Received: 16 March 2017 | Accepted: 15 May 2017

DOI: 10.1111/2041-210X.12841

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

SSDM: An R package to predict distribution of species richness and composition based on stacked species distribution models

Sylvain Schmitt[1] | Robin Pouteau[1] | Dimitri Justeau[1] | Florian de Boissieu[2] | Philippe Birnbaum[3]

[1]Botany and Applied Plant Ecology Laboratory, New Caledonian Agronomic Institute (IAC), Nouméa, New Caledonia

[2]Institute for Sustainable Development (IRD), Noumea, New Caledonia

[3]CIRAD Languedoc-Roussillon, Montpellier, France

Correspondence
Sylvain Schmitt and Robin Pouteau
Email: sylvain.schmitt@agroparistech.fr; robin.pouteau@ird.fr

Funding information
Direction for Economic and Environmental Development (DDEE)

Handling Editor: Nick Golding

Abstract

1. There is growing interest among conservationists in biodiversity mapping based on stacked species distribution models (SSDMs), a method that combines multiple individual species distribution models to produce a community-level model. However, no user-friendly interface specifically designed to provide the basic tools needed to fit such models was available until now.
2. The "ssdm" package is a computer platform implemented in R providing a range of methodological approaches and parameterisation at each step in building the SSDM: e.g. pseudo-absence selection, variable contribution and model accuracy assessment, inter-model consensus forecasting, species assembly design, and calculation of weighted endemism.
3. The object-oriented design of the package is such that: users can modify existing methods, extend the framework by implementing new methods, and share them to be reproduced by others.
4. The package includes a graphical user interface to extend the use of SSDMs to a wide range of conservation scientists and practitioners.

KEYWORDS
bioinformatics, community ecology, conservation, habitats, modelling, software

1 | INTRODUCTION

Understanding how local species richness (α-diversity) is distributed is a critical prerequisite for effective conservation strategies. Richness maps can provide the basis for selecting reserves (Cañadas et al., 2014; Moraes, Ríos-Uzeda, Moreno, Huanca-Huarachi, & Larrea-Alcázar, 2014; Murray-Smith et al., 2009; Raes, Roos, Slik, Van Loon, & ter Steege, 2009), prevention of biological invasions (Bellard et al., 2013; Gallardo, Zieritz, & Aldridge, 2015; Kelly, Leach, Cameron, Maggs, & Reid, 2014; Pouteau, Hulme, & Duncan, 2015), and mitigation of future impacts of climate change (Bellard et al., 2013; Brown, Parks, Bethell, Johnson, & Mulligan, 2015; Colombo & Joly, 2010; Fitzpatrick, Gove, Sanders, & Dunn, 2008; Midgley, Hannah, Millar, Thuiller, & Booth, 2003; Ogawa-Onishi, Berry, & Tanaka, 2010; Siqueira & Peterson, 2003).

As it is not always possible to capture the complete variation in species richness over large areas using comprehensive species inventories, a range of more pragmatic methods has been developed to extrapolate scattered local observations. They include:

1. Point-to-grid maps, that assemble natural history records (e.g. herbarium or museum specimens) within grid cells and count the number of species observed in each cell (Birnbaum et al., 2015; Cañadas et al., 2014; Droissart, Hardy, Sonké, Dahdouh-Guebas, & Stévart, 2012; Tovaranonte, Blach-Overgaard, Pongsattayapipat, Svenning, & Barfod, 2015; Wulff et al., 2013). This method has the advantage of not extrapolating data, but as natural history records are seldom evenly sampled, the accuracy of this method tends to decrease with an increase in cell resolution and hence reaches its maximum reliability at a

scale that may be too coarse for local decision-makers (Graham & Hijmans, 2006).

2. Macroecological models (MEMs), that link species richness observed over a network of comprehensive species inventories (e.g. plots, transects, quadrats) with spatially explicit environmental variables (Bhattarai & Vetaas, 2003; Sánchez-González & López-Mata, 2005; Tomasetto, Duncan, & Hulme, 2013). These variables are typically hypothesised to be or correlate with available energy, environmental heterogeneity, disturbance or history, with scale effects and some level of stochasticity. Macroecological models have contributed substantially to our understanding of large-scale ecology and biodiversity, and predict site-level richness well, probably better and more consistently than multiple species distribution models (SDMs), which have substantial problems dealing with rare species (Graham & Hijmans, 2006; Guisan & Rahbek, 2011). However, MEMs have the disadvantage of requiring a large number of inventories to be accurately calibrated and appear to be unable to extrapolate beyond known communities (Ferrier & Guisan, 2006).
3. Stacked species distribution models (SSDMs), that combine multiple individual SDMs to produce a community-level model (Ferrier & Guisan, 2006). A major strength of an SSDM compared to a point-to-grid map or a MEM is that an SSDM can predict species assemblages, which the two others cannot. An SDM (also referred as to "ecological niche model," "habitat suitability model," and "predictive habitat distribution models") refers to the process of using a statistical method to predict the distribution of a species in geographical space on the basis of a mathematical representation of its known distribution in environmental space (Guisan & Thuiller, 2005). Diversity mapping based on multiple SDMs has great potential for conservationists and the growing interest in the method is obvious in the literature (e.g. Benito, Cayuela, & Albuquerque, 2013; Brown et al., 2015; Colombo & Joly, 2010; D'Amen, Dubuis, et al., 2015; D'Amen, Pradervand, & Guisan, 2015; Fitzpatrick et al., 2008; Mateo, de la Estrella, Felicisimo, Muñoz, & Guisan, 2012; Midgley et al., 2003; Moraes et al., 2014; Murray-Smith et al., 2009; Ogawa-Onishi et al., 2010; Pérez & Font, 2012; Pouteau, Bayle, et al., 2015; Raes et al., 2009; Schmidt-Lebuhn, Knerr, & González-Orozco, 2012; Siqueira & Peterson, 2003).

