Tackling Uncertainties of Species Distribution Model Projections with Package mopa

by M. Iturbide, J. Bedia, and J.M. Gutiérrez

Abstract Species Distribution Models (SDMs) constitute an important tool to assist decision-making in environmental conservation and planning in the context of climate change. Nevertheless, SDM projections are affected by a wide range of uncertainty factors (related to training data, climate projections and SDM techniques), which limit their potential value and credibility. The new package **mopa** provides tools for designing comprehensive multi-factor SDM ensemble experiments, combining multiple sources of uncertainty (e.g. baseline climate, pseudo-absence realizations, SDM techniques, future projections) and allowing to assess their contribution to the overall spread of the ensemble projection. In addition, **mopa** is seamlessly integrated with the climate4R bundle and allows straightforward retrieval and post-processing of state-of-the-art climate datasets (including observations and climate change projections), thus facilitating the proper analysis of key uncertainty factors related to climate data.

Introduction

Species Distribution Models (SDMs) are statistical tools used for the generation of probabilistic predictions of the presence of biological entities in the geographical space (Guisan and Zimmermann, 2000; Elith and et al, 2006). SDMs operate through the establishment of an empirical link between known presence locations and the physical characteristics of their environment. A particular case are *Climate Envelope Models* (CEMs), where appropriate climatic variables are used as predictors to characterize the climatic conditions where a species can potentially live —typically in the form of *bioclimatic* variables (Nix, 1986; Busby, 1991)—. In the context of climate change, 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 and Wang, 2006; Jeschke and Strayer, 2008). These studies require suitable climate products to produce models at an adequate spatial resolution and varying geographical extents –up to global–, including historical climate databases (i.e. high resolution gridded observations) and future climate projections for different emission scenarios. However, the intricacy of climate data retrieval and post-processing of the existing climate products (e.g. the global and regional climate change projections available from the Earth System Grid Federation, ESGF, Taylor et al., 2011) has resulted in a wide use of ready-to-use products without considering their limitations for a particular case study (see Bedia et al., 2013). In this paper we fill this gap with the package **mopa** (Species Distribution MOdeling with Pseudo-Absences), which has been developed in the framework of the climate4R bundle for climate data access and post-processing, thus facilitating the use of state-of-the-art global and regional climate data for SDM projections.

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). A common approach to tackle different sources of uncertainty is based on producing *ensembles* of future SDM projections that encompass a wide range of variability by considering multiple choices of each of the factors/components involved in the modeling and projection chain (Araújo and New, 2007; Buisson et al., 2010; Bagchi et al., 2013). However, 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 training data (including the baseline climate) used to fit the SDMs characterizing the ecological niche (Mateo et al., 2010; Bedia et al., 2013; Baker et al., 2016), the varying extrapolation ability outside the training period/spatial extent of the different SDM techniques (known as SDM *transferability* in time/space; Bedia et al., 2011; Fronzek et al., 2011), the Global/Regional Climate Model (GCM/RCM) projections and biases (Turco et al., 2013) and others (see e.g.: Falloon et al., 2014, for an overview). Moreover, the *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).

The package **mopa** here presented has been designed to facilitate the design and analysis of comprehensive multi-factor SDM ensemble experiments, exploring different uncertainty factors such

as presence data sets, pseudo-absence realizations, baseline climate, modeling algorithms, and future climate projections. Moreover, **mopa** provides variance partition tools to assess the contribution of the different factors to the overall uncertainty/spread of the ensemble projection. We illustrate the functionality of the package with the case-study presented in Iturbide et al. (2018), focusing on the impact of the pseudo-absence data in the future distribution of a specific Oak phylogenetic group in Europe resulting from an ensemble of SDM projections considering three factors: 1) different SDMs techniques, 2) different realizations of randomly generated pseudo-absence data and 3) different climate projections produced over Europe from a ensemble of RCMs. The analyses undertaken with **mopa** reveal the sensitivity of SDMs to the pseudo-absence samples, affecting model stability and transferability to new climate conditions, with important implications for the construction of the final ensemble projections. We use and provide publicly available data to guarantee the reproducibility of the results.

mopa and the climate4R bundle for climate data access

The numerous climate databases available (both baselines and future projections) 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 (e.g. bias correction) prior to SDM development. This is also a major barrier for research reproducibility and data exchange. The climate4R bundle is 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 to worked examples and tutorials can be found for instance in Cofino et al. (2017), Bedia et al. (2017) and Frías et al. (2018). With this regard, **mopa** was developed as part of the climate4R ecosystem, so that typical climate data operations for SDM applications and conversion features to the data type handled by **mopa** are provided. Additionally, **mopa** includes a user guide with an end-to-end worked example of climate data retrieval, transformation and SDM development: `help(package = "mopa")`.

The "niche" of mopa within the "SDM ecosystem" in R

The popularity of R and its excellent 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 and Araújo, 2016), **biomod2** (Thuiller et al., 2016), **dismo** (Hijmans et al., 2017) and **SDMTools** (Van der Wal 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—. On the contrary, **mopa** is oriented towards the design and analysis of multi-factor ensembles of future SDM projections (considering as potential factors the presence data sets, the pseudo-absence realizations, the baseline climate, the modeling algorithms, and the future climate projections). The analysis of the resulting ensemble allows, for instance, assessing the problem of SDM transferability, which can not be properly evaluated during model calibration.

Besides, unlike previously existing packages, **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; 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 and Bivand, 2005; Bivand et al., 2013), allowing their usage at any stage of the modeling process (e.g. for data visualization and post-processing), and also extensibility to other SDM tools available in **sdm**, **biomod2**, ..., also handling the same spatial data classes.

