h., Bathols, J. & Scott, J. (2012) CliMond: Global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution* **3**, 53–64, cited By (since 1996)53.

Kuhn, M. (2011) *Variable selection using the* `caret` *package*.

Lawrence, R.L., Wood, S.D. & Sheley, R.L. (2006) Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment* **100**, 356–362.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **28**, 385–393.

Liu, C., White, M., Newell, G. & Griffioen, P. (2013) Species distribution modelling for conservation planning in victoria, australia. *Ecological Modelling* **249**, 68–74.

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.

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

Lobo, J.M. & Tognelli, M.F. (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation* **19**, 1–7.

Maiorano, L., Cheddadi, R., Zimmermann, N.E., Pellissier, L., Petitpierre, B., Pottier, J., Laborde, H., Hurdu, B.I., Pearman, P.B., Psomas, A., Singarayer, J.S., Broennimann, O., Vittoz, P., Dubuis, A., Edwards, M.E., Binney, H.A. & Guisan, A. (2013) Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe. *Global Ecology and Biogeography* **22**, 302–317.

Manel, S., Dias, J.M., Buckton, S.T. & Ormerod, S.J. (1999) Alternative methods for predicting species distribution: an illustration with himalayan river birds. *Journal of Applied Ecology* **36**, 734–747.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* **38**, 921–931.

Maraun, D. (2012) Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophysical Research Letters* **39**, L06706.

Mateo, R.G., Croat, T.B., Felicísimo, A.M. & Muñoz, J. (2010a) Profile or group discriminative techniques? generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions* **16**, 84–94.

Mateo, R.G., Felicísimo, A.M. & Muñoz, J. (2011) Modelos de distribución de especies: Una revisión sintética. *Revista chilena de historia natural* **84**, 217–240.

Mateo, R.G., Felicísimo, A.M. & Muñoz, J. (2010b) Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science* **21**, 908–922.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* **41**, 811–823.

Meehl, G.A., Stocker, T., Collins, W., Friedlingstein, P., Gaye, A., Gregory, J., Kitoh, A., Knutti, R., Murphy, J., Noda, A., Raper, S., Watterson, I., Weaver, A. & Zhao, Z.C. (2007) Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. *Climate Change 2007: The Physical Science Basis.* (eds. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor & H.L. Miller), Cambridge University Press, Cambridge, UK, and New York, NY, USA.

Mellert, K., Deffner, V., Küchenhoff, H. & Kölling, C. (2015) Modeling sensitivity to climate change and estimating the uncertainty of its impact: A probabilistic concept for risk assessment in forestry. *Ecological Modelling* **316**, 211–216.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2017) *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.6-8.

Milborrow, S. (2015) *earth: Multivariate Adaptive Regression Splines.* Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper.

Moss, R.H., Edmonds, J.A., Hibbard, K.A., Manning, M.R., Rose, S.K., van Vuuren, D.P., Carter, T.R., Emori, S., Kainuma, M., Kram, T., Meehl, G.A., Mitchell, J.F.B., Nakicenovic, N., Riahi, K., Smith, S.J., Stouffer, R.J., Thomson, A.M., Weyant, J.P. & Wilbanks, T.J. (2010) The next generation of scenarios for climate change research and assessment. *Nature* **463**, 747–756.

Muñoz, J. & Felicísimo, A.M. (2004) Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science* **15**, 285–292.

Naimi, B. & Araujo, M.B. (2016) sdm: a reproducible and extensible r platform for species distribution modelling. *Ecography* **39**, 368–375.

Nakićenović, N. (2000) Greenhouse Gas Emissions Scenarios. *Technological Forecasting and Social Change* **65**, 149–166.

Niehörster, F. (2008) Ensembles egmam 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_IPCM4_20C3M_1_D`.

Nix, H.A. (1986) *Atlas of Elapid snakes of Australia*, chap. A biogeographic analysis of Australian Elaphid snakes. Australian Government Publishing Service, Canberra, Australia.