Stacking individual species predictions can be applied to both rough probabilities (pSSDM) and binary predictions from SDMs (bSSDM) (e.g. Calabrese, Certain, Kraan, & Dormann, 2014; D'Amen, Dubuis, et al., 2015; D'Amen, Pradervand, et al., 2015; Dubuis et al., 2011). Macroecological models and pSSDMs both tend to perform similarly and to overestimate at sites with low species richness and underestimate at sites with high species richness (Calabrese et al., 2014). In contrast, bSSDMs tend to overpredict species richness, which is associated with generally higher and asymmetric prediction errors than MEMs, and may be affected by the choice of threshold for making binary predictions (Benito et al., 2013; Calabrese et al., 2014; Cord, Klein, Gernandt, de la Rosa, & Dech, 2014; D'Amen, Pradervand, et al., 2015; Dubuis et al., 2011).

Several authors also reported that SSDMs consistently overpredict species richness compared to MEMs because SSDMs reconstruct communities on the basis of species-specific abiotic filters without considering macroecological constraints to the general properties of the community as a whole (Guisan and Rahbek, 2011; Hortal, De Marco, Santos, & Diniz-Filho, 2012). These constraints are thought to be of increasing importance in structuring communities at increasing resolution and should thus be accounted for in fine-scale biodiversity assessments (Thuiller, Pollock, Gueguen, & Münkemüller, 2015). To remedy this problem, Guisan and Rahbek (2011) proposed the integrated framework SESAM (spatially explicit species assemblage modelling). The idea is to apply four successive filters in the assembly process: (1) dispersal filtering; (2) abiotic habitat filtering using SDMs; (3) macroecological constraints using MEMs; and (4) biotic filtering by applying ecological assembly rules (e.g. maximum species richness) (Guisan & Rahbek 2011). A commonly used assembly rule is the "probability ranking" rule (PRR): community composition is determined by ranking the species in decreasing order of their predicted probability up to the richness prediction (D'Amen, Dubuis, et al., 2015; D'Amen, Pradervand, et al., 2015). The core assumption behind this rule is that species with the highest habitat suitability are competitively superior. Other assembly rules include the "trait range" rule (D'Amen, Dubuis, et al., 2015) and the "checkerboard unit" rule (D'Amen, Pradervand, et al., 2015).

More recently, the core assumptions on which SESAM is based (SSDMs overpredict richness compared to MEMs) have been called into question by the convincing demonstration based on probability theory performed by Calabrese et al. (2014). These authors developed an innovative maximum-likelihood approach to adjust SSDM occurrence probabilities based on an estimate or prediction of site-level species richness. Supported by this innovative method, they argued that overprediction originates from a statistical rather than an ecological bias introduced using thresholding schemes to produce SSDMs. Thus, this statistical artefact could be caused by species prevalence and/or "regression dilution."

Since the publication of the SESAM framework, several other comprehensive modelling frameworks linking ecological theory, empirical data, and statistical models have been developed to predict communities, including the integrated framework of Boulangeat, Gravel, and Thuiller (2012), the metacommunity—space, environment, time model (M-SET; Mokany, Harwood, Williams, & Ferrier, 2012), and joint species distribution models (JSDMs; Pollock et al., 2014). These frameworks offer innovative ways to improve our understanding of community assembly processes at large spatial scales and for many species at once, based on species co-occurrence indices obtained from extensive community surveys and sometimes species-specific dispersal abilities. However, these recent frameworks received no further considerations as it would be virtually impossible to unite all community-level frameworks in a single software architecture and SESAM is still one of the best known, least complex, and least data-demanding frameworks produced to date.

While SSDMs provide increasingly promising predictions, no user-friendly interface specifically designed to provide the basic tools needed to build an SSDM was available until now (Table 1). Here, we present a new package named "ssdm" which is a free and open source object-oriented platform for stacked species distribution modelling

TABLE 1 A non-exhaustive list of software packages designed to perform species distribution modelling with their main advantages and limitations

Software	Graphical user interface	Developed in R	Designed to fit SSDMs	Evaluation of species composition	References
BIOENSEMBLES	X				Diniz-Filho et al. (2009)
BIOMOD2		X			Thuiller et al. (2009)
MODECO	X				Guo and Liu (2010)
OPENMODELLER	X				de Souza Muñoz et al. (2009)
SDM	X[a]	X	X[a]		Naimi and Araújo (2016)
SSDM	X	X	X	X	This article

[a]Included in the package description in Naimi and Araújo (2016) but not available in the latest package release (version 1.0-10).

implemented in R (R Core Team, 2015). R is perhaps the most commonly used software for ecological analysis in which state-of-the-art methods can easily be incorporated. The "SSDM" package provides a standardised and unified structure for visualizing and handling species distribution data and models. It also provides a range of cutting-edge methods including nine statistical methods and makes it possible to build ensembles of forecasts to account for inter-model variability. The user-friendly interface is likely to extend the use of SSDMs to a wide range of conservation scientists and practitioners.

2 | MODEL FLOW

The workflow of the package "SSDM" is based on three levels: (1) an individual SDM is fitted by linking the occurrences of a single species to environmental predictor variables based on the response curve of a single statistical method; (2) for each species, an ensemble SDM (ESDM) can be created from the outputs of several statistical methods to create a model that captures components of each; and (3) species assemblage from an SSDM is predicted by stacking several SDM or ESDM outputs (Figure 1).