Input data pre-processing

Climate data

SDM 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, here we use a set of present and future bioclimatic variables widely used in SDM applications based on precipitation and temperature climatologies (Busby, 1991), using the function `biovars` of package **dismo**. To this aim, we first exploit the climate4R functionalities to load and post-process observed precipitation and

temperature climatologies from the E-OBS gridded observational dataset (Haylock et al., 2008) and the simulations of 7 Regional Climate Model (RCMs) of the project ENSEMBLES (van der Linden and Mitchell, 2009, http://www.ensembles-eu.org) for the control (20C3M, 1971-2000) and future (A1B, 2071-2100) scenarios, including the application of bias-correction ("delta" method, e.g. Winkler et al., 1997; Zahn and von Storch, 2010).

```
> install.packages("mopa")
> library(mopa)
```

```
> destfile <- tempfile()
> url <- paste0("https://raw.githubusercontent.com/SantanderMetGroup/",
+   "mopa/master/data/biostack.rda")
> download.file(url, destfile)
> load(destfile, verbose = TRUE)
```

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) due to their differing adaptive responses to changes in their environment (Hernandez 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, Hernandez et al., 2006; Mateo et al., 2010). 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., 2002). 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
```

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.

Pseudo-absence generation

Most of SDMs require data not only from known presences of the biological entity, but also absence data in order to model the binary response presence/absence as a function of the different environmental variables. While the sampling efforts are typically focused on recording presence localities (atlases, natural history collections, targeted samplings, ...), in most cases there is no explicit information about the absence of the species. Therefore *pseudo-absence* generation is often required for SDM construction, by sampling the background of the study domain. Different methods have been proposed to this aim, whose choice has an important effect on the final SDM results, as highlighted in different previous studies (e.g.: Wisz and Guisan, 2009; Iturbide et al., 2015). However, there is no consensus on the best sampling design for generating pseudo–absences.

Pseudo-absence sampling in **mopa** is performed by the `pseudoAbsence` 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 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.

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

Table 1: Arguments of function `pseudoAbsences` controlling the parameter values involved in pseudo-absence generation.

Random Sampling (RS). The RS method is the simplest and most frequent way of generating pseudo-absences (Iturbide et al., 2015). 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 random sampling, a stratified sampling approach can be performed, based on homogeneous environmental conditions. To this aim, a clustering of the environmental space is applied following Senay et al. (2013) by setting argument `kmean` to `TRUE`:

```
> pa_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. 1):

```
> # Generates Fig. 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_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")
```

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 2), providing unrivalled functionality for the development and inter-comparison of multiple pseudo-absence setups for SDM refinement and ensemble prediction generation.

SDM fitting and prediction

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

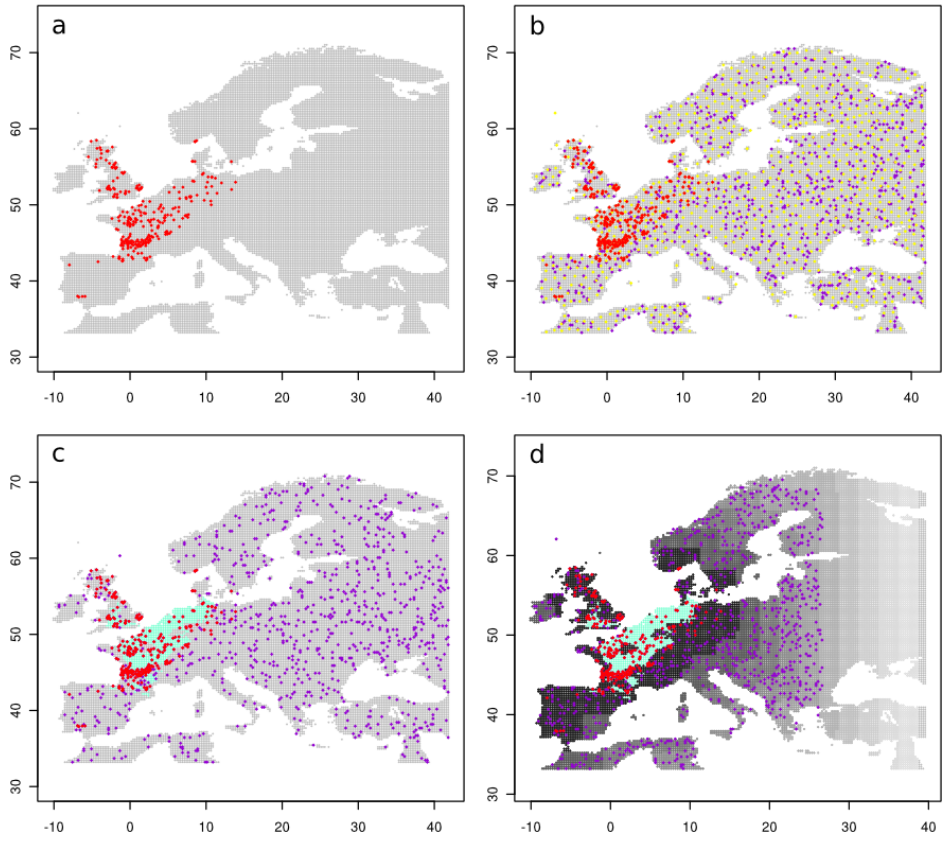

Figure 1: Pseudo-absence dataset maps, as generated by function `pseudoAbsences`. **(a)** Known presence locations of the Oak phylogeny H1 (red points) and initial background for pseudo-absence sampling (grey grid points). **(b)** pseudo-absences generated using the RS method randomly (purple points) 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).

used in SDM applications (see summary in Table 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 and 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"))
```

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 its transferability and biological meaning Van der Wal and Shoo (2009). With this regard, Iturbide et al. (2015) 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. With this regard, 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 (Fig. 2):

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

Table 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.

SDM technique	`algorithm` value	**pkg**`::function`	**Reference**
Generalized Linear Model	`"glm"`	**stats**`::glm`	Part of R
Random Forest	`"rf"`	**ranger**`::ranger`	Wright and Ziegler (2017)
Multivariate Adaptive Regression Splines	`"mars"`	**earth**`::earth`	Milborrow (2017)
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)

Table 3: SDM techniques available in **mopa** through the function `mopaTrain`. The corresponding `algorithm` argument values are also indicated.

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. 2).