Norris, D., Rocha-Mendes, F., Frosini de Barros Ferraz, S., Villani, J. & Galetti, M. (2011) How to not inflate population estimates? spatial density distribution of white-lipped peccaries in a continuous atlantic forest. *Animal Conservation* **14**, 492–501.

Oney, B., Reineking, B., O'Neill, G. & Kreyling, J. (2013) Intraspecific variation buffers projected climate change impacts on *Pinus contorta*. *Ecology and Evolution* **3**, 437–449.

Pal, J., Giorgi, F., Bi, X., Elguindi, N., Solmon, F., Rauscher, S., Gao, X., Francisco, R., Zakey, A., Winter, J., Ashfaq, M., Syed, F., Sloan, L., Bell, J., Diffenbaugh, N., Karmacharya, J., Konaré, A., Martinez, D., da Rocha, R. & Steiner, A. (2007) Regional Climate Modeling for the Developing World: The ICTP RegCM3 and RegCNET. *Bulletin of the American Meteorological Society* **88**, 1395–1409.

Pearce, J.L. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* **133**, 225–245.

Pearman, P.B., D'Amen, M., Graham, C.H., Thuiller, W. & Zimmermann, N.E. (2010) Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography* **33**, 990–1003.

Pebesma, E.J. & Bivand, R.S. (2005) Classes and methods for spatial data in R. *R News* **5**, 9–13.

Peterson, A.T. (2011) Ecological niche conservatism: a time-structured review of evidence. *Journal of Biogeography* **38**, 817–827.

Peterson, A.T. & Nakazawa, Y. (2008) Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. *Global Ecology and Biogeography* **17**, 135–144.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R.H. & Stockwell, D.R.B. (2002) Future projections for mexican faunas under global climate change scenarios. *Letters to Nature* **416**, 626–629.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martinez-Meyer, E., Nakamura, M. & Araujo, M.B. (2011) *Ecological niches and geographic distributions.* No. 49 in Monographs in population biology, Princeton University.

Petit, R.J., Brewer, S., Bordács, S., Burg, K., Cheddadi, R., Coart, E., Cottrell, J., Csaikl, U.M., van Dam, B., Deans, J.D., Espinel, S., Fineschi, S., Finkeldey, R., Glaz, I., Goicoechea, P.G., Jensen, J.S., König, A.O., Lowe, A.J., Madsen, S.F., Mátyás, G., Munro, R.C., Popescu, F., Slade, D., Tabbener, H., de Vries, S.G.M., Ziegenhagen, B., de Beaulieu, J.L. & Kremer, A. (2002a) Identification of refugia and post-glacial colonisation routes of european white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* **156**, 49–74.

Petit, R.J., Csaikl, U.M., Bordács, S., Burg, K., Coart, E., Cottrell, J., van Dam, B., Deans, J.D., Dumolin-Lapégue, S., Fineschi, S., Finkeldey, R., Gillies, A., Glaz, I., Goicoechea, P.G., Jensen, J.S., König, A.O., Lowe, A.J., Madsen, S.F., Mátyás, G., Munro, R.C., Olalde, M., Pemonge, M.H., Popescu, F., Slade, D., Tabbener, H., Taurchini, D., de Vries, S.G.M., Ziegenhagen, B. & Kremer, A. (2002b) Chloroplast DNA variation in european white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management* **156**, 5–26.

Petit, R.J., Latouche-Halle, C., Pemonge, M. & Kremer, A. (2002c) Chloroplast

DNA variation of oaks in france and the influence of forest fragmentation on genetic diversity. *Forest Ecology and Management* **156**, 115–129.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C. & Guisan, A. (2016) Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecology and Biogeography* pp. n/a–n/a.

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

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**, 181–197.

Phillips, S.J., Dudík, M. & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the 21st International Conference on Machine Learning* pp. 655–662.

Pliscoff, P., Luebert, F., Hilger, H.H. & Guisan, A. (2014) Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment. *Ecological Modelling* **288**, 166–177.