2.1 | Data inputs

2.1.1 | Natural history records

Most statistical methods included in the "SSDM" package (introduced below) require presence/absence datasets. When a sampling scheme did not account for species absences (presence-only data), the package selects pseudo-absences (randomly selected sites where a species is assumed to be absent) or background data. Three modalities can be set to generate pseudo-absences: (1) the selection strategy: either within the extent of the set of environmental rasters or within a user-specified distance from each presence; (2) the number of selected pseudo-absences: either a user-specified number or a number equal to the number of presences available for each species; and (3) the number of times the pseudo-absence selection is repeated to reduce potential errors due to randomisation in selection (Barbet-Massin, Jiguet, Albert, & Thuiller, 2012). When pseudo-absences are selected repeatedly, the package merges the results of all runs by averaging habitat suitability probabilities and the associated accuracy metrics. Default parameters have been set to recommendations from Barbet-Massin et al. (2012) adapted to each statistical method (e.g. 10 runs of 1,000 randomly selected pseudo-absences are performed for GLM). The R package for spatial thinning of species occurrences "SPTHIN" (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015) was integrated to deal with natural history records deviation from opportunistic sampling scheme prone to spatial autocorrelation. The aim of thinning is to remove the fewest possible records needed to reduce the effect of sampling bias, while retaining the greatest possible amount of information.

2.1.2 | Environmental variables

All raster formats supported by the R "RGDAL" package can be used with the "SSDM" package to describe the environment occupied by the species thereby facilitating data management and exchange with conventional GIS packages (Bivand et al., 2016). The "SSDM" package accepts both continuous (e.g. climate maps, digital elevation models, bathymetric maps) and categorical environmental variables (e.g. land cover maps, soil type maps) as inputs. The package also allows normalisation of environmental variables, which may be useful to improve the fit of certain statistical methods (e.g. artificial neural networks).

Rasters of environmental variables must have the same coordinate reference system but the spatial extent and resolution of the environmental layers can differ. During processing, the package will deal with between-variables discrepancies in spatial extent and resolution by rescaling all environmental rasters to the smallest common spatial extent, and then upscaling them to the coarsest resolution using nearest neighbour interpolation.

2.2 | Statistical methods

2.2.1 | Individual species distribution models

The "SSDM" package includes the main statistical methods used to model species distributions: general additive models, generalised

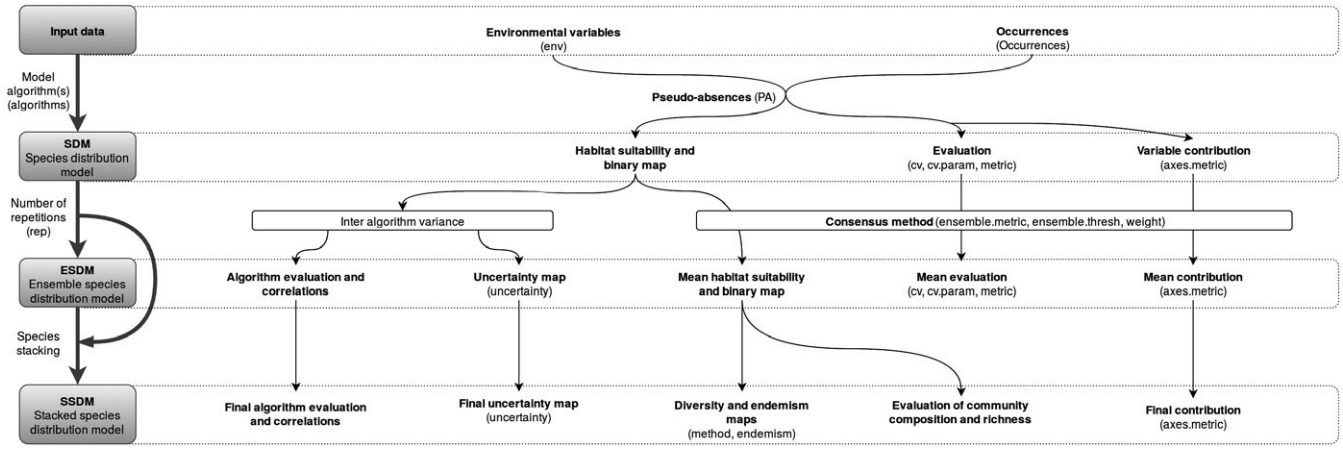

FIGURE 1 Flow chart of the "SSDM" package

linear models (GLM), multivariate adaptive regression splines, classification tree analysis, generalised boosted models, maximum entropy, artificial neural networks (ANN), random forests, and support vector machines. The default parameters of the dependent R package of each statistical method were conserved but most of them can be reset (Table 2).

A major assumption behind the concept of SDM is that species are in equilibrium with their environment and so barriers to species dispersal are consequently ignored by the most standard SDM implementations (Guisan & Thuiller, 2005). Hence, an SDM may overestimate the geographical area that a species occupies if its distribution is at least partially shaped by dispersal barriers. In order to account for this potential bias, the package contains an option to restrict SDM predictions to a user-specified distance around each presence (a habitat suitability of 0 is then assigned to the remainder of the study area) (Crisp, Laffan, Linder, & Monro, 2001).

For each species, the package can store two results in raster format: (1) a continuous raster map giving the habitat suitability for presence-only data, and the probability of presence for presence/absence data; and (2) a binary presence/absence raster based on the threshold of habitat suitability that maximises a user-specified accuracy metric (see below).

2.2.2 | Ensemble species distribution models (ESDMs)

Because uncertainty in distribution projections can skew policy making and planning, one recommendation is to fit a number of alternative statistical methods and to explore the range of projections across the different SDMs, and then to find a consensus among SDM projections (Gritti, Duputie, Massol, & Chuine, 2013; Marmion, Parviainen, Luoto, Heikkinen, & Thuiller, 2009). Two consensus methods are implemented in the "SSDM" package: (1) a simple average of the SDM outputs; and (2) a weighted average based on a user-specified metric or group of metrics (described below). The package also provides an uncertainty map representing the between-methods variance. The degree of agreement between each pair of statistical methods can be assessed through a correlation matrix that gives the Pearson's coefficient.