```
> # Train TS model and generate Fig. 2
> trainTS <- mopaTrain(y = pa_TS, x = biostack$baseline, weighting = TRUE,
                k = 10, algorithm = c("glm", "mars"), diagrams = TRUE)
```

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 use 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")
```

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")
```

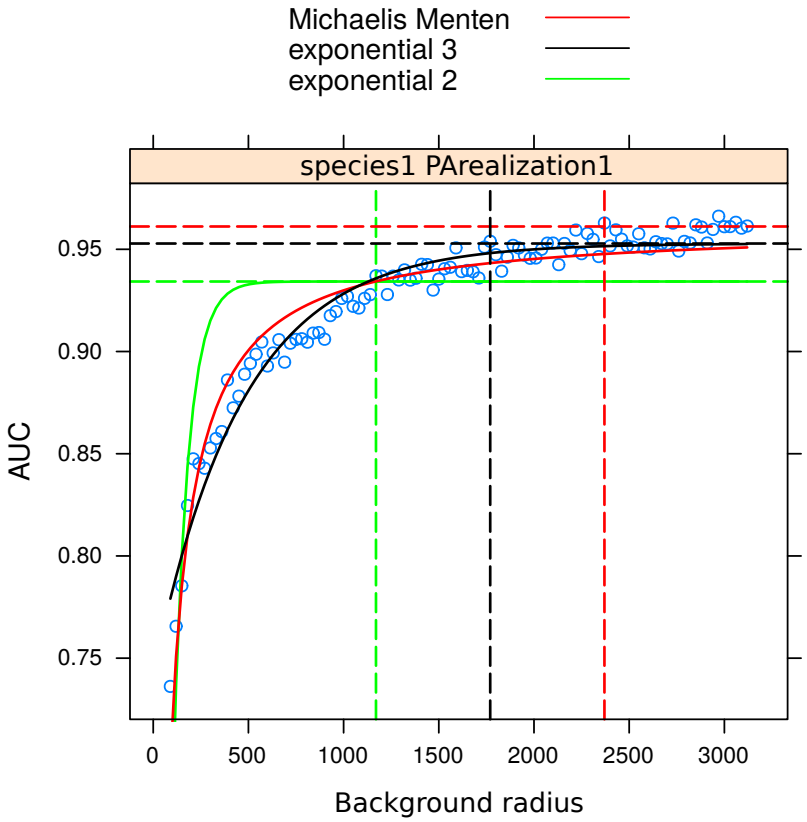

Figure 2: Asymptotic model fitting in SDMs using the TS approach for pseudo-absence generation. The blue points are the AUC values (y-axis) obtained by the SDMs for different background radius extents (x axis). Non-linear fits to the three asymptotic models considered (Michaelis Menten, 2 and 3-parameter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

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 to function `varImp` from package **caret** (Kuhn, 2017).

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)
```

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. 3).

```
> glm_runs_rcm1 <- extractFromPrediction(runs_rcm1, "glm")
> # Generates Fig. 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))
```

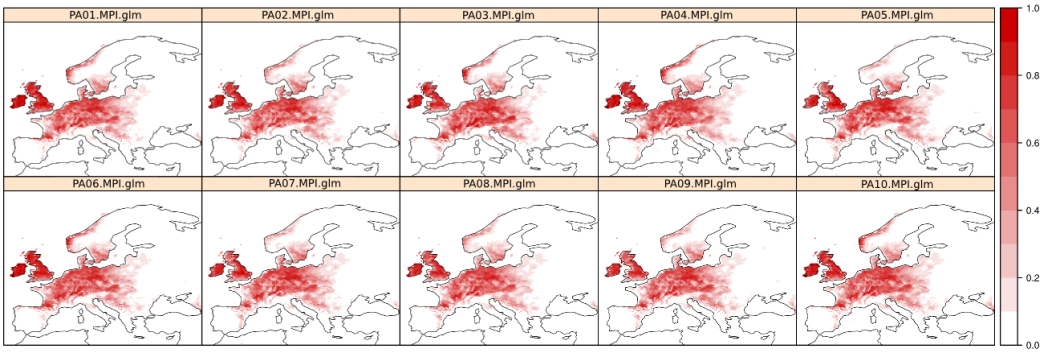

Figure 3: Future species distribution projections (2071-2100) according to the MPI RCM projections, considering 10 different pseudo-absence realizations of the RS method, as stored in the object `glm_runs_rcm1`.

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.

Partition of the uncertainty into components using ANOVA

The relative contribution of each component to the total ensemble spread/variability is implemented in **mopa** using an ANOVA approach, through the function `varianceAnalysis`, following the method in Déqué et al. (2012), also applied by 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, the RCM R and the combination of both PR, so $V = P + R + PR$. Let i be the index of the pseudo-absence realization ($i = 1, \ldots, 10$), j the index of the RCM ($j = 1, \ldots, 7$), and X_{ij} is the response (e.g.: the predicted distribution for the particular realization and climate projection). Then:

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

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

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

is the interaction term of the realization with the RCM (PR). The following example shows the analysis performed for the pseudo-absence realizations (`component1 = "PA"`) and the climate projections (`component2 = "newClim"`) 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. 4 and 5).

```
> var.glm <- varianceAnalysis(predictions = ensemble.future,
                component1 = "PA", component2 = "newClim", fixed = c("glm"))
> # Generates Fig. 4
> spplot(var.glm$mean,
                at = seq(0,1,0.1),
                col.regions = colorRampPalette(c("white", "red3")),
                sp.layout= list(wrld, first = FALSE, lwd = 0.5))
> # Generates Fig. 5
> spplot(var.glm$variance,
                col.regions = rev(gray.colors(10, end = 1)),
                at = seq(0, 100, 10),
                sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

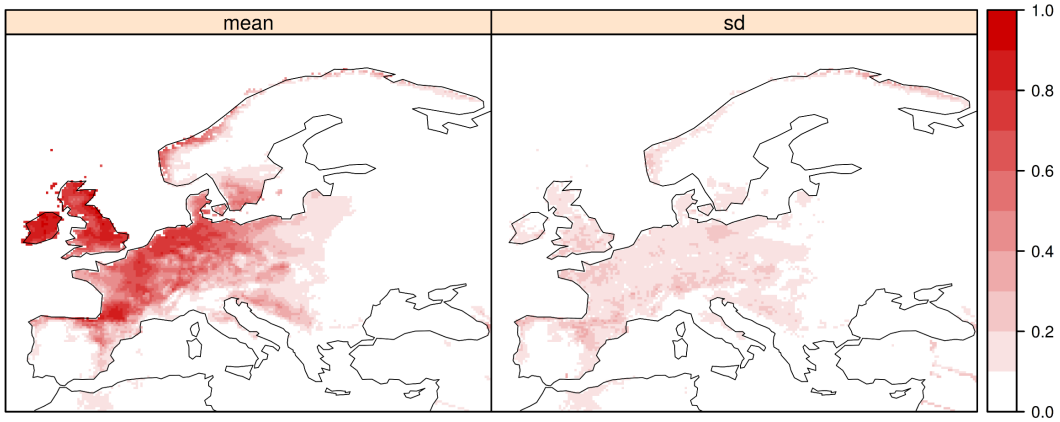

Figure 4: Mean and standard deviation of the SDM ensemble projections (GLM), formed by 7 RCMs $\times$ 10 pseudo-absence realizations (RS method, object `var.glm$mean`).

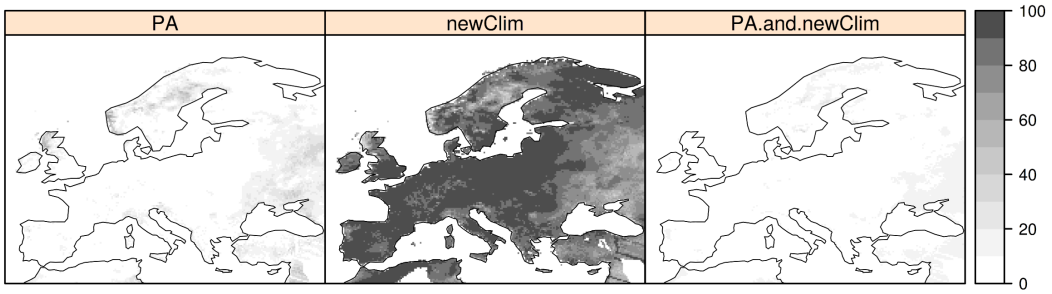

Figure 5: Variance percentage explained by each component: pseudo-absence realization (*PA*), RCM future climate projections (*newClim*) and their joint contribution (*PA.and.newClim*), considering GLM projections (object `var.glm$var`).

Figures 4 and 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 4, while Figure 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 Figure 5 unveil that the RCM choice (component `newClim`) 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 MARS instead of GLM as the statistical modeling technique (Figs. 6 and 7):

```
> var.mars <- varianceAnalysis(predictions = ensemble.future,
                component1 = "PA", component2 = "newClim", fixed = c("mars"))
> # Generates Fig. 6
```

```
>  spplot(var.mars$mean,
                at = seq(0,1,0.1),
                col.regions = colorRampPalette(c("white", "red3")),
                sp.layout= list(wrld, first = FALSE, lwd = 0.5))
> # Generates Fig. 7
> spplot(var.mars$variance,
                at = seq(0, 100, 10),
                col.regions = rev(gray.colors(10, end = 1)),
                sp.layout= list(wrld, first = FALSE, lwd = 0.5))
```

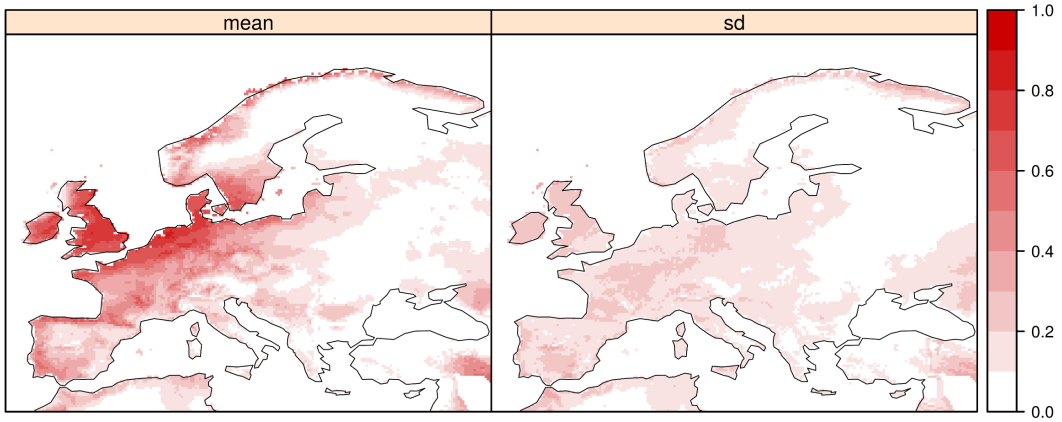

Figure 6: Same as Fig. 4, but considering MARS instead of GLM as statistical modeling technique for SDM production (object `var.mars$mean`).

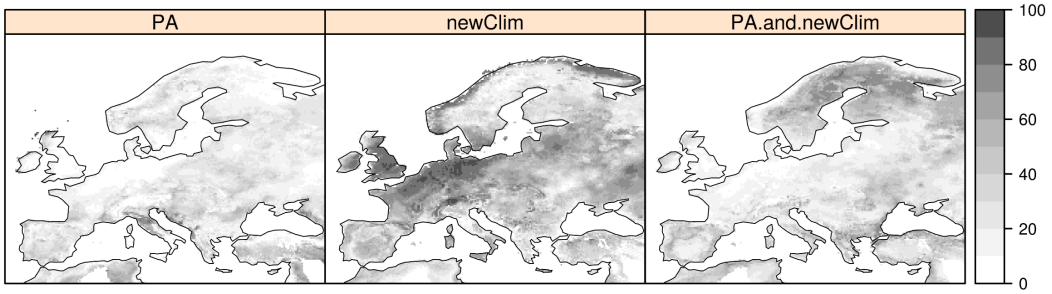

Figure 7: Same as Fig. 5, but considering MARS instead of GLM as the statistical modeling technique for SDM production (object `var.mars$var`).

Unlike GLM, in the case of MARS the ensemble spread (Fig. 6) is greatly affected by the pseudo-absence realization in a wide area of the study domain (Fig. 7), specially in peripheral regions. This is unequivocally diagnosed after applying function `varianceSummary`, which provides a summary of the results, including a graph (Fig. 8) and allowing the comparison of multiple results for a particular uncertainty component. This summary is based on the spatial subsetting of the study area, by specifying the number of subsets with argument `regions`. The output boxplot (Fig. 8) shows the spatial spread of the results (variance proportion explained by a component and the total standard deviation) in each region.

As a result, in Figure 8, we compare GLM and MARS (`var.glm` and `var.mars`) with regard to the variance proportion explained by the RCM choice (`component = 2L`), so that the percentage not explained by it, is associated to the pseudo–absence realization. From this summary, we can confirm a significantly higher sensitivity of MARS to the pseudo–absence sample across all regions.