Porfirio, L.L., Harris, R.M.B., Lefroy, E.C., Hugh, S., Gould, S.F., Lee, G., Bindoff, N.L. & Mackey, B. (2014) Improving the Use of Species Distribution Models in Conservation Planning and Management under Climate Change. *PLoS ONE* **9**.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **9**, 181–199.

R Core Team (2015) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Radu, R., Déqué, M. & Somot, S. (2008) Spectral nudging in a spectral regional climate model. *Tellus A* **60**, 898–910.

Räisänen, J. (2007) How reliable are climate models? *Tellus A* **59**, 2–29.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography* **33**, 1689–1703.

Randin, C.F., Engler, R., Normand, S., Zappa, M., Zimmermann, N.E., Pearman, P.B., Vittoz, P., Thuiller, W. & Guisan, A. (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Global Change Biology* **15**, 1557–1569.

Ripley, B. (2016) *tree: Classification and Regression Trees.* R package version 1.0-37.

Roeckner, E. (2007) ENSEMBLES ECHAM5-MPI-OM 20C3M run2, monthly mean values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_MPEH5_20C3M_2_MM`.

Roeckner, E. (2008) Ensembles stream2 echam5c-mpi-om 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_MPEH5C_20C3M_1_D`.

Royer, J. (2006) ENSEMBLES CNRM-CM3 20C3M run1, daily values. CERA Database, World Data Center for Climate, Hamburg, Germany, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_CNCM3_20C3M_1_D`.

Royer, J. (2008) Ensembles stream2 cnrm-cm33 20c3m run1, daily values. CERA Database, World Data Center for Climate, `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_CNCM33_20C3M_1_D`.

Rödder, D., Schmidtlein, S., Veith, M. & Lötters, S. (2009) Alien Invasive Slider Turtle in Unpredicted Habitat: A Matter of Niche Shift or of Predictors Studied? *PLOS ONE* **4**, e7843.

Samuelsson, P., Jones, C.G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U., Jansson, C., Kjellström, E., Nikulin, G. & Wyser, K. (2011) The Rossby Centre Regional Climate model RCA3: model description and performance. *Tellus A* **63**, 4–23.

San-Martín, D., Manzanas, R., Brands, S., Herrera, S. & Gutiérrez, J.M. (2016) Reassessing Model Uncertainty for Regional Projections of Precipitation with an Ensemble of Statistical Downscaling Methods. *Journal of Climate* **30**, 203–223.

Sanchez, E., Gallardo, C., Gaertner, M.A., Arribas, A. & Castro, M. (2004) Future climate extreme events in the Mediterranean simulated by a regional climate model: a first approach. *Global and Planetary Change* **44**, 163–180.

Santika, T. & Hutchinson, M.F. (2009) The effect of species response form on species distribution model prediction and inference. *Ecological Modelling* **220**, 2365–2379.

Scholkopf, B. & Smola, A.J. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA.

Senay, S.D., Worner, S.P. & Ikeda, T. (2013) Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE* **8**, e71218.

Serra-Varela, M.J., Grivet, D., Vincenot, L., Broennimann, O., Gonzalo-Jiménez, J. & Zimmermann, N.E. (2015) Does phylogeographical structure relate to climatic niche divergence? a test using maritime pine (pinus pinaster ait.). *Global Ecology and Biogeography* **24**, 1302–1313.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. & Kattan, M.W. (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21**, 128–138.

Stokland, J.N., Halvorsen, R. & Stø a, B. (2011) Species distribution modelling—Effect of design and sample size of pseudo-absence observations. *Ecological Modelling* **222**, 1800–1809.

Swets, J. (1988) Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Terribile, L.C., Diniz-Filho, J.A.F. & De Marco jr, P. (2010) How many studies are necessary to compare niche-based models for geographic distributions? Inductive reasoning may fail at the end. *Brazilian journal of biology = Revista brasleira de biologia* **70**, 263–269.

Therneau, T., Atkinson, B. & Ripley, B. (2017) *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-11.

Thuiller, W., Araújo, M.B., Pearson, R.G., Whittaker, R.J., Brotons, L. & Lavorel, S. (2004) Biodiversity conservation: Uncertainty in predictions of extinction risk. *Nature* **430**.