TABLE 2 Statistical methods implemented in the first release of the "SSDM" package and their dependent packages

Statistical method	Dependent package	References
GAM	MGCV	Wood (2006)
GLM	STATS	R Core Team (2015)
MARS	EARTH	Milborrow (2016)
MAXENT	DISMO	Hijmans, Phillips, Leathwick, and Elith (2016)
CTA	RPART	Therneau, Atkinson, and Ripley (2015)
GBM	GBM	Ridgeway (2015)
ANN	NNET	Venables and Ripley (2002)
RF	RANDOMFOREST	Liaw and Wiener (2002)
SVM	E1071	Meyer, Dimitriadou, Hornik, Weingessel, and Leisch (2015)

2.2.3 | Stacked species distribution models

The final maps of local species richness and composition can be computed using six different methods: (1) by summing discrete presence/absence maps (bSSDM) derived from one of the six metrics available to compute binary maps detailed in the next section (e.g. Benito et al., 2013; Brown et al., 2015; Fitzpatrick et al., 2008; Midgley et al., 2003; Moraes et al., 2014; Ogawa-Onishi et al., 2010; Raes et al., 2009); (2) by summing discrete presence/absence maps obtained by drawing repeatedly from a Bernoulli distribution (see Dubuis et al., 2011; Calabrese et al., 2014 for further details); (3) by summing continuous habitat suitability maps (pSSDM) (e.g. Mateo et al., 2012; Murray-Smith et al., 2009; Pouteau, Bayle, et al., 2015; Schmidt-Lebuhn et al., 2012); (4) by applying the PRR of the SESAM framework (a number of species equal to the prediction of species richness is selected on the basis of decreasing probability of presence calculated by the SDMs)

with species richness as estimated by a pSSDM (referred to as "PRR.pSSDM") (D'Amen, Dubuis, et al., 2015); (5) by applying the PRR with species richness as estimated by a MEM ("PRR.MEM") (D'Amen, Dubuis, et al., 2015; D'Amen, Pradervand, et al., 2015; Guisan & Rahbek, 2011); and (6) using the maximum-likelihood adjustment approach proposed by Calabrese et al. (2014).

As the computation of multiple ESDM (one per species) can be time consuming, the R "PARALLEL" package has been included to optimise the use of a multi-core processor or a computer cluster (R Core Team, 2015). Computed maps can be exported in GeoTIFF then imported into other GIS software packages for further data analysis and visualisation.

2.3 | Additional outputs

2.3.1 | Model accuracy assessment

A range of metrics to evaluate models have been integrated in the "SSDM" package using the "SDMTOOLS" package (VanDerWal, Falconi, Januchowski, Shoo, & Storlie, 2014). They include the area under the receiving operating characteristic (ROC) curve (AUC), Cohen's kappa coefficient, the omission rate, the sensitivity (true positive rate) and the specificity (true negative rate) (Fielding & Bell, 1997). These metrics are all based on the confusion matrix (also called "error matrix," that represents the instances in a predicted class vs. the instances in an actual class) and, consequently, require prior conversion of habitat suitability maps into binary presence/absence maps. The optimal threshold to split presences and absences on the basis of habitat suitability probabilities can be set to the probability that maximises: Cohen's kappa coefficient, the correct classification rate, the true skill statistic (TSS), sensitivity/specificity equality (SES), the lowest prediction occurrence probability or the shortest distance between the ROC curve and the upper left corner of the ROC plot. Recommendations by Liu, Berry, Dawson, and Pearson (2005), Liu, White, and Newell (2013) for thresholding were set to default in the package (TSS or SES for presence-only and presence-absence datasets respectively). To ensure independence between the training and evaluation sets for cross-validation, three methods are available to split the initial dataset: (1) "holdout," in which the initial dataset is partitioned into separate training and evaluation sets by a user-defined fraction, (2) "*k*-folds," in which the initial dataset is partitioned into *k* folds being *k*-1 times the training set and once the evaluation set, and (3) "leave-one-out," in which each point is successively used for evaluation.

To assess the accuracy of an SSDM, the package provides the opportunity to compare modelled species assemblages with species pools from independent inventories observed in the field. Six evaluation metrics can be computed: (1) the species richness error, i.e. the difference between the predicted and observed species richness; (2) assemblage prediction success, i.e. the proportion of correct predictions; (3) Cohen's kappa of the assemblage, i.e. the proportion of specific agreement; (4) assemblage specificity, i.e. the proportion of true negatives (species that are both predicted and observed to be absent); (5) assemblage sensitivity, i.e. the proportion of true positives (species that are both predicted and observed as present); and (6) the Jaccard index, a widely used metric of community similarity (Pottier et al., 2013).

2.3.2 | Importance analysis of environmental variables

The "SSDM" package provides two measures of the relative contribution of environmental variables on a species-by-species basis, which quantifies the relevance of an environmental variable to determine species distribution. The first measure is based on a jackknife approach that evaluates the change in accuracy between a full model and a model in which each environmental variable is omitted in turn (Phillips, Anderson, & Schapire, 2006). All metrics available in the package can be used to assess the change in accuracy. The second measure is based on Pearson's correlation coefficient between a full model and a model with each environmental variable omitted in turn (Thuiller, Lafourcade, Engler, & Araújo, 2009). These measures, which are calculated on a species-by-species basis, are averaged in SSDMs.