```
        # Generates Fig. 8
> varianceSummary("glm" = var.glm, "mars" = var.mars,
                component = 2L, regions = c(6, 6), drawBoxplot = TRUE)
```

Alternatively, a `SpatialPolygons` object (package **sp**) can be passed to argument `regions` in order to focus the analysis on specific areas of interest. For illustrative purposes, in this example we use the climatic regions defined in the EU-funded PRUDENCE project (Christensen and Christensen, 2007), which is available at the climate4R package **visualizeR** (Frías et al., 2018):

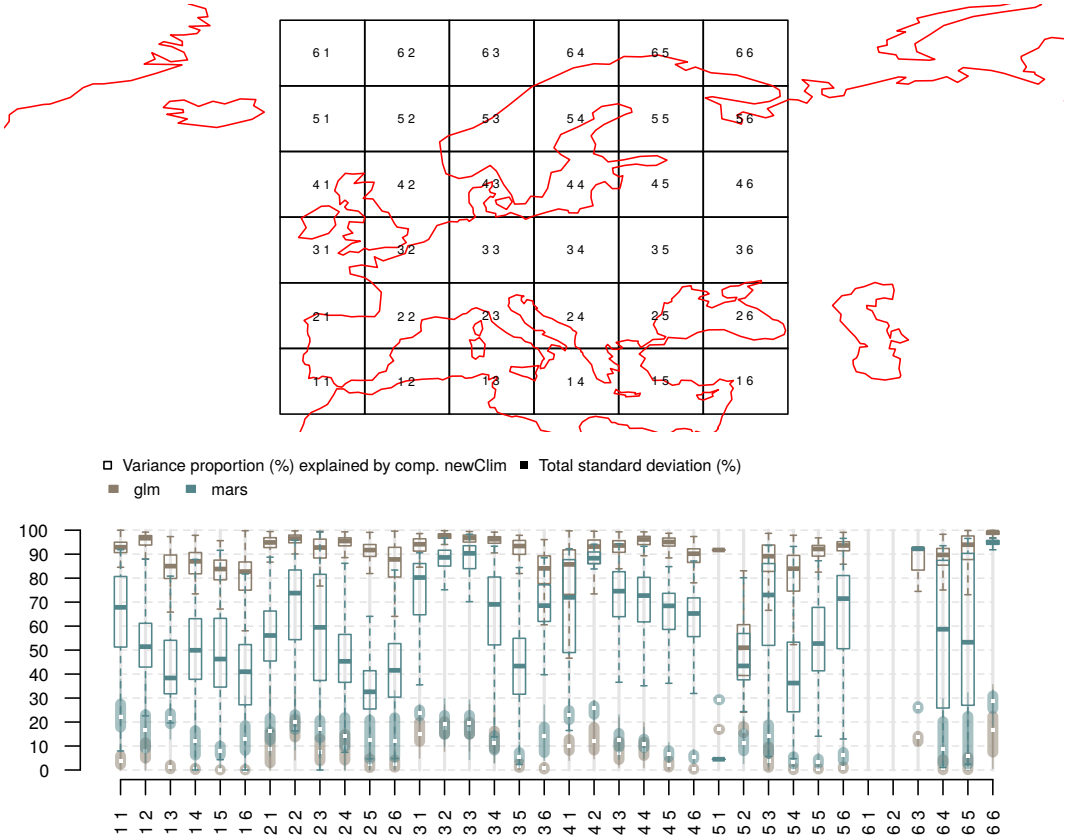

Figure 8: Summary of the variance analysis results generated with function `varianceSummary`, where GLM and MARS techniques (brown and blue respectively) are compared. Boxes account for the spatial spread of the results in each region. Empty boxes show the variance proportion explained by the RCM choice (component `newClim`) and filled boxes show the overall spread, this is, the standard deviation of the predicted probability expressed as a percentage. Thus, empty boxes show how is the total spread (filled boxes) distributed between components (`PA` and `newClim`). The x-axis corresponds to the regions shown in the map at the top.

```
> regiondir <- tempfile()
> download.file(paste0("https://github.com/SantanderMetGroup/",
+   "visualizeR/raw/devel/data/PRUDENCEregions.rda"), destfile = regiondir)
> load(regiondir)
> varianceSummary("glm" = var.glm, "mars" = var.mars,
                component = 2L, drawBoxplot = FALSE, regions = PRUDENCEregions)
```

Additionally, if argument `drawBoxplot` is set as `FALSE`, a simpler graph is obtained displaying the points of the spatial mean. This might be useful when multiple results are being compared in the same graph.

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 ANOVA 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 contribution of pseudo–absences in front SDM techniques to the overall spread is achieved by adding a new component argument to `varianceAnalysis`, while the RCM projection (MPI in this example) is kept as a fixed factor:

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

In case further uncertainty components are considered for predicting distributions (named in **mopa** as `SP`, `baseClim` and `foldModels`), these could also be analyzed by keeping several fixed factors, each corresponding to a component that is not being analyzed. This is explained in detail in the help document of function "varianceAnalysis".