Thuiller, W., Georges, D., Engler, R. & Breiner, F. (2016) *biomod2: Ensemble Platform for Species Distribution Modeling.* R package version 3.3-7.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the*

*National Academy of Sciences of the United States of America* **102**, 8245–8250.

Townsend Peterson, A., Papeş, M. & Eaton, M. (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography* **30**, 550–560.

Turco, M., Sanna, A., Herrera, S., Llasat, M.C. & Gutiérrez, J.M. (2013) Large biases and inconsistent climate change signals in ENSEMBLES regional projections. *Climatic Change* **120**, 859–869.

Urban, M.C. (2015) Accelerating extinction risk from climate change. *Science* **348**, 571–573.

van der Linden, P. & Mitchell, J. (2009) ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project — European Environment Agency (EEA). Tech. rep., Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK.

Van der Wal, J. & Shoo, L.P. (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling* pp. 589–594.

van Meijgaard, E., van Ulft, L., van de Berg, W., Bosveld, F., van den Hurk, B., Lenderink, G. & Siebesma, A. (2008) The KNMI regional atmospheric climate model RACMO, version 2.1. Tech. Rep. 302, R. Neth. Meteorol. Inst., De Bilt, Netherlands, `http://www.knmi.nl/bibliotheek/knmipubTR/TR302.pdf`.

VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L. & Storlie, C. (2014) *SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises.* R package version 1.1-221.

Varela, S., Rodríguez, J. & Lobo, J.M. (2009) Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. *Journal of Biogeography* **36**, 1645–1655.

Vaughan, I.P. & Ormerod, S.J. (2005) The continuing challenges of testing species distribution models: Testing distribution models. *Journal of Applied Ecology* **42**, 720–730.

Verbyla, D.L. & Litvaitis, J.A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management* **13**, 783–787.

Václavík, T. & Meentemeyer, R.K. (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling* **220**, 3248–3258.

Warren, D.L., Glor, R.E., Turelli, M. & Funk, D. (2008) Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution* **62**, 2868–2883.

Warren, D.L. & Seifert, S.N. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications: A Publication of the Ecological Society of America* **21**, 335–342.

Wenger, S.J., Som, N.A., Dauwalter, D.C., Isaak, D.J., Neville, H.M., Luce, C.H., Dunham, J.B., Young, M.K., Fausch, K.D. & Rieman, B.E. (2013) Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biology* **19**, 3343–3354.

Wickham, H. & Chang, W. (2016) *devtools: Tools to Make Developing R Packages Easier*. R package version 1.12.0.

Wilby, R.L. & Dessai, S. (2010) Robust adaptation to climate change. *Weather* **65**, 180–185.

Winkler, J.A., Palutikof, J.P., Andresen, J.A. & Goodess, C.M. (1997) The Simulation of Daily Temperature Time Series from GCM Output. Part II: Sensitivity Analysis of an Empirical Transfer Function Methodology. *Journal of Climate* **10**, 2514–2532.

Wisz, M.S. & Guisan, A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? an information-theoretic approach based on simulated data. *BMC Ecology* **9**, 8.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & Group, N.P.S.D.W. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions* **14**, 763–773.

Wright, M.N. (2016) *ranger: A Fast Implementation of Random Forests*. R package version 0.6.0.

Zahn, M. & von Storch, H. (2010) Decreased frequency of North Atlantic polar lows associated with future climate warming. *Nature* **467**, 309–312.

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

Zhang, L., Liu, S., Sun, P., Wang, T., Wang, G., Zhang, X. & Wang, L. (2015) Consensus Forecasting of Species Distributions: The Effects of Niche Model Performance and Niche Properties. *PLoS ONE* **10**.

Zhang, Q. & Zhang, X. (2012) Impacts of predictor variables and species models on simulating Tamarix ramosissima distribution in Tarim Basin, northwestern China. *Journal of Plant Ecology* **5**, 337–345.