2.3.3 | Endemism mapping

In addition to species richness, endemism is an important feature for conservation as it refers to species being unique to the defined geographical location (Crisp et al., 2001; Moraes et al., 2014; Raes et al., 2009). The "SSDM" package offers the opportunity to map local species endemism using two metrics: (1) the weighted endemism index (WEI); and (2) the corrected weighted endemism index (CWEI) (Crisp et al., 2001):

$$\mathrm{WEI}_c = \sum_{i=1}^{n_c} \frac{1}{r_{i,c}} \quad (1)$$

WEI for the cell *c* is calculated by summing the inverse of the geographical range size $r_{i,c}$ for each of the n_c species. WEI seeks to avoid the problem that an arbitrary region or range-size threshold is used to define what constitutes an endemic species. WEI avoids using a threshold for endemism by applying a simple continuous weighting function, assigning high weights to species with small ranges, and progressively smaller weights to species with larger ranges.

$$\mathrm{CWEI}_c = \frac{\mathrm{WEI}_c}{\mathrm{RS}_c} \quad (2)$$

CWEI is an alternative measure to reduce the correlation between richness and endemism. CWEI for cell *c* is calculated as the weighted endemism index WEI_c divided by the richness score RS_c so that CWEI_c represents the average degree of endemism of the species recorded in an area.

3 | GRAPHICAL USER INTERFACE

The "SSDM" package offers a user-friendly interface built with the web application framework for R Shiny (Chang, Cheng, Allaire, Xie, &

McPherson, 2016). The graphical user interface is launched with the function *gui()*. The interface is divided into three steps: data loading, modelling, and results display. The "Load" tab allows a new dataset or a previously saved model to be loaded. The "Modelling" tab proposes three types of models: an individual SDM, an ESDM, or a SSDM. The "Modelling" tab contains three sub-tabs offering levels of parameterisation that are more or less detailed depending on the user's level of expertise: (1) basic, to select the statistical method(s), the number of runs per statistical method, the model evaluation metric(s), and the methods to be used to map diversity and endemism; (2) intermediate, to set pseudo-absence selection (number and strategy), the cross-validation method, the metric used to estimate the relative contribution of environmental variables, the ESDM consensus method, and the SSDM stacking method; and (3) advanced, to set the parameters of the statistical methods. The "Results" tab summarises graphic modelling outputs: model maps (species habitat suitability, species richness and endemism), the relative contributions of environmental variables, assessment of model accuracy, and between-methods correlation (Figure 2). The interface includes a panel to save results maps in GeoTIFF format (.tif) compatible with most GIS software, and other numerical results as comma separated values (.csv) files.

4 | EXAMPLES

4.1 | Vulnerability to invasive species at global scale

The occurrences of 100 of the world's worst invasive alien species (as defined by the Invasive Species Specialist Group of the International Union for Conservation of Nature; http://www.issg.org/) were gathered from the Global Biodiversity Information Facility (http://www.gbif.org/). Occurrences flagged as invalid, or doubtful coordinates, or mismatching country, or doubtful taxon, were removed. The set of 19 WorldClim climate variables (all continuous) at a 2.5 arcmin resolution were used as environmental variables (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). Multicollinearity of variables was addressed by examining cross-correlations. For variables with Pearson's correlations of $r > .8$, the variable that decreased model accuracy the most when omitted from the full model (i.e. the most "meaningful" variable) was retained. Next, an SSDM using the sum of individual probabilities (pSSDM) as stacking method and with all other model settings set to default was fitted. The output provides a picture of how richness in 100 of the world's worst invasive alien species could be distributed without any barriers to spread or competitive interactions (Figure 3).

4.2 | *Endemism of the genus* Psychotria *in New Caledonia*

Psychotria (Rubiaceae) is the second most speciose genus on the megadiverse archipelago of New Caledonia (Southwest Pacific Ocean) (Barrabé et al., 2014). Occurrences of all native species described as belonging to this genus were extracted from the Noumea (NOU) VIROT database and the Paris herbaria (P) SONNERAT database. Six environmental variables (five continuous and one categorical) at 100 m resolution were used to fit an SSDM: elevation, potential insolation, slope steepness, substrate type, windwarness, and a topographical wetness index (see Pouteau, Bayle, et al., 2015 for further details). Continuous variables were correlated with a Pearson's $r < .80$. A WEI map was built with all model settings set to default. The output