```
> help(varianceAnalysis)
```

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 **raster** package. Here we calculate and plot the ensemble mean and standard deviation of the final SDM ensemble projections (Fig. 9):

```
> 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. 9
> 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.

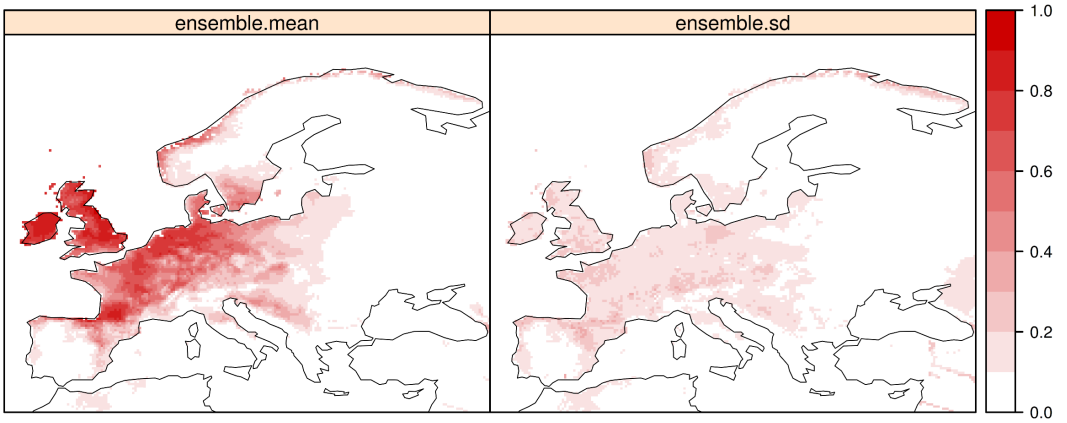

Figure 9: Future ensemble forecast (mean and standard deviation) of the suitability of the oak phylogeny H11 under climate conditions given by 7 different RCMs.

Summary

The impacts of climate change on the biological systems are of current concern worldwide, and future SDMs have become a key tool for the vulnerability and impact assessment 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 models is retained for ensemble generation and map production.

In this case-study, we illustrate the development of 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, enabling the identification of those members of the ensemble yielding consistent and plausible future estimates for final SDM building. With this regard, the ability to quantitatively assess the individual contribution of each factor to the overall SDM spread, as implemented in function `varianceAnalysis` proved crucial in the evaluation. While previously existing R packages already provide functionality for SDM building and their assessment during the calibration stage, we have shown that model performance, as evaluated by ordinary cross-validation, is not coupled to model transferabilty into future climate, being therefore this essential 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. Furthermore, the inter-operability of **mopa** with other SDM-related R packages enables maximum flexibility and eases the use of R for SDM applications in the framework of complex modeling exercises, for which multiple aspects have a varying contribution to the overall uncertainty.

Acknowledgements

We are grateful to the one anonymous referee for her/his valuable comments. We acknowledge the ENSEMBLES project (GOCE-CT-2003-505539), supported by the European Commission's 6th Framework Program for providing publicly the RCM simulations and observational data used in this study. We are also grateful to Rémy Petit and François Ehrenmann for providing the distribution of Oak phylogenies.

Bibliography

O. Allouche, A. Tsoar, and R. Kadmon. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43:1223–1232, 2006. URL https://doi.org/10.1111/j.1365-2664.2006.01214.x. [p7]

M. B. Araújo and M. New. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22:42–47, 2007. URL https://doi.org/10.1016/j.tree.2006.09.010. [p1]

M. B. Araújo, M. Cabeza, W. Thuiller, L. Hannah, and P. H. Williams. Would climate change drive species out of reserves? an assessment of existing reserve-selection methods. *Global Change Biology*, 10:1618–1626, 2004. URL https://doi.org/10.1111/j.1365-2486.2004.00828.x. [p1]

M. B. Araújo, R. J. Whittaker, R. J. Ladle, and M. Erhard. Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, 14:529–538, 2005. URL https://doi.org/10.1111/j.1466-822x.2005.00182.x. [p1]

R. Bagchi, M. Crosby, B. Huntley, D. G. Hole, S. H. M. Butchart, Y. Collingham, M. Kalra, J. Rajkumar, A. Rahmani, M. Pandey, H. Gurung, L. T. Trai, N. Van Quang, and S. G. Willis. Evaluating the effectiveness of conservation site networks under climate change: Accounting for uncertainty. *Global Change Biology*, 19:1236–1248, 2013. URL https://doi.org/10.1111/gcb.12123. [p1]

D. J. Baker, A. J. Hartley, S. H. M. Butchart, and S. G. Willis. Choice of baseline climate data impacts projected species' responses to climate change. *Global Change Biology*, 22:991–1003, 2016. URL https://doi.org/10.1111/gcb.13273. [p1]

M. Barbet-Massin, F. Jiguet, C. H. Albert, and W. Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 3:327–338, 2012. URL https://doi.org/10.1111/j.2041-210x.2011.00172.x. [p6]

L. J. Beaumont, L. Hughes, and A. J. Pitman. Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters*, 11:1135–1146, 2008. URL https://doi.org/10.1111/j.1461-0248.2008.01231.x. [p1]

J. Bedia, J. Busqué, and J. M. Gutiérrez. Predicting plant species distribution across an alpine rangeland in northern spain: a comparison of probabilistic methods. *Applied Vegetation Science*, 14:415–432, 2011. URL https://doi.org/10.1111/j.1654-109x.2011.01128.x. [p1]

J. Bedia, S. Herrera, and J. M. Gutiérrez. Dangers of using global bioclimatic datasets for ecological niche modeling. limitations for future climate projections. *Global and Planetary Change*, 107:1–12, 2013. URL https://doi.org/10.1016/j.gloplacha.2013.04.005. [p1]