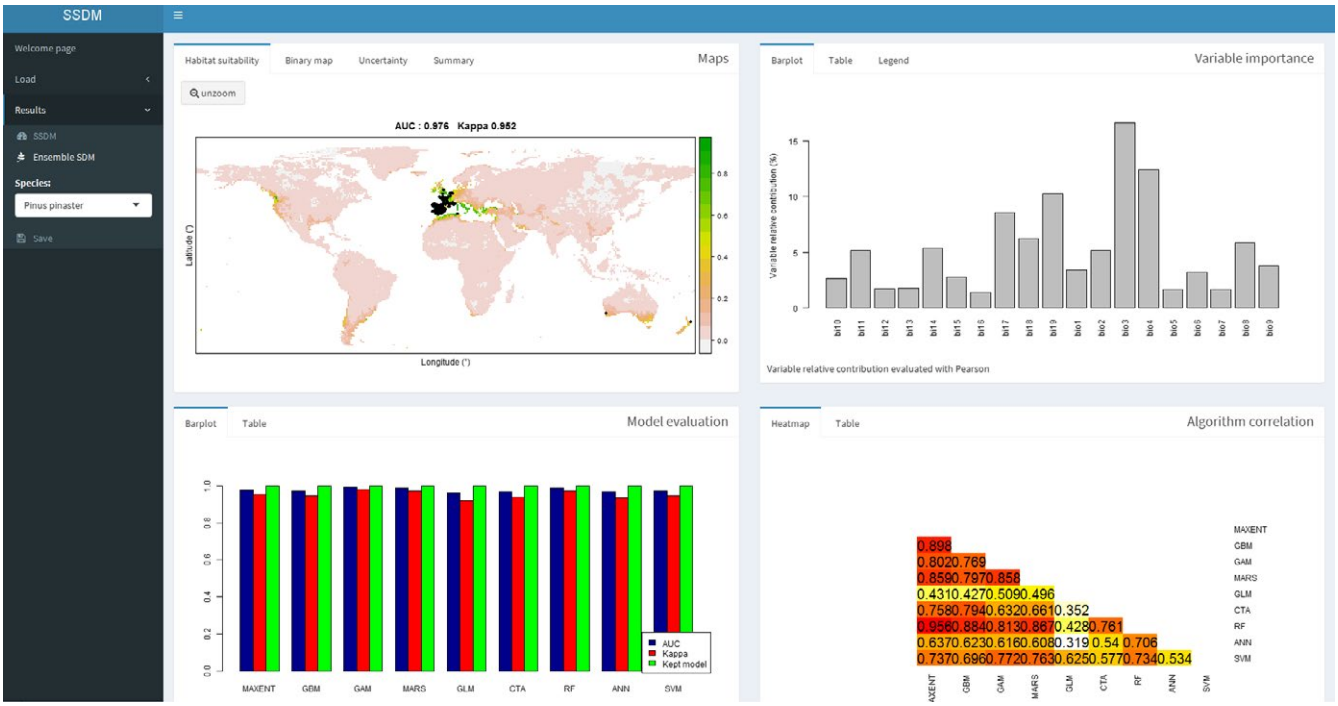

FIGURE 2 Screenshot of the results dashboard displayed by the graphical user interface of the "SSDM" package

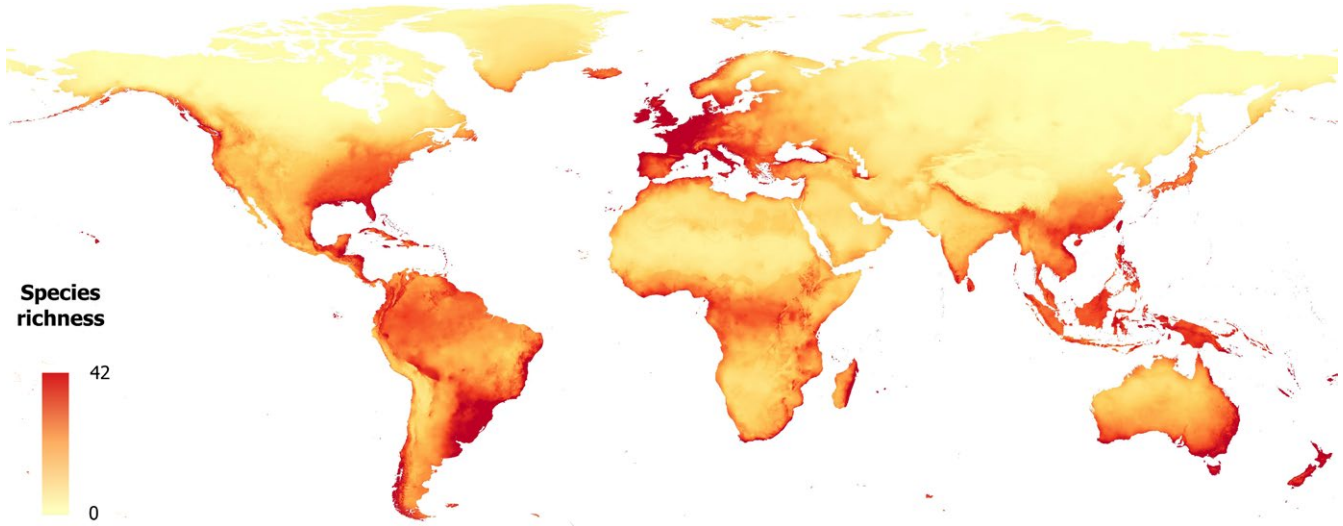

FIGURE 3 World map of the vulnerability to the 100 world's worst invasive species generated with the "SSDM" package

provides a picture of how the level of endemism of this focal genus is spatially organised in New Caledonia (Figure 4).

5 | INSTALLATION

The "SSDM" package is free and open source (version 0.2.3 with GPL v3 license). It is available from the CRAN repository https://cran.r-project.org/web/packages/SSDM/index.html, and can be installed either from CRAN or within the R environment using the command *install.packages("SSDM")*. The project is hosted on Github (https://github.com/sylvainschmitt/SSDM), which allows future users to openly contribute to the project.

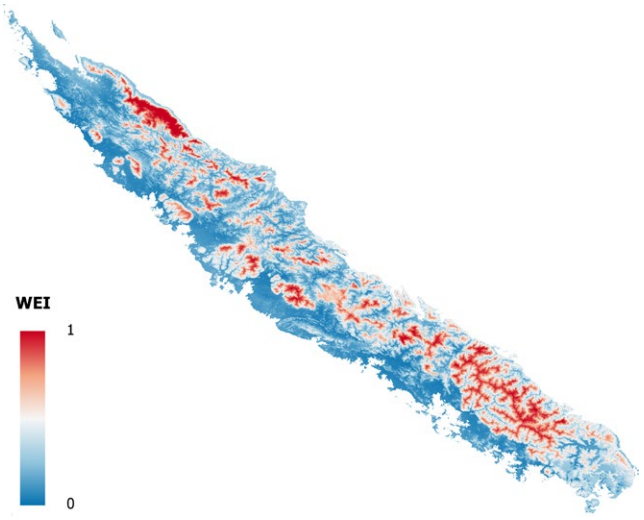

FIGURE 4 Weighted endemism map of the genus Psychotria in New Caledonia generated with the "SSDM" package

ACKNOWLEDGEMENTS

We are grateful to Maxime Réjou-Méchain (IRD) and Thomas Ibanez (IAC) for their useful comments on an earlier draft of the manuscript, to Laure Barrabé (IAC) and Frédéric Rigault (IRD) for gathering and pre-processing the occurrences of *Psychotria* used in the second example, to Jérôme Lefèvre (IRD) and the IRD high performance computing platform in Noumea for making the infrastructure available for parallelisation tests, and to Daphne Goodfellow for English revisions. We also would like to thank the "BIOMOD2" package for inspiration. The implementation of the "SSDM" package was funded by the Direction for Economic and Environmental Development (DDEE) of the North Province of New Caledonia. This manuscript benefited from the helpful suggestions made by three anonymous referees.