J. Bedia, N. Golding, A. Casanueva, M. Iturbide, C. Buontempo, and J. M. Gutiérrez. Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Services*, 2017. URL https://doi.org/10.1016/j.cliser.2017.04.001. [p2]

C. Beierkuhnlein, D. Thiel, A. Jentsch, E. Willner, and J. Kreyling. Ecotypes of european grass species respond differently to warming and extreme drought. *Journal of Ecology*, 99:703–713, 2011. URL https://doi.org/10.1111/j.1365-2745.2011.01809.x. [p3]

R. S. Bivand, E. Pebesma, and V. Gomez-Rubio. *Applied Spatial Data Analysis with R, Second Edition*. Springer-Verlag, 2013. URL http://www.asdar-book.org/. [p2]

L. Buisson, W. Thuiller, N. Casajus, S. Lek, and G. Grenouillet. Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, 16:1145–1157, 2010. URL https://doi.org/10.1111/j.1365-2486.2009.02000.x. [p1]

J. R. Busby. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, chapter BIOCLIM - a bioclimatic analysis and prediction system. CSIRO, 1991. [p1, 2]

J. H. Christensen and O. B. Christensen. A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, 81(1):7–30, 2007. ISSN 0165-0009, 1573-1480. URL https://doi.org/10.1007/s10584-006-9210-7. [p11]

A. Cofino, J. Bedia, M. Iturbide, M. Vega, S. Herrera, J. Fernández, M. D. Frías, R. Manzanas, and J. M. Gutiérrez. The ECOMS User Data Gateway: Towards Seasonal Forecast Data Provision and Research Reproducibility in the Era of Climate Services. *Climate Services*, in press, 2017. URL https://doi.org/10.1016/j.cliser.2017.07.001. [p2]

M. Déqué, S. Somot, E. Sanchez-Gomez, C. M. Goodess, D. Jacob, G. Lenderink, and O. B. Christensen. The spread amongst ENSEMBLES regional scenarios: Regional climate models, driving general circulation models and interannual variability. *Climate Dynamics*, 38:951–964, 2012. URL https://doi.org/10.1007/s00382-011-1053-x. [p9]

J. Elith and et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129–151, 2006. URL https://doi.org/10.1111/j.2006.0906-7590.04596.x. [p1]

P. Falloon, A. Challinor, S. Dessai, L. Hoang, J. Johnson, and A.-K. Koehler. Ensembles and uncertainty in climate change impacts. *Frontiers in Environmental Science*, 2:33, 2014. URL https://doi.org/10.3389/fenvs.2014.00033. [p1]

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67, 1991. [p6]

S. Fronzek, T. R. Carter, and M. Luoto. 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, 2011. URL https://doi.org/10.5194/nhess-11-2981-2011. [p1]

M. D. Frías, M. Iturbide, R. Manzanas, J. Bedia, J. Fernández, S. Herrera, A. S. Cofiño, and J. M. Gutiérrez. An R package to visualize and communicate uncertainty in seasonal climate prediction. *Environmental Modelling & Software*, 99(Supplement C):101–110, 2018. ISSN 1364-8152. URL https://doi.org/10.1016/j.envsoft.2017.09.008. [p2, 11]

A. Guisan and N. E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological modelling*, 135:147–186, 2000. URL https://doi.org/10.1016/s0304-3800(00)00354-9. [p1]

A. Guisan, T. C. Edwards, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 157:89–100, 2002. URL https://doi.org/10.1016/s0304-3800(02)00204-1. [p6]

A. Hamann and T. Wang. Potential effects of climate change on ecosystem and tree species distribution in british columbia. *Ecology*, 87:2773–2786, 2006. URL https://doi.org/10.1890/0012-9658(2006)87[2773:peocco]2.0.co;2. [p1]

M. R. Haylock, N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research*, 113:D20119, 2008. URL https://doi.org/10.1029/2008jd010201. [p3]

P. A. Hernandez, C. H. Graham, L. L. Master, and D. L. Albert. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29: 773–785, 2006. URL https://doi.org/10.1111/j.0906-7590.2006.04700.x. [p3]

R. J. Hijmans. *raster: Geographic Data Analysis and Modeling*, 2015. URL https://CRAN.R-project.org/package=raster. R package version 2.4-20. [p2]

R. J. Hijmans, S. Phillips, J. Leathwick, and J. Elith. *dismo: Species Distribution Modeling*, 2017. URL https://CRAN.R-project.org/package=dismo. R package version 1.1-4. [p2, 7]

M. Iturbide, J. Bedia, S. Herrera, O. del Hierro, M. Pinto, and J. M. Gutiérrez. A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312:166–174, 2015. URL https://doi.org/10.1016/j.ecolmodel.2015.05.018. [p2, 3, 4, 5, 6]

M. Iturbide, J. Bedia, and J. M. Gutiérrez. Background sampling and transferability of species distribution model ensembles under climate change. *Global and Planetary Change*, 2018. URL https://doi.org/10.1016/j.gloplacha.2018.03.008. In press. [p2]

J. M. Jeschke and D. L. Strayer. Usefulness of bioclimatic models for studying climate change and invasive species. In *Year in Ecology and Conservation Biology*, volume 1134 of *Annals of the New York Academy of Sciences*, pages 1–24. Blackwell Publishing, 9600 Garsington RD, Oxford OX4 2DQ, Oxen, England, 2008. [p1]

M. Kuhn. *caret: Classification and Regression Training*, 2017. URL https://CRAN.R-project.org/package=caret. R package version 6.0-76. [p8]

R. G. Mateo, Ángel M. Felicísimo, and J. Muñoz. 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, 2010. URL https://doi.org/10.1111/j.1654-1103.2010.01198.x. [p1, 3]

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. URL https://CRAN.R-project.org/package=e1071. R package version 1.6-8. [p4, 7]

S. Milborrow. *earth: Multivariate Adaptive Regression Splines*, 2017. URL https://CRAN.R-project.org/package=earth. R package version 4.5.0. [p7]