AUTHORS' CONTRIBUTIONS

S.S., R.P., D.J., and P.B. conceived and designed the software; S.S., D.J., and F.B. implemented the package; S.S. and R.P. led the writing of the manuscript. All the authors contributed critically to the draft and gave final approval for publication.

DATA ACCESSIBILITY

The occurrences of 100 of the world's worst invasive alien species: Global Biodiversity Information Facility https://doi.org/10.15468/dl.2mvxxk. The set of 19 WorldClim climate variables: http://www.worldclim.org/current (2.5 min). Psychotria data has not been archived because the locations of the endangered species cannot be disclosed. The methods used to produce Figure 4 can be fully reproduced using the Cryptocaria data included into the "SSDM" package with the associated vignette.

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*, 541–545.

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.

Barrabé, L., Maggia, L., Pillon, Y., Rigault, F., Mouly, A., Davis, A. P., & Buerki, S. (2014). New Caledonian lineages *Psychotria* (Rubiaceae) reveal different evolutionary histories and the largest documented plant radiation for the archipelago. *Molecular Phylogenetics and Evolution*, *71*, 15–35.

Bellard, C., Thuiller, W., Leroy, B., Genovesi, P., Bakkenes, M., & Courchamp, F. (2013). Will climate change promote future invasions? *Global Change Biology*, *19*, 3740–3748.

Benito, B. M., Cayuela, L., & Albuquerque, F. S. (2013). The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: Guidelines to build better diversity models. *Methods in Ecology and Evolution*, *4*, 327–335.

Bhattarai, K. R., & Vetaas, O. R. (2003). Variation in plant species richness of different life forms along a subtropical elevation gradient in the Himalayas, east Nepal. *Global Ecology and Biogeography*, *12*, 327–340.

Birnbaum, P., Ibanez, T., Pouteau, R., Vandrot, H., Hequet, V., Blanchard, E., & Jaffré, T. (2015). Environmental correlates for tree occurrences, species distribution and richness on a high-elevation tropical island. *AOB Plants*, *7*, plv075.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., & Rouault, E. (2016). Bindings for the geospatial data abstraction library. R package version 1.1-10. Retrieved from https://CRAN.R-project.org/package=rgdal

Boulangeat, I., Gravel, D., & Thuiller, W. (2012). Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, *15*, 584–593.

Brown, K. A., Parks, K. E., Bethell, C. A., Johnson, S. E., & Mulligan, M. (2015). Predicting plant diversity patterns in Madagascar: Understanding the effects of climate and land cover in a biodiversity hotspot. *PLoS ONE*, *10*, e0122721.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models: Stacking species distribution models. *Global Ecology and Biogeography*, *23*, 99–112.

Cañadas, E. M., Fenu, G., Peñas, J., Lorite, J., Mattana, E., & Bacchetta, G. (2014). Hotspots within hotspots: Endemic plant richness, environmental drivers, and implications for conservation. *Biological Conservation*, *170*, 282–291.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2016). shiny: Web application framework for R. R package version 0.13.2. Retrieved from https://CRAN.R-project.org/package=shiny

Colombo, A. F., & Joly, C. A. (2010). Brazilian Atlantic Forest lato sensu: The most ancient Brazilian forest, and a biodiversity hotspot, is highly threatened by climate change. *Brazilian Journal of Biology*, *70*, 697–708.

Cord, A. F., Klein, D., Gernandt, D. S., de la Rosa, J. A. P., & Dech, S. (2014). Remote sensing data can improve predictions of species richness by stacked species distribution models: A case study for Mexican pines. *Journal of Biogeography*, *41*, 736–748.

Crisp, M. D., Laffan, S., Linder, H. P., & Monro, A. (2001). Endemism in the Australian flora. *Journal of Biogeography*, *28*, 183–198.

D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, *42*, 1255–1266.

D'Amen, M., Pradervand, J.-N., & Guisan, A. (2015). Predicting richness and composition in mountain insect communities at high resolution: A new test of the SESAM framework. *Global Ecology and Biogeography*, *24*, 1443–1453.

de Souza Muñoz, M. E., De Giovanni, R., de Siqueira, M. F., Sutton, T., Brewer, P., Pereira, R. S., … Canhos, V. P. (2009). openModeller: A generic approach to species' potential distribution modelling. *GeoInformatica*, *15*, 111–135.

Diniz-Filho, J. A. F., Bini, L. M., Rangel, T. F., Loyola, R. D., Hof, C., Nogués-Bravo, D., & Araújo, M. B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, *32*, 897–906.

Droissart, V., Hardy, O. J., Sonké, B., Dahdouh-Guebas, F., & Stévart, T. (2012). Subsampling herbarium collections to assess geographic diversity gradients: A case study with endemic Orchidaceae and Rubiaceae in Cameroon. *Biotropica*, *44*, 44–52.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: A comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions*, *17*, 1122–1131.

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, *43*, 393–404.

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.

Fitzpatrick, M. C., Gove, A. D., Sanders, N. J., & Dunn, R. R. (2008). Climate change, plant migration, and range collapse in a global biodiversity hotspot: The Banksia (Proteaceae) of Western Australia. *Global Change Biology*, *14*, 1337–1352.

Gallardo, B., Zieritz, A., & Aldridge, D. C. (2015). The importance of the human footprint in shaping the global distribution of terrestrial, freshwater and marine invaders. *PLoS ONE*, *10*, e0125801.

Graham, C. H., & Hijmans, R. J. (2006). A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, *15*, 578–587.

Gritti, E. S., Duputie, A., Massol, F., & Chuine, I. (2013). Estimating consensus and associated uncertainty between inherently different species distribution models. *Methods in Ecology and Evolution*, *4*, 442–452.

Guisan, A., & Rahbek, C. (2011). SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, *38*, 1433–1444.