B. Naimi and M. B. Araújo. sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, 39:368–375, 2016. URL https://doi.org/10.1111/ecog.01881. [p2]

H. A. Nix. *Atlas of Elapid Snakes of Australia*, chapter A biogeographic analysis of Australian Elaphid snakes. Australian Government Publishing Service, Canberra, Australia, 1986. [p1]

E. J. Pebesma and R. S. Bivand. *Classes and Methods for Spatial Data in R*, 2005. URL https://cran.r-project.org/doc/Rnews/Rnews_2005-2.pdf. [p2]

A. T. Peterson, J. Soberón, R. G. Pearson, R. P. Anderson, E. Martinez-Meyer, M. Nakamura, and M. B. Araujo. *Ecological Niches and Geographic Distributions*. Monographs in population biology. Princeton University, 2011. ISBN 978-0-691-13688-2. [p1]

R. J. Petit, U. M. Csaikl, S. Bordács, K. Burg, E. Coart, J. Cottrell, B. van Dam, J. D. Deans, S. Dumolin-Lapègue, S. Fineschi, R. Finkeldey, A. Gillies, I. Glaz, P. G. Goicoechea, J. S. Jensen, A. O. König, A. J. Lowe, S. F. Madsen, G. Mátyás, R. C. Munro, M. Olalde, M.-H. Pemonge, F. Popescu, D. Slade, H. Tabbener, D. Taurchini, S. G. M. de Vries, B. Ziegenhagen, and A. Kremer. 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, 2002. URL https://doi.org/10.1016/s0378-1127(01)00645-4. [p3]

C. F. Randin, T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. Are niche-based species distribution models transferable in space? *Journal of Biogeography*, 33:1689–1703, 2006. URL https://doi.org/10.1111/j.1365-2699.2006.01466.x. [p3]

B. Ripley. *tree: Classification and Regression Trees*, 2016. URL https://CRAN.R-project.org/package=tree. R package version 1.0-37. [p7]

D. San-Martín, R. Manzanas, S. Brands, S. Herrera, and J. M. Gutiérrez. Reassessing Model Uncertainty for Regional Projections of Precipitation with an Ensemble of Statistical Downscaling Methods. *Journal of Climate*, 30:203–223, 2016. URL https://doi.org/10.1175/jcli-d-16-0366.1. [p9]

S. D. Senay, S. P. Worner, and T. Ikeda. Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE*, 8:e71218, 2013. URL https://doi.org/10.1371/journal.pone.0071218. [p2, 4]

M. J. Serra-Varela, D. Grivet, L. Vincenot, O. Broennimann, J. Gonzalo-Jiménez, and N. E. Zimmermann. Does phylogeographical structure relate to climatic niche divergence? a test using maritime pine (Pinus pinaster Ait.). *Global Ecology and Biogeography*, 24:1302–1313, 2015. URL https://doi.org/10.1111/geb.12369. [p3]

J. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988. URL https://doi.org/10.1126/science.3287615. [p6]

K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2011. URL https://doi.org/10.1175/bams-d-11-00094.1. [p1]

T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2017. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-11. [p7]

W. Thuiller, M. B. Araújo, R. G. Pearson, R. J. Whittaker, L. Brotons, and S. Lavorel. Biodiversity conservation: Uncertainty in predictions of extinction risk. *Nature*, 430:145–148, 2004. URL https://doi.org/10.1038/nature02716. [p1]

W. Thuiller, D. Georges, R. Engler, and F. Breiner. *biomod2: Ensemble Platform for Species Distribution Modeling*, 2016. URL https://CRAN.R-project.org/package=biomod2. R package version 3.3-7. [p2]

M. Turco, A. Sanna, S. Herrera, M.-C. Llasat, and J. M. Gutiérrez. Large biases and inconsistent climate change signals in ENSEMBLES regional projections. *Climatic Change*, 120:859–869, 2013. URL https://doi.org/10.1007/s10584-013-0844-y. [p1]

P. van der Linden and J. F. B. Mitchell. ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project — European Environment Agency (EEA). Technical report, Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK., 2009. [p3]

J. Van der Wal and L. P. Shoo. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220:589–594, 2009. URL https://doi.org/10.1016/j.ecolmodel.2008.11.010. [p6]

J. Van der Wal, L. Falconi, S. Januchowski, L. Shoo, and C. Storlie. *SDMTools: Species Distribution Modelling Tools: Tools for Processing Data Associated with Species Distribution Modelling Exercises*, 2014. URL https://CRAN.R-project.org/package=SDMTools. R package version 1.1-221. [p2]

D. L. Verbyla and J. A. Litvaitis. Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, 13:783–787, 1989. URL https://doi.org/10.1007/bf01868317. [p6]

J. A. Winkler, J. P. Palutikof, J. A. Andresen, and C. M. Goodess. 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, 1997. URL https://doi.org/10.1175/1520-0442(1997)010<2514:tsodtt>2.0.co;2. [p3]

M. S. Wisz and A. Guisan. Do pseudo-absence selection strategies influence species distribution models and their predictions? an information-theoretic approach based on simulated data. *BMC Ecology*, 9:8, 2009. URL https://doi.org/10.1186/1472-6785-9-8. [p4]

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77:1–17, 2017. URL https://doi.org/10.18637/jss.v077.i01. [p7]

M. Zahn and H. von Storch. Decreased frequency of North Atlantic polar lows associated with future climate warming. *Nature*, 467:309–312, 2010. URL https://doi.org/10.1038/nature09388. [p3]

M. Iturbide, J.M. Gutiérrez
https://orcid.org/0000-0002-5048-0941
https://orcid.org/0000-0002-2766-6297
Meteorology group. Insituto de Física de Cantabria (IFCA)
CSIC - Universidad de Cantabria. Avda. de los Castros, s/n
39005. Santander. Spain
miturbide@ifca.unican.es

J. Bedia
https://orcid.org/0000-0001-6219-4312
Predictia Intelligent Data Solutions S.L.
http://www.predictia.es/en
Avda. los Castros s/n. Edificio I+D+i. S345
39005. Santander. Spain