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

Guo, Q., & Liu, Y. (2010). ModEco: An integrated software package for ecological niche modeling. *Ecography*, *33*, 637–642.

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., Phillips, S., Leathwick, J., & Elith, J. (2016). dismo: Species distribution modelling. R package version 1.0-15. Retrieved from https://CRAN.R-project.org/package=dismo

Hortal, J., De Marco, Jr. P., Santos, A. M. C., & Diniz-Filho, J. A. F. (2012). Integrating biogeographical processes and local community assembly. *Journal of Biogeography*, *39*, 627–628.

Kelly, R., Leach, K., Cameron, A., Maggs, C. A., & Reid, N. (2014). Combining global climate and regional landscape models to improve prediction of invasion risk. *Diversity and Distributions*, *20*, 884–894.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*, 18–22.

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. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, *40*, 778–789.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*, 59–69.

Mateo, R. G., de la Estrella, M., Felicisimo, A. M., Muñoz, J., & Guisan, A. (2012). A new spin on a compositionalist predictive modelling

framework for conservation planning: A tropical case study in Ecuador. *Biological Conservation*, *160*, 150–161.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version 1.6-7. Retrieved from https://CRAN.R-package.org?package=e1071

Midgley, G. F., Hannah, L., Millar, D., Thuiller, W., & Booth, A. (2003). Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape Floristic Region. *Biological Conservation*, *112*, 87–97.

Milborrow, S. (2016). earth: Multivariate adaptive regression splines. R package version 4.4.4. Retrieved from https://CRAN.R-project.org/package=earth

Mokany, K., Harwood, T. D., Williams, K. J., & Ferrier, S. (2012). Dynamic macroecology and the future for biodiversity. *Global Change Biology*, *18*, 3149–3159.

Moraes, M. M., Ríos-Uzeda, B., Moreno, L. R., Huanca-Huarachi, G., & Larrea-Alcázar, D. (2014). Using potential distribution models for patterns of species richness, endemism, and phytogeography of palm species in Bolivia. *Tropical Conservation Science*, *7*, 45–60.

Murray-Smith, C., Brummitt, N. A., Oliveira-Filho, A. T., Bachman, S., Moat, J., Lughadha, E. M. N., & Lucas, E. J. (2009). Plant diversity hotspots in the Atlantic coastal forests of Brazil. *Conservation Biology*, *23*, 151–163.

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

Ogawa-Onishi, Y., Berry, P. M., & Tanaka, N. (2010). Assessing the potential impacts of climate change and their conservation implications in Japan: A case study of conifers. *Biological Conservation*, *143*, 1728–1736.

Pérez, N., & Font, X. (2012). Predicting vascular plant richness patterns in Catalonia (NE Spain) using species distribution models. *Applied Vegetation Science*, *15*, 390–400.

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

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406.

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C. F., … Guisan, A. (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, *22*, 52–63.

Pouteau, R., Bayle, E., Blanchard, E., Birnbaum, P., Cassan, J.-J., Hequet, V., … Vandrot, H. (2015). Accounting for the indirect area effect in stacked species distribution models to map species richness in a montane biodiversity hotspot. *Diversity and Distributions*, *21*, 1329–1338.

Pouteau, R., Hulme, P. E., & Duncan, R. P. (2015). Widespread native and alien plant species occupy different habitats. *Ecography*, *68*, 462–471.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raes, N., Roos, M. C., Slik, J. W. F., Van Loon, E. E., & ter Steege, H. (2009). Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography*, *32*, 180–192.

Ridgeway, G. (2015). gbm: Generalized boosted regression models. R package version 2.1.1. Retrieved from https://CRAN>R-project.org/package=gbm

Sánchez-González, A., & López-Mata, L. (2005). Plant species richness and diversity along an altitudinal gradient in the Sierra Nevada, Mexico. *Diversity and Distributions*, *11*, 567–575.

Schmidt-Lebuhn, A. N., Knerr, N. J., & González-Orozco, C. E. (2012). Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography*, *39*, 2072–2080.

Siqueira, M. F., & Peterson, A. T. (2003). Consequences of global change for geographic distributions of cerrado tree species. *Biota Neotropica*, *3*, 1–14.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees. R package version 4.1-10. Retrieved from https://CRAN.R-project.org/package=rpart

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD – A platform for ensemble forecasting of species distributions. *Ecography*, *32*, 369–373.

Thuiller, W., Pollock, L. J., Gueguen, M., & Münkemüller, T. (2015). From species distributions to meta-communities. *Ecology Letters*, *18*, 1321–1328.

Tomasetto, F., Duncan, R. P., & Hulme, P. E. (2013). Environmental gradients shift the direction of the relationship between native and alien plant species richness. *Diversity and Distributions*, *19*, 49–59.

Tovaranonte, J., Blach-Overgaard, A., Pongsattayapipat, R., Svenning, J.-C., & Barfod, A. S. (2015). Distribution and diversity of palms in a tropical biodiversity hotspot (Thailand) assessed by species distribution modeling. *Nordic Journal of Botany*, *33*, 214–224.

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. Retrieved from https://CRAN.R-project.org/package=SDMTools

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*, 4th ed. New York, NY: Springer.

Wood, S. N. (2006). *Generalized additive models*. Boca Raton, FL: Chapman and Hall/CRC.

Wulff, A. S., Hollingsworth, P. M., Ahrends, A., Jaffré, T., Veillon, J.-M., L'Huillier, L., & Fogliani, B. (2013). Conservation priorities in a biodiversity hotspot: Analysis of narrow endemic plant species in New Caledonia. *PLoS ONE*, *8*, e73371.

How to cite this article: Schmitt S, Pouteau R, Justeau D, de Boissieu F, Birnbaum P. ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models. *Methods Ecol Evol*. 2017;8:1795–1803. https://doi.org/10.1111/2041-210X.12841